GLOBAL EDITION

Essentials of Genetics

NINTH EDITION

William S. Klug • Michael R. Cummings
Charlotte A. Spencer • Michael A. Palladino

ALWAYS LEARNING

PEARSON

Brief Contents

ESSENTIALS *of* GENETICS

Ninth Edition
Global Edition

William S. Klug
The College of New Jersey

Michael R. Cummings
Illinois Institute of Technology

Charlotte A. Spencer
University of Alberta

Michael A. Palladino
Monmouth University

with contributions by
Darrell Killian
Colorado College

PEARSON

Senior Acquisitions Editor: Michael Gillespie
Project Manager: Margaret Young
Program Manager: Anna Amato
Development Editor: Dusty Friedman
Assistant Editor: Chloé Veylit
Executive Editorial Manager: Ginnie Simione Jutson
Program Management Team Lead: Mike Early
Project Management Team Lead: David Zielonka
Assistant Acquisitions Editor, Global Edition: Murchana Borthakur
Project Editor, Global Edition: Amrita Naskar
Manager, Media Production, Global Edition: Vikram Kumar
Senior Manufacturing Controller, Production, Global Edition: Trudy Kimber
Production Management: Rose Kernan, Cenveo® Publisher Services
Design Manager: Mark Ong
Interior Designer: Tani Hasegawa
Cover Designer: Lumina Datamatics Ltd.
Illustrators: Imagineering
Rights & Permissions Project Manager: Donna Kalal
Rights & Permissions Management: Rachel Youdelman
Photo Researcher: QBS Learning
Senior Procurement Specialist: Stacey Weinberger
Project Manager—Instructor Media: Chelsea Logan
Executive Marketing Manager: Lauren Harp

Cover Photo Credit: irin-k /Shutterstock

Acknowledgements of third party content appear on page C-1, which constitutes an extension of this copyright page.

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies throughout the world

Visit us on the World Wide Web at:
www.pearsonglobaleditions.com

© William S. Klug and Michael R. Cummings 2017

The rights of William S. Klug, Michael R. Cummings, Charlotte A. Spencer, and Michael A. Palladino to be identified as the authors of this work have been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

Authorized adaptation from the United States edition, entitled Essentials of Genetics, 9th edition, *ISBN 978-0-134-04779-9, by William S. Klug, Michael R. Cummings, Charlotte A. Spencer, and Michael A. Palladino, published by Pearson Education © 2016.*

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without either the prior written permission of the publisher or a license permitting restricted copying in the United Kingdom issued by the Copyright Licensing Agency Ltd, Saffron House, 6—10 Kirby Street, London EC 1N 8TS.

All trademarks used herein are the property of their respective owners. The use of any trademark in this text does not vest in the author or publisher any trademark ownership rights in such trademarks, nor does the use of such trademarks imply any affiliation with or endorsement of this book by such owners.

MasteringGenetics is a trademark in the U.S. and/or other countries, owned by Pearson Education, Inc. or its affiliates.

Unless otherwise indicated herein, any third-party trademarks that may appear in this work are the property of their respective owners and any references to third-party trademarks, logos or other trade dress are for demonstrative or descriptive purposes only. Such references are not intended to imply any sponsorship, endorsement, authorization, or promotion of Pearson's products by the owners of such marks, or any relationship between the owner and Pearson Education, Inc. or its affiliates, authors, licensees or distributors.

ISBN 10: 1-292-10886-X
ISBN 13: 978-1-292-10886-5

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

10 9 8 7 6 5 4 3 2 1

Typeset by Cenveo Publisher Services
Printed and bound by Vivar in Malaysia

About the Authors

William S. Klug is an Emeritus Professor of Biology at The College of New Jersey (formerly Trenton State College) in Ewing, New Jersey, where he served as Chair of the Biology Department for 17 years. He received his B.A. degree in Biology from Wabash College in Crawfordsville, Indiana, and his Ph.D. from Northwestern University in Evanston, Illinois. Prior to coming to The College of New Jersey, he was on the faculty of Wabash College as an Assistant Professor, where he first taught genetics, as well as general biology and electron microscopy. His research interests have involved ultrastructural and molecular genetic studies of development, utilizing oogenesis in *Drosophila* as a model system. He has taught the genetics course as well as the senior capstone seminar course in Human and Molecular Genetics to undergraduate biology majors for over four decades. He was the recipient in 2001 of the first annual teaching award given at The College of New Jersey, granted to the faculty member who "most challenges students to achieve high standards." He also received the 2004 Outstanding Professor Award from Sigma Pi International, and in the same year, he was nominated as the Educator of the Year, an award given by the Research and Development Council of New Jersey.

Michael R. Cummings is Research Professor in the Department of Biological, Chemical, and Physical Sciences at Illinois Institute of Technology, Chicago, Illinois. For more than 25 years, he was a faculty member in the Department of Biological Sciences and in the Department of Molecular Genetics at the University of Illinois at Chicago. He has also served on the faculties of Northwestern University and Florida State University. He received his B.A. from St. Mary's College in Winona, Minnesota, and his M.S. and Ph.D. from Northwestern University in Evanston, Illinois. In addition to this text and its companion volumes, he has also written textbooks in human genetics and general biology for nonmajors. His research interests center on the molecular organization and physical mapping of the heterochromatic regions of human acrocentric chromosomes. At the undergraduate level, he teaches courses in Mendelian and molecular genetics, human genetics, and general biology, and has received numerous awards for teaching excellence given by university faculty, student organizations, and graduating seniors.

Charlotte A. Spencer is a retired Associate Professor from the Department of Oncology at the University of Alberta in Edmonton, Alberta, Canada. She has also served as a faculty member in the Department of Biochemistry at the University of Alberta. She received her B.Sc. in Microbiology from the University of British Columbia and her Ph.D. in Genetics from the University of Alberta, followed by postdoctoral training at the Fred Hutchinson Cancer Research Center in Seattle, Washington. Her research interests involve the regulation of RNA polymerase II transcription in cancer cells, cells infected with DNA viruses, and cells traversing the mitotic phase of the cell cycle. She has taught courses in biochemistry, genetics, molecular biology, and oncology, at both undergraduate and graduate levels. In addition, she has written booklets in the Prentice Hall *Exploring Biology* series, which are aimed at the undergraduate nonmajor level.

Michael A. Palladino is Dean of the School of Science and Professor of Biology at Monmouth University in West Long Branch, New Jersey. He received his B.S. degree in Biology from Trenton State College (now known as The College of New Jersey) and his Ph.D. in Anatomy and Cell Biology from the University of Virginia. He directs an active laboratory of undergraduate student researchers studying molecular mechanisms involved in innate immunity of mammalian male reproductive organs and genes involved in oxygen homeostasis and ischemic injury of the testis. He has taught a wide range of courses for both majors and nonmajors and currently teaches genetics, biotechnology, endocrinology, and laboratory in cell and molecular biology. He has received several awards for research and teaching, including the 2009 Young Investigator Award of the American Society of Andrology, the 2005 Distinguished Teacher Award from Monmouth University, and the 2005 Caring Heart Award from the New Jersey Association for Biomedical Research. He is co-author of the undergraduate textbook *Introduction to Biotechnology*, Series Editor for the Benjamin Cummings *Special Topics in Biology* booklet series, and author of the first booklet in the series, *Understanding the Human Genome Project*.

This page intentionally left blank

Contents

Preface

Essentials of Genetics is written for courses requiring a text that is briefer and less detailed than its more comprehensive companion, *Concepts of Genetics*. While coverage is thorough and modern, *Essentials* is written to be more accessible to biology majors, as well as to students majoring in a number of other disciplines, including agriculture, animal husbandry, chemistry, nursing, engineering, forestry, psychology, and wildlife management. Because *Essentials of Genetics* is shorter than many other texts, it is also more manageable in one-quarter and trimester courses.

Goals

In this edition of *Essentials of Genetics*, the two most important goals have been to introduce pedagogic innovations that enhance learning and to provide carefully updated, highly accessible coverage of genetic topics of both historical and modern significance. As new tools and findings of genetics research continue to emerge rapidly and grow in importance in the study of all subdisciplines of biology, instructors face tough choices about what content is truly essential as they introduce the discipline to novice students. We have thoughtfully revised each chapter in light of this challenge, by selectively scaling back the detail or scope of coverage in the more traditional chapters in order to provide expanded coverage and broader context for the more modern, cutting-edge topics. Our aim is to continue to provide efficient coverage of the fundamental concepts in transmission and molecular genetics that lay the groundwork for more in-depth coverage of emerging topics of growing importance—in particular, the many aspects of the genomic revolution that is already relevant to our day-to-day lives as well as the relatively new findings involving epigenetics and noncoding RNAs.

While we have adjusted this edition to keep pace with changing content and teaching practices, we remain dedicated to the core principles that underlie this book. Specifically, we seek to

- Emphasize concepts rather than excessive detail.
- Write clearly and directly to students in order to provide understandable explanations of complex analytical topics.
- Emphasize problem solving, thereby guiding students to think analytically and to apply and extend their knowledge of genetics.
- Provide the most modern and up-to-date coverage of this exciting field.
- Propagate the rich history of genetics that so beautifully elucidates how information is acquired as the discipline develops and grows.
- Create inviting, engaging, and pedagogically useful figures enhanced by meaningful photographs to support student understanding.
- Provide outstanding interactive media support to guide students in understanding important concepts through animations, tutorial exercises, and assessment tools.

The above goals serve as the cornerstone of *Essentials of Genetics*. This pedagogic foundation allows the book to accommodate courses with many different approaches and lecture formats. While the book presents a coherent table of contents that represents one approach to offering a course in genetics, chapters are nevertheless written to be independent of one another, allowing instructors to utilize them in various sequences.

New to This Edition

In addition to streamlining core chapters and updating information throughout the text, key improvements to this edition include three additional chapters in the Special Topics in Modern Genetics unit, end of chapter questions in Special Topics chapters, and a new feature exploring scientists' evolving understanding of the concept of the gene.

- **Special Topics in Modern Genetics** We have been pleased with the popular reception to the Special Topics in Modern Genetics chapters. Our goal has been to provide abbreviated, cohesive coverage of important topics in genetics that are not always easily located in textbooks. Professors have used these focused, flexible chapters in a multitude of ways: as the backbone of lectures, as inspiration for student assignments outside of class, and as the basis of group assignments and presentations.

  New to this edition are chapters on topics of great significance in genetics:

  - Emerging Roles of RNA
  - Genetically Modified Foods
  - Gene Therapy

  For all Special Topics chapters, we have added a series of questions that send the student back into the chapter to review key ideas or that provide the basis of personal contemplations and group discussions.
- **Evolving Concept of the Gene** Also new to this edition is a short feature, integrated in appropriate chapters, that highlights how scientists' understanding of a gene has changed over time. Since we cannot see genes, we must infer just what this unit of heredity is, based on experimental findings. By highlighting how scientists' conceptualization of the gene has advanced over time, we aim to help students appreciate the process of discovery

that has led to an ever more sophisticated understanding of hereditary information.

- **Concepts Question** A new feature, found as the second question in the Problems and Discussion Questions at the end of each chapter, asks the student to review and comment on common aspects of the Key Concepts, listed at the beginning of each chapter. This feature places added emphasis on our pedagogic approach of conceptual learning.
- **MasteringGenetics** This powerful online homework and assessment program guides students through complex topics in genetics, using in-depth tutorials that coach students to correct answers with hints and feedback specific to their misconceptions. New content for *Essentials of Genetics* includes a robust library of Practice Problems—found only in MasteringGenetics—that are like end of chapter questions in scope and difficulty. These questions include wrong answer feedback specific to a student's error, helping build students' problem-solving and critical thinking skills.

New and Updated Topics

While we have revised each chapter in the text to present the most current findings in genetics, below is a list of some of the most significant new and updated topics present in this edition.

Ch. 1: Introduction to Genetics • New chapter introduction vignette emphasizing translational medicine

Ch. 2: Mitosis and Meiosis • Updated coverage of kinetechore assembly and the concept of disjunction • Expanded coverage of checkpoints in cell cycle regulation

Ch. 4: Modification of Mendelian Ratios • New section on mitochondria, human health, and aging

Ch. 5: Sex Determination and Sex Chromosomes • Updated coverage on paternal age effects (PAEs) in humans • New content regarding the primary sex ratio in humans

Ch. 6: Chromosome Mutations • New information on Fragile X Syndrome and the *FMRI* gene • New information regarding gene families as linked to gene duplications

Ch. 7: Linkage and Chromosome Mapping in Eukaryotes • Introduction of "sequence maps" in humans based on the use of DNA markers

Ch. 10: DNA Replication and Recombination • Updated coverage of DNA Pol III holoenzyme • Revised figures involving DNA synthesis • New coverage of the initiation of bacterial DNA synthesis • New information on DNA recombination • New coverage of replication of telomeric DNA • Revision of the GTS essay: Telomeres: The Key to Immortality

Ch. 11: Chromosome Structure and DNA Sequence Organization • Updated coverage of chromatin remodeling • New information on H3 histone substitution in centromeric DNA • New coverage regarding the transcript of Alu sequences

Ch. 12: The Genetic Code and Transcription • Extended coverage of promoter elements in eukaryotes • Introduction of the process of RNA editing • Revision of figures involving ribosomes and transcription

Ch. 13: Translation and Proteins • Revision of all ribosome figures • New information on initiation, elongation during translation in eukaryotes

Ch. 14: Gene Mutation, DNA Repair, and Transposition • Reorganization and updates for mutation classification • Updated coverage of xeroderma pigmentosum and DNA repair mechanisms

Ch. 15: Regulation of Gene Expression • Updated coverage of gene regulation by riboswitches • Expanded coverage of chromatin modifications • Updated coverage of promoter and enhancer structures and functions • Updated coverage of the mechanisms of transcription activation and repression

Ch. 16: The Genetics of Cancer • New coverage of the progressive nature of colorectal cancers • Revised and updated coverage of driver and passenger mutations

Ch. 17: Recombinant DNA Technology • Streamlined content on recombinant DNA techniques to deemphasize older techniques and focus on more modern methods • New figure on FISH • Expanded coverage on next-generation and third-generation sequencing • New section on gene-targeting approaches includes content and figures on gene knockout animals and transgenic animals • Revised PDQ content

Ch. 18: Genomics, Bioinformatics, and Proteomics • Updated content on the Human Microbiome Project • New content introducing exome sequencing • Updated content on personal genome projects • Revised and expanded coverage of the Encyclopedia of DNA Elements (ENCODE) Project • New figure on genome sequencing technologies • New Case Study on the microbiome as a risk factor for disease

Ch. 19: Applications and Ethics of Genetic Engineering and Biotechnology • New section on synthetic biology for bioengineering applications • New material and figure on deducing fetal genome sequences from maternal blood • Revised and updated content on prenatal genetic testing • Moved content on GM crops to ST 5 • Moved content on gene therapy to ST 6 • Updated discussion on synthetic genomes • Revised and streamlined content on DNA microarrays given the changing role of microarrays in gene testing (relative to whole-genome, exome, and RNA sequencing) • New content on genetic analysis by sequencing individual

genomes for clinical purposes and single-cell sequencing • Revised ethics section to include additional discussion on the analysis of whole-genome sequences, preconception testing, DNA patents, and destiny predictions • Major revision of end of chapter questions • New GTS essay on the privacy and anonymity of genomic data • New Case Study on genetically modified bacteria for cancer treatment

Ch. 20: Developmental Genetics • New introductory section on the key steps to the differentiated state • New section on the role of binary switch genes and regulatory programs in controlling organ formation, including new figures

Ch. 21: Quantitative Genetics and Multifactorial Traits • New section on limitations of heritability studies • Updated coverage of multifactorial genotypes and expanded coverage of the tomato genome and implications for future improvement in tomato strains • Revised coverage of eQTLS

Ch. 22: Population and Evolutionary Genetics • Revised and updated section on detecting genetic variation and the application of new technology to detect variation in DNA and in genomes • Extensively revised and updated section on the process of speciation • The section on use of phylogenetics to investigate evolutionary history has been improved and expanded with new examples • Information on human evolution has been completely revised and updated with new information about the genomics of extinct human species and their relationship to our species • Five new figures have been added throughout the chapter to accompany the added text

Special Topic 1: Epigenetics • Heavily revised section on imprinting • New ideas on the role of epigenetics in cancer accompany the coverage of the role of somatic mutation in cancer • New section on epigenetic modification of behavior in model organisms and humans

Special Topic 2: Emerging Roles of RNA • New chapter that focuses on the recently discovered functions of RNAs with an emphasis on noncoding RNAs • An introduction to CRISPR/Cas technology in gene editing • Explanation of mechanisms of microRNA and long noncoding RNA gene regulation • Discussion of extracellular RNAs in cell—cell communication and disease diagnosis • Coverage of RNA-induced transcriptional silencing

Special Topic 3: DNA Forensics • New coverage describing how DNA can be inadvertently transferred to a crime scene, leading to false arrests • New coverage of DNA phenotyping

Special Topic 4: Genomics and Personalized Medicine • New coverage on personal genomics and cancer, including a new story of one person's successful experience using "omics" profiling to select a personalized cancer treatment • Updated coverage of personalized medicine and disease diagnostics • Updated coverage of recent studies using "omics" profiles to predict and monitor disease states

Special Topic 5: Genetically Modified Foods • New chapter on genetically modified foods—the genetic technology behind them, the promises, debates, and controversies

Special Topic 6: Gene Therapy • New chapter on the modern aspects of gene therapy • Provides up-to-date applications of gene therapy in humans

Emphasis on Concepts

Essentials of Genetics focuses on conceptual issues in genetics and uses problem solving to develop a deep understanding of them. We consider a concept to be a cognitive unit of meaning that encompasses a related set of scientifically derived findings and ideas. As such, a concept provides broad mental imagery, which we believe is a very effective way to teach science, in this case, genetics. Details that might be memorized, but soon forgotten, are instead subsumed within a conceptual framework that is easily retained. Such a framework may be expanded in content as new information is acquired and may interface with other concepts, providing a useful mechanism to integrate and better understand related processes and ideas. An extensive set of concepts may be devised and conveyed to eventually encompass and represent an entire discipline—and this is our goal in this genetics textbook.

To aid students in identifying the conceptual aspects of a major topic, each chapter begins with a section called ***Chapter Concepts,*** which identifies the most important ideas about to be presented. Then, throughout each chapter, ***Essential Points*** are provided that establish the key issues that have been discussed. And in the ***How Do We Know?*** question that starts each chapter's problem set, students are asked to identify the experimental basis of important genetic findings presented in the chapter. As an extension of the learning approach in biology called "Science as a Way of Knowing," this feature enhances students' understanding of many key concepts covered in each chapter.

Collectively, these features help to ensure that students easily become aware of and understand the major conceptual issues as they confront the extensive vocabulary and the many important details of genetics. Carefully designed figures also support this approach throughout the book.

Emphasis on Problem Solving

Helping students develop effective problem-solving skills is one of the greatest challenges of a genetics course. The feature called ***Now Solve This,*** integrated throughout each chapter, asks students to link conceptual understanding in a more immediate way to problem solving. Each entry provides a problem for the student to solve that is closely related to the current text discussion. A pedagogic hint is

then provided to aid in arriving at the correct solution. All chapters conclude with ***Insights and Solutions,*** a popular and highly useful section that provides sample problems and solutions that demonstrate approaches useful in genetic analysis. These help students develop analytical thinking and experimental reasoning skills. Digesting the information in *Insights and Solutions* primes students as they move on to the lengthier ***Problems and Discussion Questions*** section that concludes each chapter. Here, we present questions that review topics in the chapter and problems that ask students to think in an analytical and applied way about genetic concepts. Problems are of graduated difficulty, with the most demanding near the end of each section. The addition of MasteringGenetics extends our focus on problem solving online, and it allows students to get help and guidance while practicing how to solve problems.

Continuing Features

The Ninth Edition has maintained a number of popular features that are pedagogically useful for students as they study genetics. Collectively, these create a platform that seeks to challenge students to think more deeply about, and thus understand more comprehensively, the information he or she has just finished studying.

- **Exploring Genomics** Appearing in numerous chapters, this feature illustrates the pervasiveness of genomics in the current study of genetics. Each entry asks students to access one or more genomics-related Web sites that collectively are among the best publicly available resources and databases. Students work through interactive exercises that ensure their familiarity with the type of genomic or proteomic information available. Exercises instruct students on how to explore specific topics and how to access significant data. Questions guide student exploration and challenge them to further explore the sites on their own. Importantly, *Exploring Genomics* integrates genomics information throughout the text, as this emerging field is linked to chapter content. This feature provides the basis for individual or group assignments in or out of the classroom.
- **Genetics, Technology, and Society Essays** Appearing in many chapters, this feature provides a synopsis of a topic related to a current finding in genetics that impacts directly on our current society. After each essay, a section entitled "Your Turn" appears in which questions are posed to students along with various resources to help answer them. This innovation provides yet another format to enhance classroom interactions.
- **Case Studies** This feature appears at the end of each chapter and provides the basis for enhanced classroom interactions. In each entry, a short scenario related to one of the chapter topics is presented, followed by several questions. These ask students to apply their newly acquired knowledge to real-life issues that may be explored in small-group discussions or serve as individual assignments.

For the Instructor

MasteringGenetics— http://www.masteringgenetics.com

MasteringGenetics engages and motivates students to learn and allows you to easily assign automatically graded activities. Tutorials provide students with personalized coaching and feedback. Using the gradebook, you can quickly monitor and display student results. MasteringGenetics easily captures data to demonstrate assessment outcomes. Resources include:

- In-depth tutorials that coach students with hints and feedback specific to their misconceptions.
- A new, robust library of **Practice Problems** offers more opportunities to assign challenging problems for student homework or practice. These questions include targeted wrong answer feedback to help students learn from their mistakes. They appear only in MasteringGenetics.
- An item library of assignable questions including end of chapter problems, test bank questions, and reading quizzes. You can use publisher-created prebuilt assignments to get started quickly. Each question can be easily edited to match the precise language you use.
- A gradebook that provides you with quick results and easy-to-interpret insights into student performance.

TestGen EQ Computerized Testing Software

Test questions are available as part of the TestGen EQ Testing Software, a text-specific testing program that is networkable for administering tests. It also allows instructors to view and edit questions, export the questions as tests, and print them out in a variety of formats.

For the Student

MasteringGenetics— http://www.masteringgenetics.com

Used by over a million science students, the Mastering platform is the most effective and widely used online tutorial, homework, and assessment system for the sciences. Perform better on exams with MasteringGenetics. As an instructor-assigned homework system, MasteringGenetics is designed to provide students with a variety of assessments to help them understand key topics and concepts and to build problem-solving skills. MasteringGenetics tutorials guide students through the toughest topics in genetics with self-paced tutorials that provide individualized coaching with hints and feedback specific to a student's individual misconceptions. Students can also explore MasteringGenetics' Study Area, which includes animations, the eText, *Exploring Genomics* exercises, and other study aids. The interactive eText allows students to access their text on mobile devices, highlight text, add study notes, review instructor's notes, and search throughout the text, 24/7.

Acknowledgments

Contributors

We begin with special acknowledgments to those who have made direct contributions to this text. Foremost, we are pleased to thank Dr. Darrell Killian of Colorado College for writing the Special Topic chapter on Emerging Roles of RNA. We much appreciate this important contribution. We also thank Christy Filman of the University of Colorado—Boulder, Jutta Heller of the University of Washington—Tacoma, Christopher Halweg of North Carolina State University, Pamela Osenkowski of Loyola University—Chicago, John Osterman of the University of Nebraska—Lincoln, and Fiona Rawle of the University of Toronto—Mississauga for their work on the media program. Virginia McDonough of Hope College and Cindy Malone of California State University—Northridge contributed greatly to the instructor resources. We also express special thanks to Harry Nickla, recently retired from Creighton University. In his role as author of the *Student Handbook and Solutions Manual* and the test bank, he has reviewed and edited the problems at the end of each chapter and has written many of the new entries as well. He also provided the brief answers to selected problems that appear in the Appendix.

We are grateful to all of these contributors not only for sharing their genetic expertise, but for their dedication to this project as well as the pleasant interactions they provided.

Proofreaders and Accuracy Checking

Reading the detailed manuscript of textbook deserves more thanks than words can offer. Our utmost appreciation is extended to Michelle Gaudette, Tufts University, and Kirkwood Land, University of the Pacific, who provided accuracy checking of many chapters, and to Joanna Dinsmore, who proofread the entire manuscript. They confronted this task with patience and diligence, contributing greatly to the quality of this text.

Reviewers

All comprehensive texts are dependent on the valuable input provided by many reviewers. While we take full responsibility for any errors in this book, we gratefully acknowledge the help provided by those individuals who reviewed the content and pedagogy of this edition:

Soochin Cho, *Creighton University;* Mary Colavito, *Santa Monica College;* Kurt Elliott, *Northwest Vista College;* Edison Fowlks, *Hampton University;* Yvette Gardner, *Clayton State University;* Theresa Geiman, *Loyola University—Maryland;* Christopher Harendza, *Montgomery County Community College;* Lucinda Jack, *University of Maryland;* David Kass, *Eastern Michigan University;* Kirkwood Land, *University of the Pacific;* Te-Wen Lo, *Ithaca College;* Matthew Marcello, *Pace University;* Virginia McDonough, *Hope College;* Amy McMIllan, *SUNY Buffalo State;* Sanghamitra Mohanty, *University of Texas—Austin;* Sudhir Nayak, *The College of New Jersey;* Pamela Osenkowski, *Loyola University—Chicago;* John Osterman, *University of Nebraska—Lincoln;* Pamela Sandstrom, *Unviersity of Nevada—Reno;* Adam Sowalsky, *Northeastern University;* Brian Stout, *Northwest Vista College;* James D. Tucker, *Wayne State University;* Jonathan Visick, *North Central College;* Fang-Sheng Wu, *Virginia Commonwealth University;* Lev Yampolsky, *East Tennessee State University*

Special thanks go to Mike Guidry of LightCone Interactive and Karen Hughes of the University of Tennessee for their original contributions to the media program.

As these acknowledgments make clear, a text such as this is a collective enterprise. All of the above individuals deserve to share in any success this text enjoys. We want them to know that our gratitude is equaled only by the extreme dedication evident in their efforts. Many, many thanks to them all.

Editorial and Production Input

At Pearson, we express appreciation and high praise for the editorial guidance of Michael Gillespie, whose ideas and efforts have helped to shape and refine the features of this edition of the text. Dusty Friedman, our Project Editor, has worked tirelessly to keep the project on schedule and to maintain our standards of high quality. In addition, our editorial team—Ginnie Simione-Jutson, Executive Director of Development, Chloé Veylit, Media Producer, and Tania Mlawer, Director of Editorial Content for MasteringGenetics—have provided valuable input into the current edition. They have worked creatively to ensure that the pedagogy and design of the book and media package are at the cutting edge of a rapidly changing discipline. Sudhir Nayak of The College of New Jersey provided outstanding work for the MasteringGenetics program and his input regarding genomics is much appreciated. Margaret Young and Rose Kernan supervised all of the production intricacies with great attention to detail and perseverance. Outstanding copyediting was performed by Betty Pessagno, for which we are most grateful. Lauren Harp has professionally and enthusiastically managed the marketing of the text. Finally, the beauty and consistent presentation of the art work are the product of Imagineering of Toronto. Without the work ethic and dedication of the above individuals, the text would never have come to fruition.

The publishers would like to thank the following for their contribution to the Global Edition:

Contributors

Sridev Mohapatra, BITS Pilani
Elizabeth R. Martin, D.Phil.

Reviewers

Francisco Ramos Morales, University of Seville
Adriaan Engelbrecht, University of the Western Cape
Shefali Sabharanjak, Ph.D.

This page intentionally left blank

1 Introduction to Genetics

CHAPTER CONCEPTS

- Genetics in the twenty-first century is built on a rich tradition of discovery and experimentation stretching from the ancient world through the nineteenth century to the present day.
- Transmission genetics is the process by which traits controlled by genes are transmitted through gametes from generation to generation.
- Mutant strains can be used in genetic crosses to map the location and distance between genes on chromosomes.
- The Watson–Crick model of DNA structure explains how genetic information is stored and expressed. This discovery is the foundation of molecular genetics.
- Recombinant DNA technology revolutionized genetics, was the foundation for the Human Genome Project, and has generated new fields that combine genetics with information technology.
- Biotechnology provides genetically modified organisms and their products that are used across a wide range of fields including agriculture, medicine, and industry.
- Model organisms used in genetics research are now utilized in combination with recombinant DNA technology and genomics to study human diseases.
- Genetic technology is developing faster than the policies, laws, and conventions that govern its use.

Newer model organisms in genetics include the roundworm *Caenorhabditis elegans*, the zebrafish, *Danio rerio*, and the mustard plant *Arabidopsis thaliana*.

Information from the Human Genome Project and other areas of genetics is now having far-reaching effects on our daily lives. For example, researchers and clinicians are using genomic information to improve the quality of medical care via **translational medicine,** a process in which genetic findings are directly "translated" into new and improved methods of diagnosis and treatment. One important area of focus is cardiovascular disease, which is the leading cause of death worldwide. One of the key risk factors for development of this condition is the presence of elevated blood levels of "bad" cholesterol (low-density lipoprotein cholesterol, or LDL cholesterol). Although statin drugs are effective in lowering the blood levels of LDL cholesterol and reducing the risk of heart disease, up to 50 percent of treated individuals remain at risk, and serious side-effects prevent many others from using these drugs.

To gain a share of the estimated $25 billion market for treatment of elevated LDL levels, major pharmaceutical firms are developing a new generation of more effective cholesterol-lowering drugs. However, bringing a new drug to market is risky. Costs can run over $1 billion, and many drugs (up to 1 in 3) fail clinical trials and are withdrawn. In the search for a new strategy in drug development, human genetics is now playing an increasingly vital role. Blood levels of LDL in a population vary over a threefold range, and about 50 percent of this variation is genetic. Although many genes are involved, the role of one gene, *PCSK9,* in controlling LDL levels is an outstanding example of how a genetic approach has been successful in identifying drug targets and improving the chance that a new drug will be successful. The rapid transfer of basic

research on *PCSK9* to drug development and its use in treating patients is a pioneering example of translational medicine.

Soon after the *PCSK9* gene was identified, several mutant forms of this gene were found to be associated with extremely high levels of LDL cholesterol, resulting in a condition called familial hypercholesterolemia (FH). When this work came to the attention of researchers in Texas, they wondered whether other mutations in *PCSK9* might have the opposite effect and drastically lower LDL cholesterol levels. To test this idea, they turned to data from the Dallas Heart Study, which collected detailed clinical information, including LDL levels and DNA samples, from 3500 individuals. DNA sequencing of the *PSCK9* gene from participants with extremely low LDL levels identified two mutations that reduced blood levels of LDL by 40 percent. Other work showed that carriers of these mutations had an 88 percent lower risk of heart disease.

The PCSK9 protein binds to LDL receptors on liver cells, moving the receptors into the cell where they are broken down. However, if the PCSK9 protein does not bind to an LDL receptor, the receptor is returned to the cell surface where it can remove more LDL from the bloodstream. Carriers of either of the two mutations have much lower PCSK9 protein levels. As a result, liver cells in these individuals have many more LDL receptors, which, in turn, remove more LDL from the blood. Using this information, several pharmaceutical firms have developed antibody-based drugs that bind to the PCKS9 protein and prevent its interaction with LDL receptors, which, in turn, lowers LDL cholesterol levels. Successful clinical trials show that LDL blood levels can be reduced by up to 70 percent in the test population, and one of these drugs has been shown to reduce heart attacks and strokes by 50 percent. Ongoing clinical trials are drawing to a close, and it is expected that these drugs will soon be available to treat elevated cholesterol levels.

The example of the *PCSK9* gene clearly demonstrates that coupling genetic research with drug development will play a critical and exciting role in speeding the movement of research findings into medical practice.

This introductory chapter provides an overview of genetics and a survey of the high points in its history and gives a preliminary description of its central principles and emerging developments. All the topics discussed in this chapter will be explored in far greater detail elsewhere in the book. This text will enable you to achieve a thorough understanding of modern-day genetics and its underlying principles. Along the way, enjoy your studies, but take your responsibilities as a novice geneticist very seriously.

1.1 Genetics Has a Rich and Interesting History

We don't know when people first recognized the hereditary nature of certain traits, but archaeological evidence (e.g., pictorial representations, preserved bones and skulls, and dried seeds) documents the successful domestication of animals and the cultivation of plants thousands of years ago by the artificial selection of genetic variants from wild populations. Between 8000 and 1000 B.C., horses, camels, oxen, and wolves were domesticated, and selective breeding of these species soon followed. Cultivation of many plants, including maize, wheat, rice, and the date palm, began around 5000 B.C. Such evidence documents our ancestors' successful attempts to manipulate the genetic composition of species.

During the Golden Age of Greek culture, the writings of the Hippocratic School of Medicine (500–400 B.C.) and of the philosopher and naturalist Aristotle (384–322 B.C.) discussed heredity as it relates to humans. The Hippocratic treatise *On the Seed* argued that active "humors" in various parts of the body served as the bearers of hereditary traits. Drawn from various parts of the male body to the semen and passed on to offspring, these humors could be healthy or diseased, with the diseased humors accounting for the appearance of newborns with congenital disorders or deformities. It was also believed that these humors could be altered in individuals before they were passed on to offspring, explaining how newborns could "inherit" traits that their parents had "acquired" in response to their environment.

Aristotle extended Hippocrates' thinking and proposed that the male semen contained a "vital heat" with the capacity to produce offspring of the same "form" (i.e., basic structure and capacities) as the parent. Aristotle believed that this heat cooked and shaped the menstrual blood produced by the female, which was the "physical substance" that gave rise to an offspring. The embryo developed not because it already contained the parts of an adult in miniature form (as some Hippocratics had thought) but because of the shaping power of the vital heat. Although the ideas of Hippocrates and Aristotle sound primitive and naive today, we should recall that prior to the 1800s neither sperm nor eggs had been observed in mammals.

1600–1850: The Dawn of Modern Biology

Between about 300 B.C. and A.D. 1600, there were few significant new ideas about genetics. However, between 1600 and 1850, major strides provided insight into the biological basis of life. In the 1600s, William Harvey proposed the theory of **epigenesis,** which states that an organism develops from the fertilized embryo by a succession of developmental events that eventually transform the embryo into an adult. The theory of epigenesis directly conflicted with the theory of **preformation,** which stated that the sperm or the fertilized egg contains a complete miniature adult, called a **homunculus** (Figure 1–1). Around 1830, Matthias Schleiden and Theodor Schwann proposed the **cell theory,** stating that all organisms are composed of basic structural units called cells,


© 1964 National Library of Medicine

FIGURE 1–1 Depiction of the homunculus, a sperm containing a miniature adult, perfect in proportion and fully formed.
(Hartsoeker, N. Essay de dioptrique Paris, 1694, p. 246. National Library of Medicine)

which are derived from preexisting cells. The idea of **spontaneous generation,** the creation of living organisms from nonliving components, was disproved by Louis Pasteur later in the century, and living organisms were then considered to be derived from preexisting organisms and to consist of cells.

In the mid-1800s the revolutionary work of Charles Darwin and Gregor Mendel set the stage for the rapid development of genetics in the twentieth and twenty-first centuries.

Charles Darwin and Evolution

With this background, we turn to a brief discussion of the work of Charles Darwin, who published *The Origin of Species* in 1859, describing his ideas about evolution. Darwin's geological, geographical, and biological observations convinced him that existing species arose by descent with modification from ancestral species. Greatly influenced by his voyage on the HMS *Beagle* (1831–1836), Darwin's thinking led him to formulate the theory of **natural selection,** which presented an explanation of the mechanism of evolutionary change. Formulated and proposed independently by Alfred Russel Wallace, natural selection is based on the observation that populations tend to contain more offspring than the environment can support, leading to a struggle for survival among individuals. Those individuals with heritable traits that allow them to adapt to their environment are better able to survive and reproduce than those with less adaptive traits. Over a long period of time, advantageous variations, even very slight ones, will accumulate. If a population carrying these inherited variations becomes reproductively isolated, a new species may result.

Darwin, however, lacked an understanding of the genetic basis of variation and inheritance, a gap that left his theory open to reasonable criticism well into the twentieth century. Shortly after Darwin published his book, Gregor Johann Mendel published a paper in 1866 showing how traits were passed from generation to generation in pea plants and offering a general model of how traits are inherited. His research was little known until it was partially duplicated and brought to light by Carl Correns, Hugo de Vries, and Erich Tschermak around 1900.

By the early part of the twentieth century, it became clear that heredity and development were dependent on genetic information residing in genes contained in chromosomes, which were then contributed to each individual by gametes—the so-called **chromosomal theory of inheritance.** The gap in Darwin's theory was closed, and Mendel's research has continued to serve as the foundation of genetics.

1.2 Genetics Progressed from Mendel to DNA in Less Than a Century

Because genetic processes are fundamental to life itself, the science of genetics unifies biology and serves as its core. The starting point for this branch of science was a monastery garden in central Europe in the late 1850s.

Mendel's Work on Transmission of Traits

Gregor Mendel, an Augustinian monk, conducted a decade-long series of experiments using pea plants. He applied quantitative data analysis to his results and showed that traits are passed from parents to offspring in predictable ways. He further concluded that each trait in the plant is controlled by a pair of factors (which we now call genes) and that during gamete formation (the formation of egg cells and sperm), members of a gene pair separate from each other. His work was published in 1866 but was largely unknown until it was cited in papers published by others around 1900. Once confirmed, Mendel's findings became recognized as explaining the transmission of traits in pea plants and all other higher organisms. His work forms the foundation for **genetics,** which is defined as the branch of biology concerned with the study of heredity and variation. Mendelian genetics will be discussed later in the text (see Chapters 3 and 4).

ESSENTIAL POINT

Mendel's work on pea plants established the principles of gene transmission from parent to offspring that serve as the foundation for the science of genetics. ■

The Chromosome Theory of Inheritance: Uniting Mendel and Meiosis

Mendel did his experiments before the structure and role of chromosomes were known. About 20 years after his work was published, advances in microscopy allowed researchers to identify chromosomes and establish that, in most eukaryotes, members of each species have a characteristic number of chromosomes called the **diploid number (2*n*)** in most of their cells. For example, humans have a diploid number of 46 (Figure 1–2). Chromosomes in diploid cells exist in pairs, called **homologous chromosomes.**

Researchers in the last decades of the nineteenth century also described chromosome behavior during two forms of cell division, **mitosis** and **meiosis.** In mitosis, chromosomes are copied and distributed so that each daughter cell receives a diploid set of chromosomes identical to those in the parental cell. Meiosis is associated with gamete formation. Cells produced by meiosis receive only one chromosome from each chromosome pair, and the resulting number of chromosomes is called the **haploid (*n*) number.** This reduction in chromosome number is essential if the offspring arising from the fusion of egg and sperm are to maintain the constant number of chromosomes characteristic of their parents and other members of their species.

Early in the twentieth century, Walter Sutton and Theodor Boveri independently noted that the behavior of chromosomes during meiosis is identical to the behavior of genes during gamete formation described by Mendel. For example, genes and chromosomes exist in pairs, and members of a gene pair and members of a chromosome pair separate from each other during gamete formation. Based on these parallels, Sutton and Boveri each proposed that genes are carried on chromosomes. They independently formulated the chromosome theory of inheritance, which states that inherited traits are controlled by genes residing on chromosomes faithfully transmitted through gametes, maintaining genetic continuity from generation to generation.

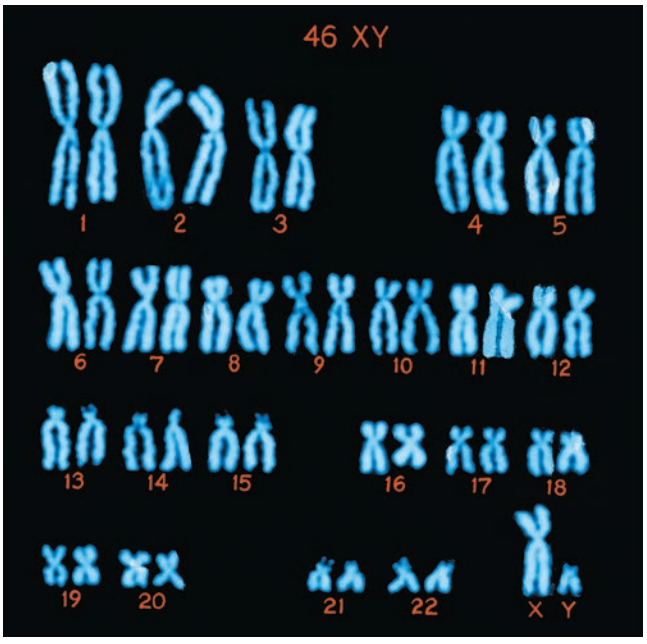


FIGURE 1–2 A colorized image of the human male chromosome set. Arranged in this way, the set is called a karyotype.

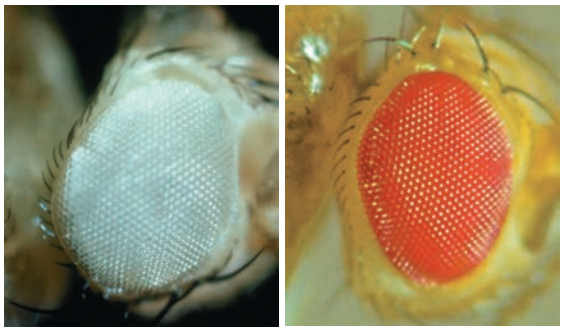


FIGURE 1–3 The white-eyed mutation in *D. melanogaster* (left) and the normal red eye color (right).

Genetic Variation

About the same time that the chromosome theory of inheritance was proposed, scientists began studying the inheritance of traits in the fruit fly, *Drosophila melanogaster*. Early in this work, a white-eyed fly (Figure 1–3) was discovered among normal (wild-type) red-eyed flies. This variant was produced by a **mutation** in one of the genes controlling eye color. Mutations are defined as any heritable change in the DNA sequence and are the source of all genetic variation.

ESSENTIAL POINT

The chromosome theory of inheritance explains how genetic information is transmitted from generation to generation. ■

The white-eye variant discovered in *Drosophila* is an **allele** of a gene controlling eye color. Alleles are defined as alternative forms of a gene. Different alleles may produce differences in the observable features, or **phenotype,** of an organism. The set of alleles for a given trait carried by an organism is called the **genotype.** Using mutant genes as markers, geneticists can map the location of genes on chromosomes.

The Search for the Chemical Nature of Genes: DNA or Protein?

Work on white-eyed *Drosophila* showed that the mutant trait could be traced to a single chromosome, confirming the idea that genes are carried on chromosomes. Once this relationship was established, investigators turned their attention to identifying which chemical component of chromosomes carries genetic information. By the 1920s, scientists knew that proteins and DNA were the major chemical

components of chromosomes. There are a large number of different proteins, and because of their universal distribution in the nucleus and cytoplasm, many researchers thought proteins were the carriers of genetic information.

In 1944, Oswald Avery, Colin MacLeod, and Maclyn McCarty, researchers at the Rockefeller Institute in New York, published experiments showing that DNA was the carrier of genetic information in bacteria. This evidence, though clear-cut, failed to convince many influential scientists. Additional evidence for the role of DNA as a carrier of genetic information came from other researchers who worked with viruses. This evidence that DNA carries genetic information, along with other research over the next few years, provided solid proof that DNA, not protein, is the genetic material, setting the stage for work to establish the structure of DNA.

1.3 Discovery of the Double Helix Launched the Era of Molecular Genetics

Once it was accepted that DNA carries genetic information, efforts were focused on deciphering the structure of the DNA molecule and the mechanism by which information stored in it produces a phenotype.

The Structure of DNA and RNA

One of the great discoveries of the twentieth century was made in 1953 by James Watson and Francis Crick, who described the structure of DNA. DNA is a long, ladder-like macromolecule that twists to form a double helix (**Figure 1–4**). Each linear strand of the helix is made up of subunits called **nucleotides.** In DNA, there are four different nucleotides, each of which contains a nitrogenous base, abbreviated A (adenine), G (guanine), T (thymine), or C (cytosine). These four bases, in various sequence combinations, ultimately encode genetic information. The two strands of DNA are exact complements of one another, so that the rungs of the ladder in the double helix always consist of A═T and G═C base pairs. Along with Maurice Wilkins, Watson and Crick were awarded a Nobel Prize in 1962 for their work on the structure of DNA. We will discuss the structure of DNA later in the text (see Chapter 9).

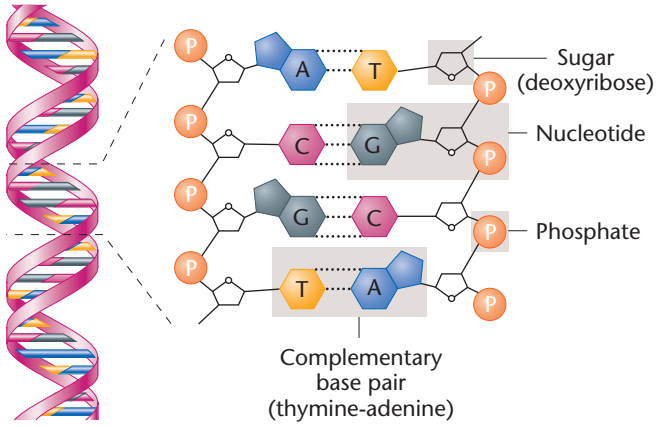


FIGURE 1–4 Summary of the structure of DNA, illustrating the arrangement of the double helix (on the left) and the chemical components making up each strand (on the right). The dotted lines on the right represent weak chemical bonds, called hydrogen bonds, which hold together the two strands of the DNA helix.

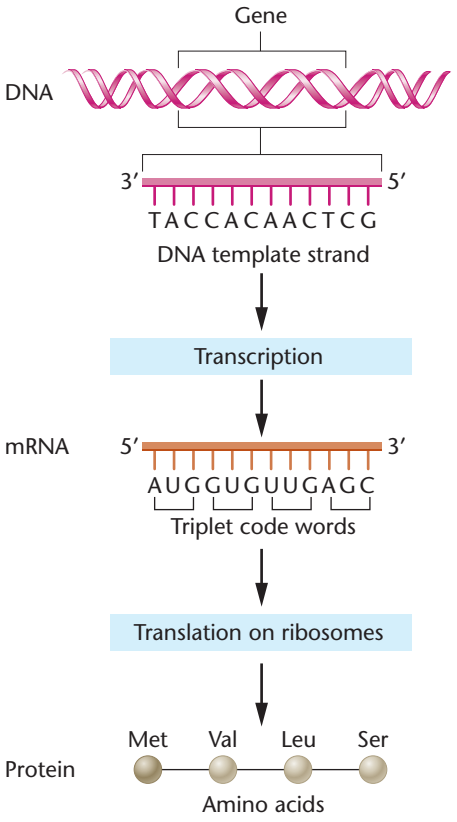


FIGURE 1–5 Gene expression consists of transcription of DNA into mRNA (top) and the translation (center) of mRNA (with the help of a ribosome) into a protein (bottom).

Another nucleic acid, RNA, is chemically similar to DNA but contains a different sugar (ribose rather than deoxyribose) in its nucleotides and contains the nitrogenous base uracil in place of thymine. RNA, however, is generally a single-stranded molecule.

Gene Expression: From DNA to Phenotype

The genetic information encoded in the order of nucleotides in DNA is expressed in a series of steps that results in the formation of a functional gene product. In the majority of cases, this product is a protein. In eukaryotic cells, the process leading to protein production begins in the nucleus with **transcription,** a process in which the nucleotide sequence in one strand of DNA is used to construct a complementary RNA sequence (top part of **Figure 1–5**). Once an RNA

molecule is produced, it moves to the cytoplasm, where the RNA—called **messenger RNA,** or **mRNA** for short—binds to **ribosomes.** The synthesis of proteins under the direction of mRNA is called **translation** (center part of Figure 1–5). The information encoded in mRNA (called the **genetic code**) consists of a linear series of nucleotide triplets. Each triplet, called a **codon,** is complementary to the information stored in DNA and specifies the insertion of a specific amino acid into a protein. Proteins (lower part of Figure 1–5) are polymers made up of amino acid monomers. There are 20 different amino acids commonly found in proteins.

Protein assembly is accomplished with the aid of adapter molecules called **transfer RNA (tRNA).** Within the ribosome, tRNAs recognize the information encoded in the mRNA codons and carry the proper amino acids for construction of the protein during translation.

We now know that gene expression can be more complex than outlined here. Some of these complexities will be discussed later in the text (see Chapters 13, 15, and Special Topic Chapter 1—Epigenetics).

Proteins and Biological Function

In most cases, proteins are the end products of gene expression. The diversity of proteins and the biological functions they perform—the diversity of life itself—arises from the fact that proteins are made from combinations of 20 different amino acids. Consider that a protein chain containing 100 amino acids can have at each position any one of 20 amino acids; the number of possible different 100 amino acid proteins, each with a unique sequence, is therefore equal to

$$20^{100}$$

Obviously, proteins are molecules with the potential for enormous structural diversity and serve as the mainstay of biological systems.

Enzymes form the largest category of proteins. These molecules serve as biological catalysts, lowering the energy of activation in reactions and allowing cellular metabolism to proceed at body temperature.

Proteins other than enzymes are critical components of cells and organisms. These include hemoglobin, the oxygen-binding molecule in red blood cells; insulin, a pancreatic hormone; collagen, a connective tissue molecule; and actin and myosin, the contractile muscle proteins. A protein's shape and chemical behavior are determined by its linear sequence of amino acids, which in turn are dictated by the stored information in the DNA of a gene that is transferred to RNA, which then directs the protein's synthesis.

Linking Genotype to Phenotype: Sickle-Cell Anemia

Once a protein is made, its biochemical or structural properties play a role in producing a phenotype. When mutation alters a gene, it may modify or even eliminate the encoded protein's usual function and cause an altered phenotype. To trace this chain of events, we will examine sickle-cell anemia, a human genetic disorder.

Sickle-cell anemia is caused by a mutant form of hemoglobin, the protein that transports oxygen from the lungs to cells in the body. Hemoglobin is a composite molecule made up of two different proteins, α-globin and β-globin, each encoded by a different gene. In sickle-cell anemia, a mutation in the gene encoding β-globin causes an amino acid substitution in 1 of the 146 amino acids in the protein. **Figure 1–6** shows the template DNA sequence, the corresponding mRNA codons, and the amino acids occupying positions 4–7 for the normal and mutant forms of β-globin. Notice that the mutation in sickle-cell anemia consists of a change in one DNA nucleotide, which leads to a change in codon 6 in mRNA from GAG to GUG, which in turn changes amino acid number 6 in β-globin from glutamic acid to valine. The other 145 amino acids in the protein are not changed by this mutation.

Individuals with two mutant copies of the β-globin gene have sickle-cell anemia. Their mutant β-globin proteins cause hemoglobin molecules in red blood cells to polymerize when the blood's oxygen concentration is low, forming long chains of hemoglobin that distort the shape of red blood cells (**Figure 1–7**). The deformed cells are fragile and break easily, reducing the number of red blood cells in circulation (anemia is an insufficiency of red blood cells). Sickle-shaped blood cells block blood flow in capillaries and small blood vessels, causing severe pain and damage to the heart, brain, muscles, and kidneys. All the symptoms of this disorder are caused by a change in a single nucleotide in a gene that changes one amino acid out of 146 in the β-globin molecule, demonstrating the close relationship between genotype and phenotype.

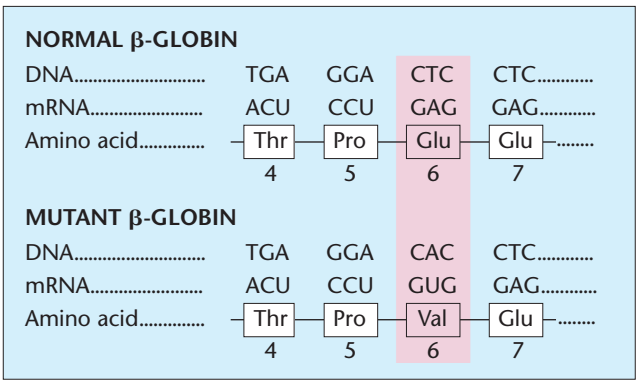


FIGURE 1–6 A single-nucleotide change in the DNA encoding β-globin (CTC→CAC) leads to an altered mRNA codon (GAG→GUG) and the insertion of a different amino acid (Glu→Val), producing the altered version of the β-globin protein that is responsible for sickle-cell anemia.

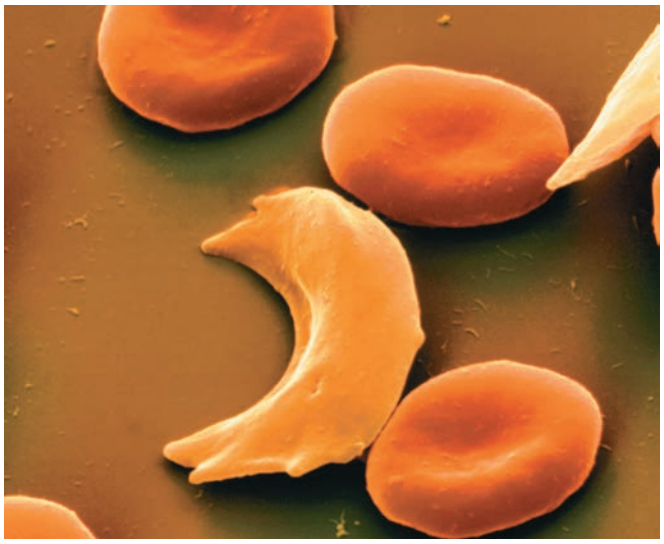


FIGURE 1–7 Normal red blood cells (round) and sickled red blood cells. The sickled cells block capillaries and small blood vessels.

ESSENTIAL POINT

The central dogma of molecular biology—that DNA is a template for making RNA, which in turn directs the synthesis of proteins—explains how genes control phenotypes. ■

1.4 Development of Recombinant DNA Technology Began the Era of DNA Cloning

The era of recombinant DNA began in the early 1970s, when researchers discovered that bacterial proteins called **restriction endonucleases,** which cut the DNA of invading viruses, could also be used to cut any organism's DNA at specific nucleotide sequences, producing a reproducible set of fragments.

Soon after, researchers discovered ways to insert the DNA fragments produced by the action of restriction enzymes into carrier DNA molecules called **vectors** to form **recombinant DNA** molecules. When transferred into bacterial cells, thousands of copies, or **clones,** of the combined vector and DNA fragments are produced during bacterial reproduction. Large amounts of cloned DNA fragments can be isolated from these bacterial host cells. These DNA fragments can be used to isolate genes, to study their organization and expression, and to study their nucleotide sequence and evolution.

Collections of clones that represent an organism's **genome,** defined as the complete haploid DNA content of a specific organism, are called genomic libraries. Genomic libraries are now available for hundreds of species.

Recombinant DNA technology has not only accelerated the pace of research but also given rise to the biotechnology industry, which has grown to become a major contributor to the U.S. economy.

1.5 The Impact of Biotechnology Is Continually Expanding

The use of recombinant DNA technology and other molecular techniques to make products is called **biotechnology.** In the United States, biotechnology has quietly revolutionized many aspects of everyday life; products made by biotechnology are now found in the supermarket, in health care, in agriculture, and in the court system. A later chapter (see Chapter 19) contains a detailed discussion of biotechnology, but for now, let's look at some everyday examples of biotechnology's impact.

Plants, Animals, and the Food Supply

The use of recombinant DNA technology to genetically modify crop plants has revolutionized agriculture. Genes for traits including resistance to herbicides, insects, and genes for nutritional enhancement have been introduced into crop plants. The transfer of heritable traits across species using recombinant DNA technology creates **transgenic organisms.** Herbicide-resistant corn and soybeans were first planted in the mid-1990s, and transgenic strains now represent about 88 percent of the U.S. corn crop and 93 percent of the U.S. soybean crop. It is estimated that more than 70 percent of the processed food in the United States contains ingredients from transgenic crops.

We will discuss the most recent findings involving genetically modified organisms later in the text (Special Topic Chapter 5—Genetically Modified Organisms).

New methods of cloning livestock such as sheep and cattle have also changed the way we use these animals. In 1996, Dolly the sheep (**Figure 1–8**) was cloned by nuclear transfer,

FIGURE 1–8 Dolly, a Finn Dorset sheep cloned from the genetic material of an adult mammary cell, shown next to her first-born lamb, Bonnie.

a method in which the nucleus of an adult cell is transferred into an egg that has had its nucleus removed. This method makes it possible to produce dozens or hundreds of genetically identical offspring with desirable traits and has many applications in agriculture, sports, and medicine.

Biotechnology has also changed the way human proteins for medical use are produced. Through use of gene transfer, transgenic animals now synthesize these therapeutic proteins. In 2009, an anticlotting protein derived from the milk of transgenic goats was approved by the U.S. Food and Drug Administration for use in the United States. Other human proteins from transgenic animals are now being used in clinical trials to treat several diseases. The biotechnology revolution will continue to expand as new methods are developed to make an increasing array of products.

Biotechnology in Genetics and Medicine

More than 10 million children or adults in the United States suffer from some form of genetic disorder, and every childbearing couple faces an approximately 3 percent risk of having a child with a genetic anomaly. The molecular basis for hundreds of genetic disorders is now known, and many of these genes have been mapped, isolated, and cloned. Biotechnology-derived whole-genome testing is now available to perform prenatal diagnosis of most if not all heritable disorders and to test parents for their status as "carriers" of inherited disorders. However, the use of genetic testing and related technologies raises ethical concerns that have yet to be fully resolved.

ESSENTIAL POINT

Biotechnology has revolutionized agriculture and the pharmaceutical industry, while genetic testing has had a profound impact on the diagnosis of genetic diseases. ■

1.6 Genomics, Proteomics, and Bioinformatics Are New and Expanding Fields

The use of recombinant DNA technology to create genomic libraries prompted scientists to consider sequencing all the clones in a library to derive the nucleotide sequence of an organism's genome. This sequence information would be used to identify each gene in the genome and establish its function.

One such project, the Human Genome Project, began in 1990 as an international effort to sequence the human genome. By 2003, the publicly funded Human Genome Project and a private, industry-funded genome project completed sequencing of the gene-containing portion of the genome.

As more genome sequences were acquired, several new biological disciplines arose. One, called **genomics** (the study of genomes), studies the structure, function, and evolution of genes and genomes. A second field, **proteomics,** identifies the set of proteins present in a cell under a given set of conditions, and studies their functions and interactions. To store, retrieve, and analyze the massive amount of data generated by genomics and proteomics, a specialized subfield of information technology called **bioinformatics** was created to develop hardware and software for processing and storing nucleotide and protein data.

Geneticists and other biologists now use information in databases containing nucleic acid sequences, protein sequences, and gene-interaction networks to answer experimental questions in a matter of minutes instead of months and years. A feature called "Exploring Genomics," located at the end of many of the chapters in this textbook, gives you the opportunity to explore these databases for yourself while completing an interactive genetics exercise.

Modern Approaches to Understanding Gene Function

Historically, a method known as **classical** or **forward genetics** was used to study and understand gene function. In this approach geneticists relied on the use of naturally occurring mutations, or intentionally induced mutations (using chemicals, X-rays, or UV light as examples) to cause altered phenotypes in model organisms, and then worked through the lab-intensive and time-consuming process of identifying the genes that caused these new phenotypes. Such characterization often led to the identification of a gene or genes of interest, and once the technology advanced, the gene sequence could be determined.

Classical genetics approaches are still used, but as genome sequencing has become routine, molecular approaches to understanding gene function have changed considerably. These modern approaches are what we will highlight in this section.

For the past two decades or so, geneticists have relied on the use of molecular techniques in an approach referred to as **reverse genetics.** In reverse genetics, the DNA sequence for a particular gene of interest is known, but the role and function of the gene are typically not well understood. For example, molecular biology techniques such as **gene knockout** render targeted genes nonfunctional in model organisms or in cultured cells, allowing scientists to investigate the fundamental question of "what happens if this gene is disrupted?" After creating a knockout, scientists look for changes in phenotype, as well as alterations at the cellular and molecular level. The ultimate goal is to determine the function of the gene being studied.

ESSENTIAL POINT

Recombinant DNA technology gave rise to several new fields, including genomics, proteomics, and bioinformatics, which allow scientists to explore the structure and evolution of genomes and the proteins they encode. ■

1.7 Genetic Studies Rely on the Use of Model Organisms

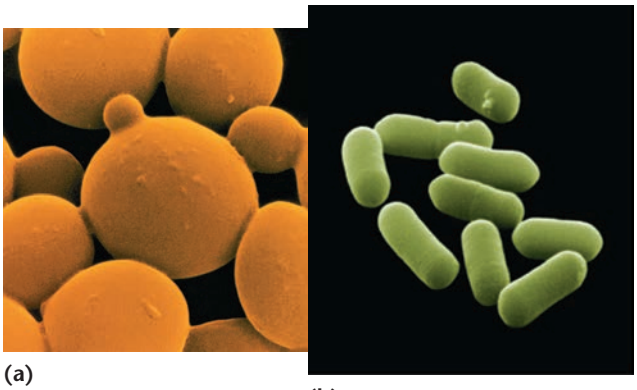
After the rediscovery of Mendel's work in 1900, research using a wide range of organisms confirmed that the principles of inheritance he described were of universal significance among plants and animals. Geneticists gradually came to focus attention on a small number of organisms, including the fruit fly (*Drosophila melanogaster*) and the mouse (*Mus musculus*) (**Figure 1–9**). This trend developed for two main reasons: first, it was clear that genetic mechanisms were the same in most organisms, and second, some organisms had characteristics that made them especially suitable for genetic research. They were easy to grow, had relatively short life cycles, produced many offspring, and their genetic analysis was fairly straightforward. Over time, researchers created a large catalog of mutant strains for these species, and the mutations were carefully studied, characterized, and mapped. Because of their well-characterized genetics, these species became **model genetic organisms,** defined as organisms used for the study of basic biological processes. In later chapters, we will see how discoveries in model organisms are shedding light on many aspects of biology, including aging, cancer, the immune system, and behavior.

The Modern Set of Genetic Model Organisms

Gradually, geneticists added other species to their collection of model organisms: viruses (such as the T phages and lambda phage) and microorganisms (the bacterium *Escherichia coli* and the yeast *Saccharomyces cerevisiae*) (**Figure 1–10**).

(a) (b)

FIGURE 1–9 The first generation of model organisms in genetic analysis included (a) the mouse, *Mus musculus,* and (b) the fruit fly, *Drosophila melanogaster*.



(a) (b)

FIGURE 1–10 Microbes that have become model organisms for genetic studies include (a) the yeast *Saccharomyces cerevisiae* and (b) the bacterium *Escherichia coli*.

More recently, additional species have been developed as model organisms, three of which are shown in the chapter opening photograph. Each species was chosen to allow study of some aspect of embryonic development. The nematode *Caenorhabditis elegans* was chosen as a model system to study the development and function of the nervous system because its nervous system contains only a few hundred cells and the developmental fate of these and all other cells in the body has been mapped out. *Arabidopsis thaliana,* a small plant with a short life cycle, has become a model organism for the study of many aspects of plant biology. The zebrafish, *Danio rerio,* is used to study vertebrate development: it is small, it reproduces rapidly, and its egg, embryo, and larvae are all transparent.

Model Organisms and Human Diseases

The development of recombinant DNA technology and the data from genome sequencing have confirmed that all life has a common origin. Because of this, genes with similar functions in different organisms tend to be similar or identical in structure and nucleotide sequence. Much of what scientists learn by studying the genetics of model organisms can therefore be applied to humans as a way of understanding and treating human diseases. In addition, the ability to create transgenic organisms by transferring genes between species has enabled scientists to develop models of human diseases in organisms ranging from bacteria to fungi, plants, and animals (**Table 1.1**).

The idea of studying a human disease such as colon cancer by using *E. coli* may strike you as strange, but the basic steps of DNA repair (a process that is defective in some forms of colon cancer) are the same in both organisms, and a gene involved (*mutL* in *E. coli* and *MLH1* in humans) is found in both organisms. More importantly, *E. coli* has the advantage of being easier to grow (the cells divide every 20 minutes), and researchers can easily create and study new mutations in the bacterial *mutL* gene in

TABLE 1.1 Model Organisms Used to Study Some Human Diseases

Organism	Human Diseases
E. coli	Colon cancer and other cancers
S. cerevisiae	Cancer, Werner syndrome
D. melanogaster	Disorders of the nervous system, cancer
C. elegans	Diabetes
D. rerio	Cardiovascular disease
M. musculus	Lesch–Nyhan disease, cystic fibrosis, fragile-X syndrome, and many other diseases

order to figure out how it works. This knowledge may eventually lead to the development of drugs and other therapies to treat colon cancer in humans.

The fruit fly, *Drosophila melanogaster,* is also being used to study a number of human diseases. Mutant genes have been identified in *D. melanogaster* that produce phenotypes with structural abnormalities of the nervous system and adult-onset degeneration of the nervous system. The information from genome-sequencing projects indicates that almost all these genes have human counterparts. For example, genes involved in a human disease of the retina called retinitis pigmentosa are identical to *Drosophila* genes involved in retinal degeneration. Study of these mutations in *Drosophila* is helping to dissect this disorder and to identify the function of the genes involved.

Another approach to studying diseases of the human nervous system is to transfer mutant human disease genes into *Drosophila* using recombinant DNA technology. The transgenic flies are then used for studying the mutant human genes themselves, other genes that affect the expression of the human disease genes, and the effects of therapeutic drugs on the action of those genes—all studies that are difficult or impossible to perform in humans. This gene transfer approach is being used to study almost a dozen human neurodegenerative disorders, including Huntington disease, Machado–Joseph disease, myotonic dystrophy, and Alzheimer disease.

Throughout the following chapters, you will encounter these model organisms again and again. Remember each time you meet them that they not only have a rich history in basic genetics research but are also at the forefront in the study of human genetic disorders and infectious diseases. As discussed in the next section, however, we have yet to reach a consensus on how and when some of this technology will be accepted as safe and ethically acceptable.

ESSENTIAL POINT

The study of model organisms for understanding human health and disease is one of many ways genetics and biotechnology are rapidly changing everyday life. ■

1.8 We Live in the Age of Genetics

Mendel described his decade-long project on inheritance in pea plants in an 1865 paper presented at a meeting of the Natural History Society of Brünn in Moravia. Less than 100 years later, the 1962 Nobel Prize was awarded to James Watson, Francis Crick, and Maurice Wilkins for their work on the structure of DNA. This time span encompassed the years leading up to the acceptance of Mendel's work, the discovery that genes are on chromosomes, the experiments that proved DNA encodes genetic information, and the elucidation of the molecular basis for DNA replication. The rapid development of genetics from Mendel's monastery garden to the Human Genome Project and beyond is summarized in a timeline in **Figure 1–11**.

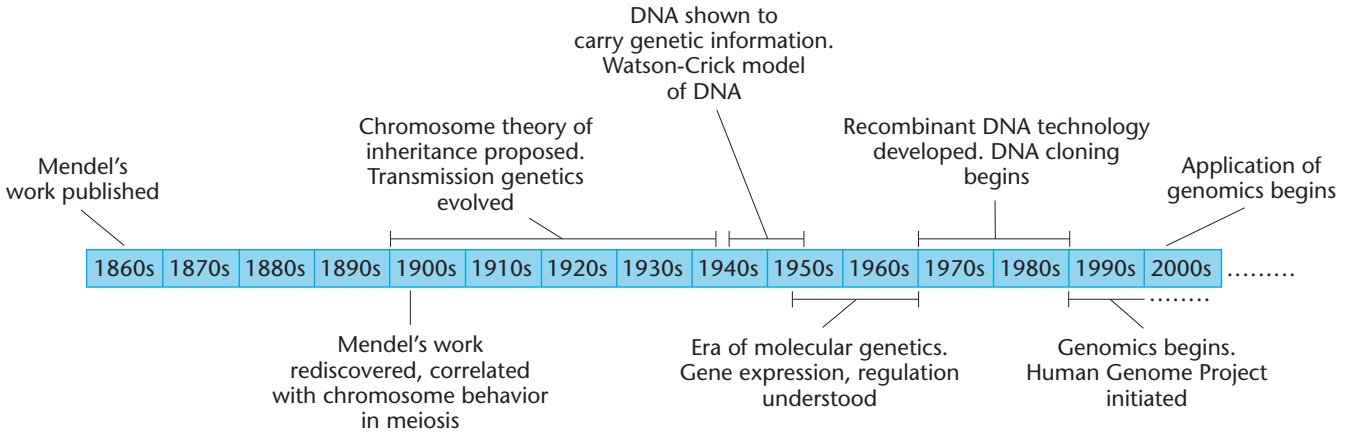


FIGURE 1–11 A timeline showing the development of genetics from Gregor Mendel's work on pea plants to the current era of genomics and its many applications in research, medicine, and society. Having a sense of the history of discovery in genetics should provide you with a useful framework as you proceed through this textbook.

The Nobel Prize and Genetics

No other scientific discipline has experienced the explosion of information and the level of excitement generated by the discoveries in genetics. This impact is especially apparent in the list of Nobel Prizes related to genetics, beginning with those awarded in the early and mid-twentieth century and continuing into the present (see inside front cover). Nobel Prizes in Medicine or Physiology and Chemistry have been consistently awarded for work in genetics and related fields. The first Nobel Prize awarded for such work was given to Thomas Morgan in 1933 for his research on the chromosome theory of inheritance. That award was followed by many others, including prizes for the discovery of genetic recombination, the relationship between genes and proteins, the structure of DNA, and the genetic code. In this century, geneticists continue to be recognized for their impact on biology in the current millennium, including Nobel Prizes awarded in 2002, 2006, 2007, and 2009. In 2010, the prize in Physiology or Medicine was given to Robert Edwards for the development of *in vitro* fertilization, and the 2012 prize was awarded to John Gurdon and Shinya Yamanaka for their work showing that adult cells can be reprogrammed to direct embryonic development and to form stem cells.

Genetics and Society

Just as there has never been a more exciting time to study genetics, the impact of this discipline on society has never been more profound. Genetics and its applications in biotechnology are developing much faster than the social conventions, public policies, and laws required to regulate their use. As a society, we are grappling with a host of sensitive genetics-related issues, including concerns about prenatal testing, genetic discrimination, ownership of genes, access to and safety of gene therapy, and genetic privacy. By the time you finish this course, you will have seen more than enough evidence to convince yourself that the present is the Age of Genetics, and you will understand the need to think about and become a participant in the dialogue concerning genetic science and its use.

ESSENTIAL POINT

Genetic technology is having a profound effect on society, but policies and legislation governing its use are lagging behind the resulting innovations. ■

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

1. How does Mendel's work on the transmission of traits relate to our understanding of genetics today?

CONCEPT QUESTION

2. Review the Chapter Concepts list on p. 17. Most of these concepts are related to the discovery of DNA as the genetic material and the subsequent development of recombinant DNA technology. Write a brief essay that discusses the impact of recombinant DNA technology on genetics as we perceive the discipline today. ■
3. What is the chromosome theory of inheritance, and how is it related to Mendel's findings?
4. Define genotype and phenotype. Describe how they are related and how alleles fit into your definitions.
5. Given the state of knowledge at the time of the Avery, MacLeod, and McCarty experiment, why was it difficult for some scientists to accept that DNA is the carrier of genetic information?
6. What is a gene?
7. What is the structure of DNA? How does it differ from that of RNA?
8. Describe the central dogma of molecular genetics and how it serves as the basis of modern genetics.
9. Until the mid-1940s, many scientists considered proteins to be the likely candidates for the genetic material. Why?
10. Outline the roles played by restriction enzymes and vectors in cloning DNA.
11. Genetics is commonly seen as being grouped into several general areas: transmission, molecular, and population/evolution. Which biological processes are studied in transmission genetics?
12. Summarize the arguments for and against patenting genetically modified organisms.
13. We all carry about 20,000 genes in our genome. So far, patents have been issued for more than 6000 of these genes. Do you think that companies or individuals should be able to patent human genes? Why or why not?
14. Why do we use model organisms to study human genetic diseases?
15. If you knew that a devastating late-onset inherited disease runs in your family (in other words, a disease that does not appear until later in life) and you could be tested for it at the age of 20, would you want to know whether you are a carrier? Would your answer be likely to change when you reach age 40?
16. Why have the advances in bioinformatics kept pace with the advances in biotechnology, while the policies and legislation regarding the ethical issues involved have lagged behind?

2 Mitosis and Meiosis

CHAPTER CONCEPTS

- Genetic continuity between generations of cells and between generations of sexually reproducing organisms is maintained through the processes of mitosis and meiosis, respectively.
- Diploid eukaryotic cells contain their genetic information in pairs of homologous chromosomes, with one member of each pair being derived from the maternal parent and one from the paternal parent.
- Mitosis provides a mechanism by which chromosomes, having been duplicated, are distributed into progeny cells during cell reproduction.
- Mitosis converts a diploid cell into two diploid daughter cells.
- The process of meiosis distributes one member of each homologous pair of chromosomes into each gamete or spore, thus reducing the diploid chromosome number to the haploid chromosome number.
- Meiosis generates genetic variability by distributing various combinations of maternal and paternal members of each homologous pair of chromosomes into gametes or spores.
- During the stages of mitosis and meiosis, the genetic material is condensed into discrete structures called chromosomes.

Chromosomes in the prometaphase stage of mitosis, derived from a cell in the flower of *Haemanthus*.

Every living thing contains a substance described as the genetic material. Except in certain viruses, this material is composed of the nucleic acid DNA. DNA has an underlying linear structure possessing segments called genes, the products of which direct the metabolic activities of cells. An organism's DNA, with its arrays of genes, is organized into structures called **chromosomes,** which serve as vehicles for transmitting genetic information. The manner in which chromosomes are transmitted from one generation of cells to the next and from organisms to their descendants must be exceedingly precise. In this chapter we consider exactly how genetic continuity is maintained between cells and organisms.

Two major processes are involved in the genetic continuity of nucleated cells: **mitosis** and **meiosis.** Although the mechanisms of the two processes are similar in many ways, the outcomes are quite different. Mitosis leads to the production of two cells, each with the same number of chromosomes as the parent cell. In contrast, meiosis reduces the genetic content and the number of chromosomes by precisely half. This reduction is essential if sexual reproduction is to occur without doubling the amount of genetic material in each new generation. Strictly speaking, mitosis is that portion of the cell cycle during which the hereditary components are equally partitioned into daughter cells. Meiosis is part of a special type of cell division that leads to the production of sex cells: **gametes** or **spores.** This process is an essential step in the transmission of genetic information from an organism to its offspring.

Normally, chromosomes are visible only during mitosis and meiosis. When cells are not undergoing division, the genetic material making up chromosomes unfolds and uncoils into a diffuse network within the nucleus, generally referred to as **chromatin.** Before describing mitosis and meiosis, we will briefly review the structure of cells, emphasizing components that are of particular significance to genetic function. We will also compare the structural differences between the prokaryotic (nonnucleated) cells of bacteria and the eukaryotic cells of higher organisms. We then devote the remainder of the chapter to the behavior of chromosomes during cell division.

2.1 Cell Structure Is Closely Tied to Genetic Function

Before 1940, our knowledge of cell structure was limited to what we could see with the light microscope. Around 1940, the transmission electron microscope was in its early stages of development, and by 1960, many details of cell ultrastructure had emerged. Under the electron microscope, cells were seen as highly organized structures whose form and function are dependent on specific genetic expression by each cell type. A new world of whorled membranes, organelles, microtubules, granules, and filaments was revealed. These discoveries revolutionized thinking in the entire field of biology. Many cell components, such as the nucleolus, ribosome, and centriole, are involved directly or indirectly with genetic processes. Other components—the mitochondria and chloroplasts—contain their own unique genetic information. Here, we will focus primarily on those aspects of cell structure that relate to genetic study. The generalized animal cell shown in **Figure 2–1** illustrates most of the structures we will discuss.

All cells are surrounded by a **plasma membrane,** an outer covering that defines the cell boundary and delimits the cell from its immediate external environment. This membrane is not passive but instead actively controls the movement of materials into and out of the cell. In addition to this membrane, plant cells have an outer covering called

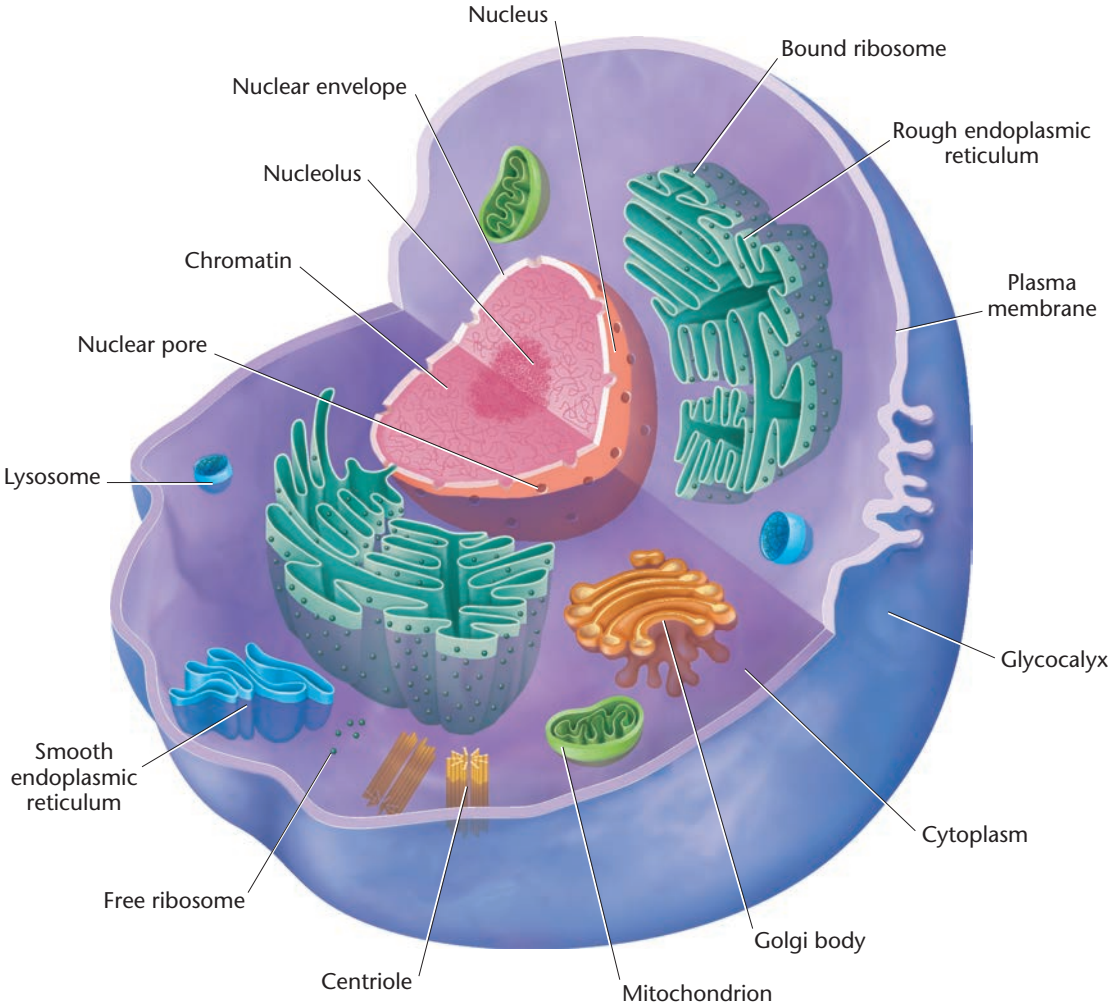


FIGURE 2–1 A generalized animal cell. The cellular components discussed in the text are emphasized here.

the **cell wall** whose major component is a polysaccharide called *cellulose*.

Many, if not most, animal cells have a covering over the plasma membrane, referred to as the **glycocalyx,** or **cell coat.** Consisting of glycoproteins and polysaccharides, this covering has a chemical composition that differs from comparable structures in either plants or bacteria. The glycocalyx provides biochemical identity at the surface of cells. For example, various cell-identity markers that you may have heard of—the AB, Rh, and MN antigens—are found on the surface of red blood cells, among other cell types. On the surface of other cells, histocompatibility antigens, which elicit an immune response during tissue and organ transplants, are present. Various **receptor molecules** are also found on the surfaces of cells. These molecules act as recognition sites that transfer specific chemical signals across the cell membrane into the cell.

Living organisms are categorized into two major groups depending on whether or not their cells contain a nucleus. The presence of a nucleus and other membranous organelles is the defining characteristic of **eukaryotic organisms.** The **nucleus** in eukaryotic cells is a membrane-bound structure that houses the genetic material, DNA, which is complexed with an array of acidic and basic proteins into thin fibers. During nondivisional phases of the cell cycle, the fibers are uncoiled and dispersed into **chromatin** (as mentioned above). During mitosis and meiosis, chromatin fibers coil and condense into **chromosomes.** Also present in the nucleus is the **nucleolus,** an amorphous component where ribosomal RNA (rRNA) is synthesized and where the initial stages of ribosomal assembly occur. The portions of DNA that encode rRNA are collectively referred to as the **nucleolus organizer region,** or the **NOR.**

Prokaryotic organisms, of which there are two major groups, lack a nuclear envelope and membranous organelles. For the purpose of our brief discussion here, we will consider the *eubacteria,* the other group being the more ancient bacteria referred to as *archaea.* In eubacteria, such as *Escherichia coli,* the genetic material is present as a long, circular DNA molecule that is compacted into an unenclosed region called the **nucleoid.** Part of the DNA may be attached to the cell membrane, but in general the nucleoid extends through a large part of the cell. Although the DNA is compacted, it does not undergo the extensive coiling characteristic of the stages of mitosis, during which the chromosomes of eukaryotes become visible. Nor is the DNA associated as extensively with proteins as is eukaryotic DNA. **Figure 2–2**, which shows two bacteria forming by cell division, illustrates the nucleoid regions containing the bacterial chromosomes. Prokaryotic cells do not have a distinct nucleolus but do contain genes that specify rRNA molecules.

The remainder of the eukaryotic cell within the plasma membrane, excluding the nucleus, is referred to as **cytoplasm** and includes a variety of extranuclear cellular

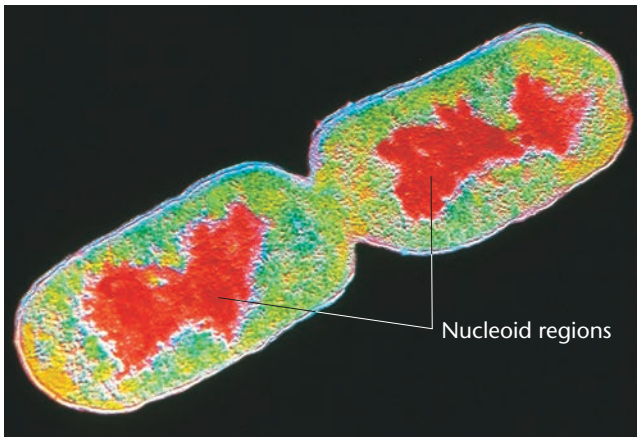


FIGURE 2–2 Color-enhanced electron micrograph of *E. coli* undergoing cell division. Particularly prominent are the two chromosomal areas (shown in red), called nucleoids, that have been partitioned into the daughter cells.

organelles. One organelle, the membranous **endoplasmic reticulum (ER),** compartmentalizes the cytoplasm, greatly increasing the surface area available for biochemical synthesis. The ER appears smooth in places where it serves as the site for synthesizing fatty acids and phospholipids; in other places, it appears rough because it is studded with ribosomes. **Ribosomes** serve as sites where genetic information contained in messenger RNA (mRNA) is translated into proteins.

Three other cytoplasmic structures are very important in the eukaryotic cell's activities: mitochondria, chloroplasts, and centrioles. **Mitochondria** are found in most eukaryotes, including both animal and plant cells, and are the sites of the oxidative phases of cell respiration. These chemical reactions generate large amounts of the energy-rich molecule adenosine triphosphate (ATP). **Chloroplasts,** which are found in plants, algae, and some protozoans, are associated with photosynthesis, the major energy-trapping process on Earth. Both mitochondria and chloroplasts contain DNA in a form distinct from that found in the nucleus. They are able to duplicate themselves and transcribe and translate their own genetic information.

Animal cells and some plant cells also contain a pair of complex structures called **centrioles.** These cytoplasmic bodies, each located in a specialized region called the **centrosome,** are associated with the organization of spindle fibers that function in mitosis and meiosis. In some organisms, the centriole is derived from another structure, the basal body, which is associated with the formation of cilia and flagella (hair-like and whip-like structures for propelling cells or moving materials).

The organization of **spindle fibers** by the centrioles occurs during the early phases of mitosis and meiosis. These fibers play an important role in the movement of chromosomes as they separate during cell division. They are composed of arrays of microtubules consisting of polymers of the protein tubulin.

ESSENTIAL POINT

Most components of cells are involved directly or indirectly with genetic processes. ■

2.2 Chromosomes Exist in Homologous Pairs in Diploid Organisms

As we discuss the processes of mitosis and meiosis, it is important that you understand the concept of homologous chromosomes. Such an understanding will also be of critical importance in our future discussions of Mendelian genetics. Chromosomes are most easily visualized during mitosis. When they are examined carefully, distinctive lengths and shapes are apparent. Each chromosome contains a constricted region called the **centromere,** whose location establishes the general appearance of each chromosome. Figure 2–3 shows chromosomes with centromere placements at different distances along their length. Extending from either side of the centromere are the arms of the chromosome. Depending on the position of the centromere, different arm ratios are produced. As Figure 2–3 illustrates, chromosomes are classified as **metacentric, submetacentric, acrocentric,** or **telocentric** on the basis of the centromere location. The shorter arm, by convention, is shown above the centromere and is called the **p arm** (p, for "petite"). The longer arm is shown below the centromere and is called the **q arm** (q because it is the next letter in the alphabet).

In the study of mitosis, several other observations are of particular relevance. First, all somatic cells derived from members of the same species contain an identical number of chromosomes. In most cases, this represents the **diploid number (2*n*),** whose meaning will become clearer below. When the lengths and centromere placements of all such chromosomes are examined, a second general feature is apparent. With the exception of sex chromosomes, they exist in pairs with regard to these two properties, and the members of each pair are called **homologous chromosomes.** So, for each chromosome exhibiting a specific length and centromere placement, another exists with identical features.

There are exceptions to this rule. Many bacteria and viruses have but one chromosome, and organisms such as yeasts and molds, and certain plants such as bryophytes (mosses), spend the predominant phase of their life cycle in the haploid stage. That is, they contain only one member of each homologous pair of chromosomes during most of their lives.

Figure 2–4 illustrates the physical appearance of different pairs of homologous chromosomes. There, the human mitotic chromosomes have been photographed, cut out of the print, and matched up, creating a display called a **karyotype.** As you can see, humans have a 2*n* number of 46 chromosomes, which on close examination exhibit a diversity of sizes and centromere placements. Note also that each of the 46 chromosomes in this karyotype is clearly a double structure consisting of two parallel *sister chromatids* connected by a common centromere. Had these chromosomes been allowed to continue dividing, the sister chromatids, which are replicas of one another, would have separated into the two new cells as division continued.

The **haploid number (*n*)** of chromosomes is equal to one-half the diploid number. Collectively, the genetic information contained in a haploid set of chromosomes constitutes the **genome** of the species. This, of course, includes copies of all genes as well as a large amount of noncoding DNA. The examples listed in Table 2.1 demonstrate the wide range of *n* values found in plants and animals.

Centromere location	Designation	Metaphase shape	Anaphase shape
Middle	Metacentric	Sister chromatids, Centromere	Migration
Between middle and end	Submetacentric	p arm, q arm	
Close to end	Acrocentric		
At end	Telocentric		

FIGURE 2–3 Centromere locations and the chromosome designations that are based on them. Note that the shape of the chromosome during anaphase is determined by the position of the centromere during metaphase.

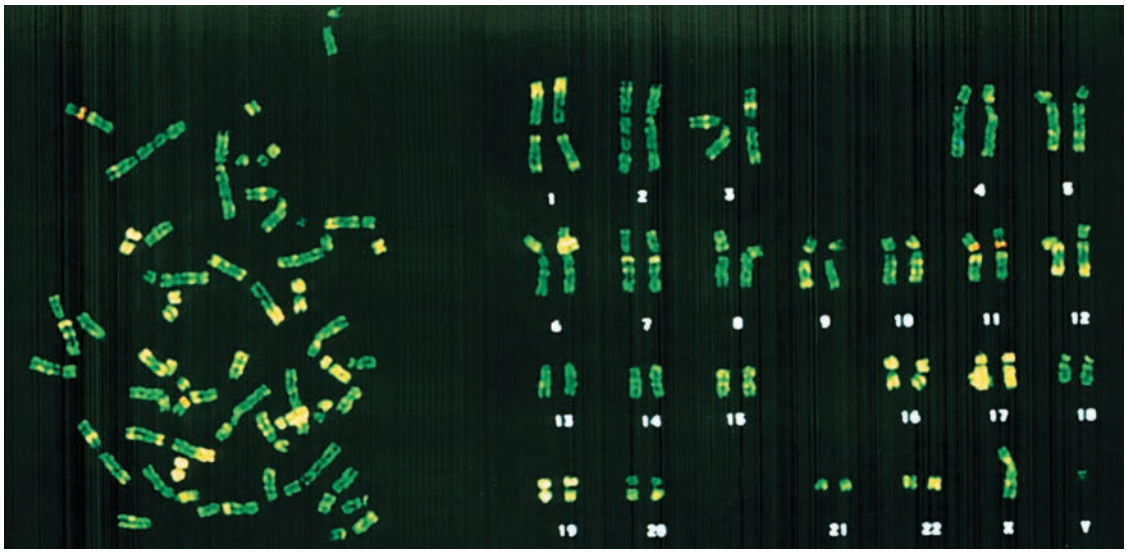


FIGURE 2–4 A metaphase preparation of chromosomes derived from a dividing cell of a human male (left), and the karyotype derived from the metaphase preparation (right). All but the X and Y chromosomes are present in homologous pairs. Each chromosome is clearly a double structure consisting of a pair of sister chromatids joined by a common centromere.

Homologous chromosomes have important genetic similarities. They contain identical gene sites along their lengths; each site is called a **locus** (pl. loci). Thus, they are identical in the traits that they influence and in their genetic potential. In sexually reproducing organisms, one member of each pair is derived from the maternal parent (through the ovum) and the other member is derived from the paternal parent (through the sperm). Therefore, each diploid organism contains two copies of each gene as a consequence of **biparental inheritance,** inheritance from two parents. As we shall see in the chapters on transmission genetics, the members of each pair of genes, while influencing the same characteristic or trait, need not be identical. In a population of members of the same species, many different alternative forms of the same gene, called **alleles,** can exist.

TABLE 2.1 The Haploid Number of Chromosomes for a Variety of Organisms

Common Name	Scientific Name	Haploid Number
Chimpanzee	*Pan troglodytes*	24
Corn	*Zea mays*	10
Fruit fly	*Drosophila melanogaster*	4
Garden pea	*Pisum sativum*	7
House mouse	*Mus musculus*	20
Human	*Homo sapiens*	23
Pink bread mold	*Neurospora crassa*	7
Roundworm	*Caenorhabditis elegans*	6
Yeast	*Saccharomyces cerevisiae*	16

The concepts of haploid number, diploid number, and homologous chromosomes are important for understanding the process of meiosis. During the formation of gametes or spores, meiosis converts the diploid number of chromosomes to the haploid number. As a result, haploid gametes or spores contain precisely one member of each homologous pair of chromosomes—that is, one complete haploid set. Following fusion of two gametes at fertilization, the diploid number is reestablished; that is, the zygote contains two complete haploid sets of chromosomes. The constancy of genetic material is thus maintained from generation to generation.

There is one important exception to the concept of homologous pairs of chromosomes. In many species, one pair, consisting of the **sex-determining chromosomes,** is often not homologous in size, centromere placement, arm ratio, or genetic content. For example, in humans, while females carry two homologous X chromosomes, males carry one Y chromosome in addition to one X chromosome (Figure 2–4). These X and Y chromosomes are not strictly homologous. The Y is considerably smaller and lacks most of the gene loci contained on the X. Nevertheless, they contain homologous regions and behave as homologs in meiosis so that gametes produced by males receive either one X or one Y chromosome.

ESSENTIAL POINT

In diploid organisms, chromosomes exist in homologous pairs, where each member is identical in size, centromere placement, and gene sites. One member of each pair is derived from the maternal parent, and one is derived from the paternal parent. ■

2.3 Mitosis Partitions Chromosomes into Dividing Cells

The process of mitosis is critical to all eukaryotic organisms. In some single-celled organisms, such as protozoans and some fungi and algae, mitosis (as a part of cell division) provides the basis for asexual reproduction. Multicellular diploid organisms begin life as single-celled fertilized eggs called **zygotes.** The mitotic activity of the zygote and the subsequent daughter cells is the foundation for the development and growth of the organism. In adult organisms, mitotic activity is the basis for wound healing and other forms of cell replacement in certain tissues. For example, the epidermal cells of the skin and the intestinal lining of humans are continuously sloughed off and replaced. Cell division also results in the continuous production of reticulocytes that eventually shed their nuclei and replenish the supply of red blood cells in vertebrates. In abnormal situations, somatic cells may lose control of cell division and form a tumor.

The genetic material is partitioned into daughter cells during nuclear division, or **karyokinesis.** This process is quite complex and requires great precision. The chromosomes must first be exactly replicated and then accurately partitioned. The end result is the production of two daughter nuclei, each with a chromosome composition identical to that of the parent cell.

Karyokinesis is followed by cytoplasmic division, or **cytokinesis.** This less complex process requires a mechanism that partitions the volume into two parts, then encloses each new cell in a distinct plasma membrane. As the cytoplasm is reconstituted, organelles replicate themselves, arise from existing membrane structures, or are synthesized *de novo* (anew) in each cell.

Following cell division, the initial size of each new daughter cell is approximately one-half the size of the parent cell. However, the nucleus of each new cell is not appreciably smaller than the nucleus of the original cell. Quantitative measurements of DNA confirm that there is an amount of genetic material in the daughter nuclei equivalent to that in the parent cell.

Interphase and the Cell Cycle

Many cells undergo a continuous alternation between division and nondivision. The events that occur from the completion of one division until the completion of the next division constitute the **cell cycle** (Figure 2–5). We will consider **interphase,** the initial stage of the cell cycle, as the interval between divisions. It was once thought that the biochemical activity during interphase was devoted solely to the cell's growth and its normal function. However, we now know that another biochemical step critical to the ensuing mitosis occurs during interphase: *the replication of the DNA of each chromosome.* This period, during which DNA is synthesized, occurs before the cell enters mitosis and is called the **S phase.** The initiation and completion of synthesis can be detected by monitoring the incorporation of radioactive precursors into DNA.

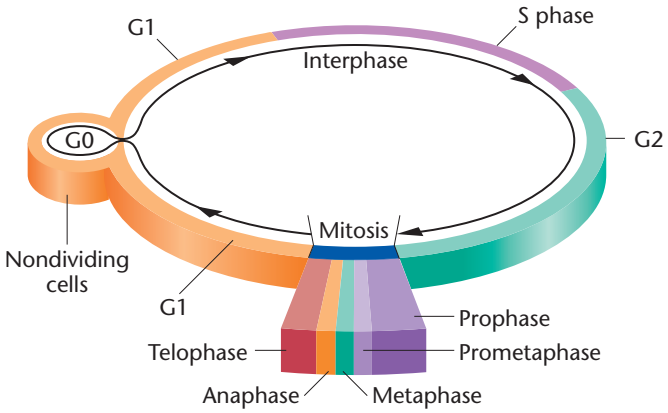


FIGURE 2–5 The stages comprising an arbitrary cell cycle. Following mitosis, cells enter the G1 stage of interphase, initiating a new cycle. Cells may become nondividing (G0) or continue through G1, where they become committed to begin DNA synthesis (S) and complete the cycle (G2 and mitosis). Following mitosis, two daughter cells are produced, and the cycle begins anew for both of them.

Investigations of this nature demonstrate two periods during interphase when no DNA synthesis occurs, one before and one after the S phase. These are designated **G1 (gap I)** and **G2 (gap II),** respectively. During both of these intervals, as well as during S, intensive metabolic activity, cell growth, and cell differentiation are evident. By the end of G2, the volume of the cell has roughly doubled, DNA has been replicated, and mitosis (M) is initiated. Following mitosis, continuously dividing cells then repeat this cycle (G1, S, G2, M) over and over, as shown in Figure 2–5.

Much is known about the cell cycle based on *in vitro* (literally, "in glass") studies. When grown in culture, many cell types in different organisms traverse the complete cycle in about 16 hours. The actual process of mitosis occupies only a small part of the overall cycle, often less than an hour. The lengths of the S and G2 phases of interphase are fairly consistent in different cell types. Most variation is seen in the length of time spent in the G1 stage. Figure 2–6 shows the relative length of these

Interphase			Mitosis
G1	S	G2	M
5	7	3	1

Hours

Pro	Met	Ana	Tel
36	3	3	18

Minutes

FIGURE 2–6 The time spent in each interval of one complete cell cycle of a human cell in culture. Times vary according to cell types and conditions.

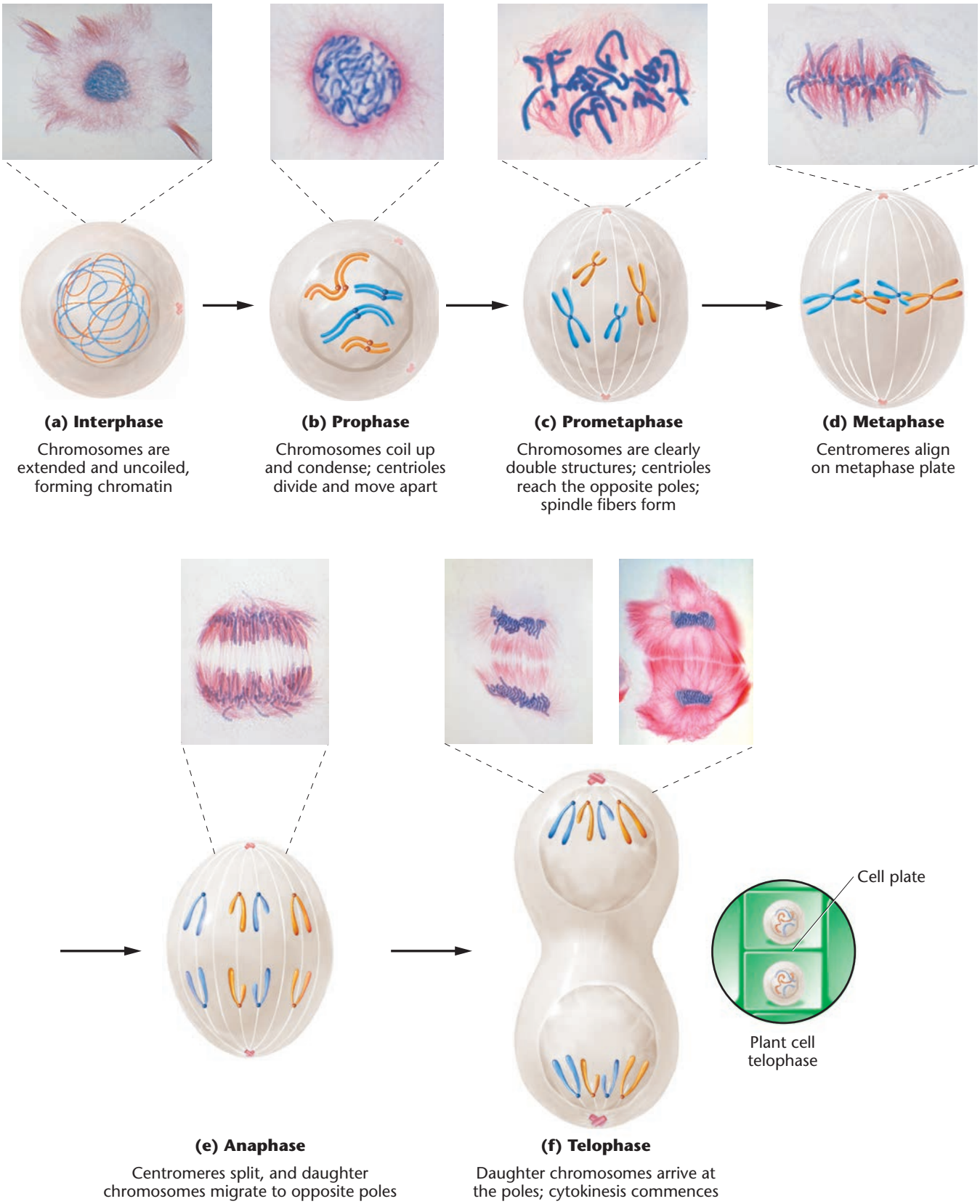


FIGURE 2–7 Drawings depicting mitosis in an animal cell with a diploid number of 4. The events occurring in each stage are described in the text. Of the two homologous pairs of chromosomes, one pair consists of longer, metacentric members and the other of shorter, submetacentric members. The maternal chromosome and the paternal chromosome of each pair are shown in different colors. In (f), a drawing of late telophase in a plant cell shows the formation of the cell plate and lack of centrioles. The cells shown in the light micrographs came from the flower of *Haemanthus,* a plant that has a diploid number of 8.

intervals as well as the length of the stages of mitosis in a human cell in culture.

G1 is of great interest in the study of cell proliferation and its control. At a point during G1, all cells follow one of two paths. They either withdraw from the cycle, become quiescent, and enter the **G0 stage** (see Figure 2–5), or they become committed to proceed through G1, initiating DNA synthesis, and completing the cycle. Cells that enter G0 remain viable and metabolically active but are not proliferative. Cancer cells apparently avoid entering G0 or pass through it very quickly. Other cells enter G0 and never reenter the cell cycle. Still other cells in G0 can be stimulated to return to G1 and thereby reenter the cell cycle.

Cytologically, interphase is characterized by the absence of visible chromosomes. Instead, the nucleus is filled with chromatin fibers that are formed as the chromosomes uncoil and disperse after the previous mitosis [**Figure 2–7(a)**]. Once G1, S, and G2 are completed, mitosis is initiated. Mitosis is a dynamic period of vigorous and continual activity. For discussion purposes, the entire process is subdivided into discrete stages, and specific events are assigned to each one. These stages, in order of occurrence, are prophase, prometaphase, metaphase, anaphase, and telophase. They are diagrammed with corresponding photomicrographs in Figure 2–7.

Prophase

Often, over half of mitosis is spent in **prophase** [**Figure 2–7(b)**], a stage characterized by several significant occurrences. One of the early events in prophase of all animal cells is the migration of two pairs of centrioles to opposite ends of the cell. These structures are found just outside the nuclear envelope in an area of differentiated cytoplasm called the centrosome (introduced in Section 2.1). It is believed that each pair of centrioles consists of one mature unit and a smaller, newly formed daughter centriole.

The centrioles migrate and establish poles at opposite ends of the cell. After migration, the centrosomes, in which the centrioles are localized, are responsible for organizing cytoplasmic microtubules into the spindle fibers that run between these poles, creating an axis along which chromosomal separation occurs. Interestingly, the cells of most plants (there are a few exceptions), fungi, and certain algae seem to lack centrioles. Spindle fibers are nevertheless apparent during mitosis.

As the centrioles migrate, the nuclear envelope begins to break down and gradually disappears. In a similar fashion, the nucleolus disintegrates within the nucleus. While these events are taking place, the diffuse chromatin fibers have begun to condense, until distinct threadlike structures, the chromosomes, become visible. It becomes apparent near the end of prophase that each chromosome is actually a double structure split longitudinally except at a single point of constriction, the centromere. The two parts of each chromosome are called **sister chromatids** because the DNA contained in each of them is genetically identical, having formed from a single replicative event. Sister chromatids are held together by a multi-subunit protein complex called **cohesin.** This molecular complex is originally formed between them during the S phase of the cell cycle when the DNA of each chromosome is replicated. Thus, even though we cannot see chromatids in interphase because the chromatin is uncoiled and dispersed in the nucleus, the chromosomes are already double structures, which becomes apparent in late prophase. In humans, with a diploid number of 46, a cytological preparation of late prophase reveals 46 chromosomes randomly distributed in the area formerly occupied by the nucleus.

Prometaphase and Metaphase

The distinguishing event of the two ensuing stages is the migration of every chromosome, led by its centromeric region, to the equatorial plane. The equatorial plane, also referred to as the *metaphase plate,* is the midline region of the cell, a plane that lies perpendicular to the axis established by the spindle fibers. In some descriptions, the term **prometaphase** refers to the period of chromosome movement [**Figure 2–7(c)**], and the term **metaphase** is applied strictly to the chromosome configuration following migration.

Migration is made possible by the binding of spindle fibers to the chromosome's **kinetochore,** an assembly of multilayered plates of proteins associated with the centromere. This structure forms on opposite sides of each paired centromere, in intimate association with the two sister chromatids. Once properly attached to the spindle fibers, cohesin is degraded by an enzyme, appropriately named *separase,* and the sister chromatid arms disjoin, except at the centromere region. A unique protein family called **shugoshin** (from the Japanese meaning guardian spirit) protects cohesin from being degraded by separase at the centromeric regions. The involvement of the cohesin and shugoshin complexes with a pair of sister chromatids during mitosis is depicted in **Figure 2–8**.

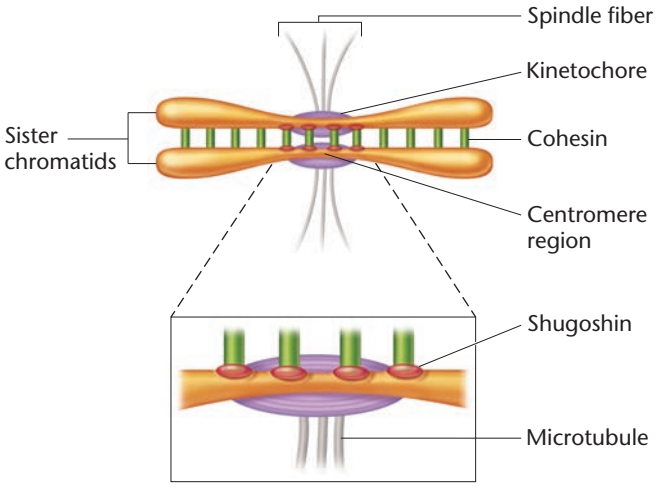


FIGURE 2–8 The depiction of the alignment, pairing, and disjunction of sister chromatids during mitosis, involving the molecular complexes cohesin and shugoshin and the enzyme separase.

We know a great deal about the molecular interactions involved in kinetechore assembly along the centromere. This is of great interest because of the consequences when mutations alter the proteins that make up the kinetechore complex. Altered kinetechore function potentially leads to errors during chromosome migration, altering the diploid content of daughter cells. A more detailed account will be presented later in the text, once we have provided more information about DNA and the proteins that make up chromatin (see Chapter 11).

We also know a great deal about spindle fibers. They consist of microtubules, which themselves consist of molecular subunits of the protein tubulin. Microtubules seem to originate and "grow" out of the two centrosome regions at opposite poles of the cell. They are dynamic structures that lengthen and shorten as a result of the addition or loss of polarized tubulin subunits. The microtubules most directly responsible for chromosome migration make contact with, and adhere to, kinetochores as they grow from the centrosome region. They are referred to as **kinetochore microtubules** and have one end near the centrosome region (at one of the poles of the cell) and the other end anchored to the kinetochore. The number of microtubules that bind to the kinetochore varies greatly between organisms. Yeast (*Saccharomyces*) has only a single microtubule bound to each plate-like structure of the kinetochore. Mitotic cells of mammals, at the other extreme, reveal 30 to 40 microtubules bound to each portion of the kinetochore.

At the completion of metaphase, each centromere is aligned at the metaphase plate with the chromosome arms extending outward in a random array. This configuration is shown in **Figure 2–7(d)**.

Anaphase

Events critical to chromosome distribution during mitosis occur during **anaphase,** the shortest stage of mitosis. During this phase, sister chromatids of each chromosome, held together only at their centromere regions, *disjoin* (separate) from one another—an event described as **disjunction**—and are pulled to opposite ends of the cell. For complete disjunction to occur: (1) shugoshin must be degraded, reversing its protective role; (2) the cohesin complex holding the centromere region of each sister chromosome is then cleaved by separase; and (3) sister chromatids of each chromosome are pulled toward the opposite poles of the cell (Figure 2–8). As these events proceed, each migrating chromatid is now referred to as a **daughter chromosome.**

The location of the centromere determines the shape of the chromosome during separation, as you saw in Figure 2–3. The steps that occur during anaphase are critical in providing each subsequent daughter cell with an identical set of chromosomes. In human cells, there would now be 46 chromosomes at each pole, one from each original sister pair. **Figure 2–7(e)** shows anaphase prior to its completion.

Telophase

Telophase is the final stage of mitosis and is depicted in **Figure 2–7(f)**. At its beginning, two complete sets of chromosomes are present, one set at each pole. The most significant event of this stage is cytokinesis, the division or partitioning of the cytoplasm. Cytokinesis is essential if two new cells are to be produced from one cell. The mechanism of cytokinesis differs greatly in plant and animal cells, but the end result is the same: two new cells are produced. In plant cells, a **cell plate** is synthesized and laid down across the region of the metaphase plate. Animal cells, however, undergo a constriction of the cytoplasm, much as a loop of string might be tightened around the middle of a balloon.

It is not surprising that the process of cytokinesis varies in different organisms. Plant cells, which are more regularly shaped and structurally rigid, require a mechanism for depositing new cell wall material around the plasma membrane. The cell plate laid down during telophase becomes a structure called the **middle lamella.** Subsequently, the primary and secondary layers of the cell wall are deposited between the cell membrane and middle lamella in each of the resulting daughter cells. In animals, complete constriction of the cell membrane produces the **cell furrow** characteristic of newly divided cells.

Other events necessary for the transition from mitosis to interphase are initiated during late telophase. They generally constitute a reversal of events that occurred during prophase. In each new cell, the chromosomes begin to uncoil and become diffuse chromatin once again, while the nuclear envelope reforms around them, the spindle fibers disappear, and the nucleolus gradually reforms and becomes visible in the nucleus during early interphase. At the completion of telophase, the cell enters interphase.

NOW SOLVE THIS

2–1 With the initial appearance of the feature we call "Now Solve This," a short introduction is in order. The feature occurs several times in this and all ensuing chapters, each time providing a problem related to the discussion just presented. A "Hint" is then offered that may help you solve the problem. Here is the first problem:

(a) If an organism has a diploid number of 16, how many chromatids are visible at the end of mitotic prophase?

(b) How many chromosomes are moving to each pole during anaphase of mitosis?

HINT: *This problem involves an understanding of what happens to each pair of homologous chromosomes during mitosis, asking you to apply your understanding of chromosome behavior to an organism with a diploid number of 16. The key to its solution is your awareness that throughout mitosis, the members of each homologous pair do not pair up, but instead behave independently.*

Cell-Cycle Regulation

The cell cycle, culminating in mitosis, is fundamentally the same in all eukaryotic organisms. This similarity in many diverse organisms suggests that the cell cycle is governed by a genetically regulated program that has been conserved throughout evolution. Because disruption of this regulation may underlie the uncontrolled cell division characterizing malignancy, interest in how genes regulate the cell cycle is particularly strong.

A mammoth research effort has paid high dividends, and we now have knowledge of many genes involved in the control of the cell cycle. This work was recognized by the awarding of the 2001 Nobel Prize in Medicine or Physiology to Lee Hartwell, Paul Nurse, and Tim Hunt. As with other studies of genetic control over essential biological processes, investigation has focused on the discovery of mutations that interrupt the cell cycle and on the effects of those mutations. As we shall return to this subject in much greater detail later in the text during our consideration of the molecular basis of cancer (see Chapter 16), what follows is a very brief overview.

Many mutations are now known that exert an effect at one or another stage of the cell cycle. First discovered in yeast, but now evident in all organisms, including humans, such mutations were originally designated as ***cell division cycle (cdc)* mutations.** The normal products of many of the mutated genes are enzymes called **kinases** that can add phosphates to other proteins. They serve as "master control" molecules functioning in conjunction with proteins called **cyclins.** Cyclins bind to these kinases (creating *cyclin-dependent kinases*), activating them at appropriate times during the cell cycle. Activated kinases then phosphorylate other target proteins that regulate the progress of the cell cycle. The study of *cdc* mutations has established that the cell cycle contains at least three **cell-cycle checkpoints** where the processes culminating in normal mitosis are monitored, or "checked," by these master control molecules before the next stage of the cycle is allowed to commence. These checkpoints will be discussed in Chapter 16.

The importance of cell-cycle control can be demonstrated by considering what happens when this regulatory system is impaired. Let's assume, for example, that the DNA of a cell has incurred damage leading to one or more mutations impairing cell-cycle control. If allowed to proceed through the cell cycle as one of the population of dividing cells, this genetically altered cell would divide uncontrollably—a key step in the development of a cancer cell. If instead the cell cycle is arrested at one of the checkpoints, the cell can repair the DNA damage or permanently stop the cell from dividing, thereby preventing its potential malignancy.

ESSENTIAL POINT

Mitosis is subdivided into discrete stages that initially depict the condensation of chromatin into the diploid number of chromosomes, each of which is initially a double structure, each composed of a pair of sister chromatids. During mitosis, sister chromatids are pulled apart and directed toward opposite poles, after which cytoplasmic division creates two new cells with identical genetic information. ■

2.4 Meiosis Creates Haploid Gametes and Spores and Enhances Genetic Variation in Species

Whereas in diploid organisms, mitosis produces two daughter cells with full diploid complements, **meiosis** produces gametes or spores that are characterized by only one haploid set of chromosomes. During sexual reproduction, haploid gametes then combine at fertilization to reconstitute the diploid complement found in parental cells. Meiosis must be highly specific since, by definition, haploid gametes or spores must contain precisely one member of each homologous pair of chromosomes. When successfully completed, meiosis provides the basis for maintaining genetic continuity from generation to generation.

Another major accomplishment of meiosis is to ensure that during sexual reproduction an enormous amount of genetic variation is produced among members of a species. Such variation occurs in two forms. First, meiosis produces gametes with many unique combinations of maternally and paternally derived chromosomes among the haploid complement, thus assuring that following fertilization, a large number of unique chromosome combinations are possible. As we will see (Chapter 3), this process is the underlying basis of Mendel's principles of segregation and independent assortment. The second source of variation is created by the meiotic event referred to as **crossing over,** which results in genetic exchange between members of each homologous pair of chromosomes prior to one or the other finding its way into a haploid gamete or spore. This creates intact chromosomes that are mosaics of the maternal and paternal homologs from which they arise, further enhancing genetic variation. Sexual reproduction therefore significantly reshuffles the genetic material, producing highly diverse offspring.

Meiosis: Prophase I

As in mitosis, the process in meiosis begins with a diploid cell duplicating its genetic material in the interphase stage preceding chromosome division. To achieve haploidy, two divisions are thus required. The meiotic achievements, as described above, are largely dependent on the behavior of chromosomes during the initial stage of the first division, called **prophase I.** Recall that in mitosis the paternally and maternally derived members of each homologous pair of chromosomes behave autonomously during division. Each chromosome is duplicated, creating genetically identical **sister chromatids,** and

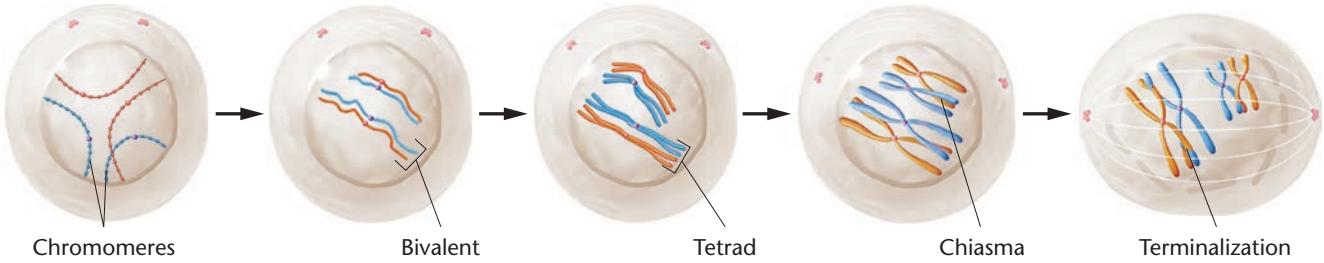


FIGURE 2–9 The events characterizing meiotic prophase I for the chromosomes depicted in Figure 2–7.

subsequently, one chromatid of each pair is distributed to each new cell. The major difference in meiosis is that once the chromatin characterizing interphase has condensed into visible structures, the homologous chromosomes are not autonomous but are instead seen to be paired up, having undergone the process called **synapsis.** Figure 2–9 illustrates this process as well as the ensuing events of prophase I. Each synapsed pair of homologs is initially called a **bivalent,** and the number of bivalents is equal to the haploid number. In Figure 2–9, we have depicted two homologous pairs of chromosomes and thus two bivalents. As the homologs condense and shorten, each bivalent gives rise to a unit called a **tetrad,** consisting of two pairs of sister chromatids, each of which is joined at a common centromere. Remember that one pair of sister chromatids is maternally derived, and the other pair paternally derived. The presence of tetrads is visible evidence

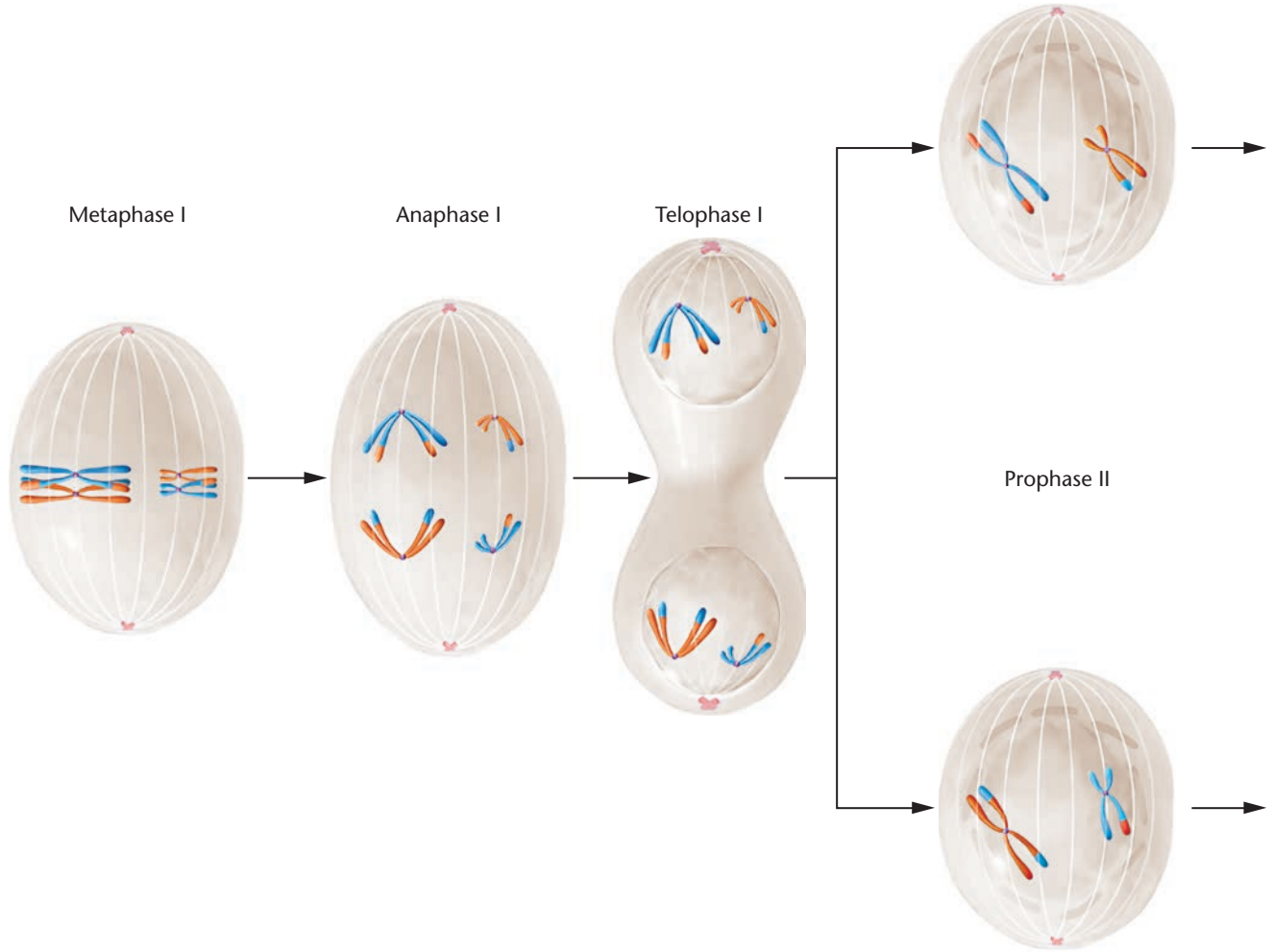


FIGURE 2–10 The major events in meiosis in an animal cell with a diploid number of 4, beginning with metaphase I. Note that the combination of chromosomes in the cells produced following telophase II is dependent on the random alignment of each tetrad and dyad on the equatorial plate during metaphase I and metaphase II. Several other combinations, which are not shown, can also be formed. The events depicted here are described in the text.

that *both* homologs have, in fact, duplicated. As prophase progresses within each tetrad, each pair of sister chromatids is seen to pull apart. However, one or more areas remain in contact where chromatids are intertwined. Each such area, called a **chiasma** (pl., chiasmata), is thought to represent a point where **nonsister chromatids** (one paternal and one maternal chromatid) have undergone genetic exchange through the process of crossing over. Since crossing over is thought to occur one or more times in each tetrad, mosaic chromosomes are routinely created during every meiotic event. During the final period of prophase I, the nucleolus and nuclear envelope break down, and the two centromeres of each tetrad attach to the recently formed spindle fibers.

Metaphase I, Anaphase I, and Telophase I

The remainder of the meiotic process is depicted in **Figure 2–10**. After meiotic prophase I, steps similar to those of mitosis occur. In the first division, **metaphase I,** the chromosomes have maximally shortened and thickened. The terminal chiasmata of each tetrad are visible and appear to be the only factor holding the nonsister chromatids together. Each tetrad interacts with spindle fibers, facilitating movement to the metaphase plate. The alignment of each tetrad prior to the first anaphase is random. Half of each tetrad is pulled randomly to one or the other pole, and the other half then moves to the opposite pole.

During the stages of meiosis I, a single centromere holds each pair of sister chromatids together. It does *not* divide. At **anaphase I,** one-half of each tetrad (the dyad) is pulled toward each pole of the dividing cell. This separation process is the physical basis of disjunction, the separation of chromosomes from one another. Occasionally, errors in meiosis occur and separation is not achieved. The term **nondisjunction** describes such an error. At the completion of a normal anaphase I, a series of dyads equal to the haploid number is present at each pole.

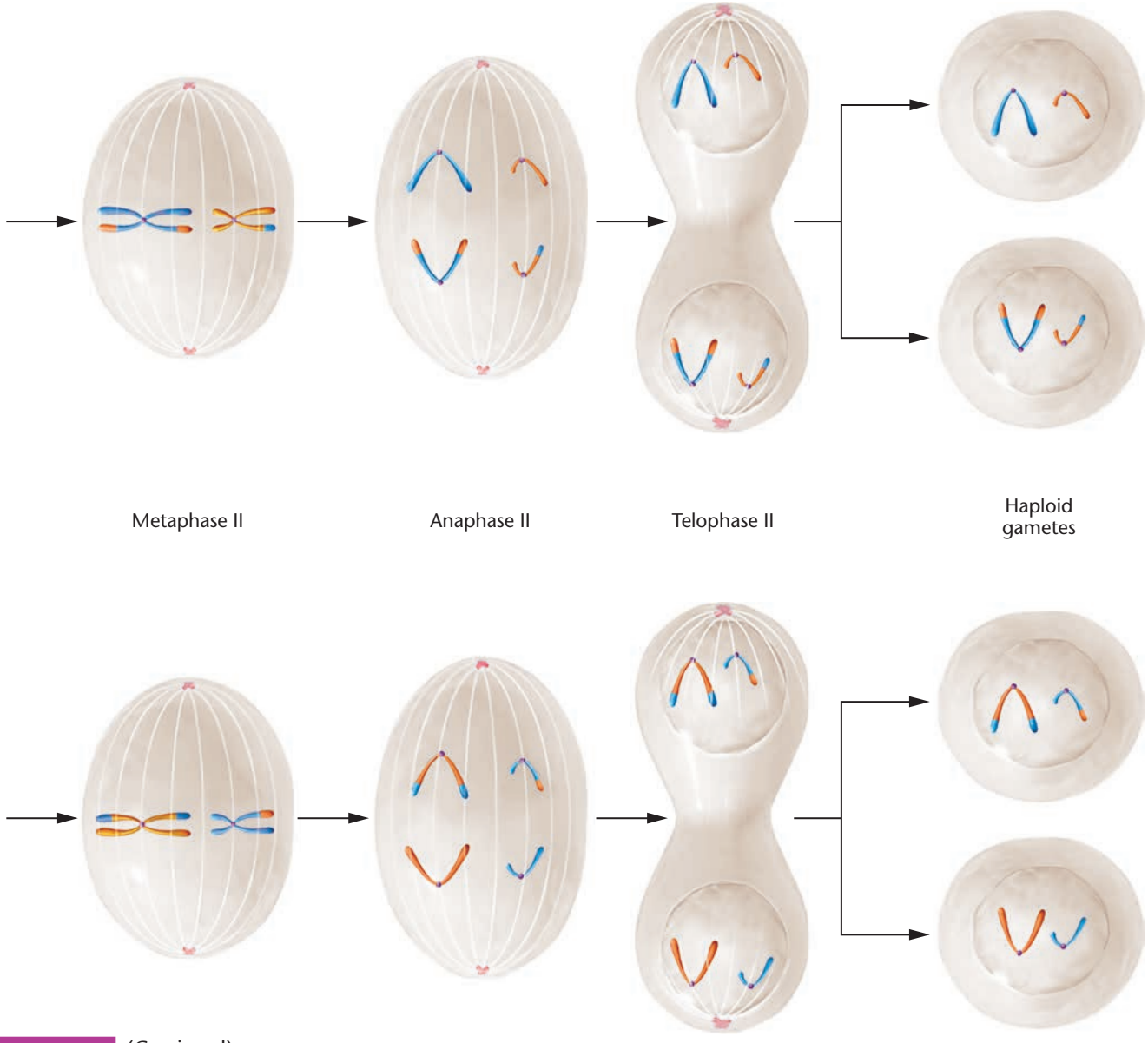


FIGURE 2–10 (Continued)

If crossing over had not occurred in the first meiotic prophase, each dyad at each pole would consist solely of either paternal or maternal chromatids. However, the exchanges produced by crossing over create mosaic chromatids of paternal and maternal origin.

In many organisms, **telophase I** reveals a nuclear membrane forming around the dyads. Next, the nucleus enters into a short interphase period. If interphase occurs, the chromosomes do not replicate since they already consist of two chromatids. In other organisms, the cells go directly from anaphase I to meiosis II. In general, meiotic telophase is much shorter than the corresponding stage in mitosis.

The Second Meiotic Division

A second division, **meiosis II,** is essential if each gamete or spore is to receive only one chromatid from each original tetrad. The stages characterizing meiosis II are shown in the right half of Figure 2–10. During **prophase II,** each dyad is composed of one pair of sister chromatids attached by a common centromere. During **metaphase II,** the centromeres are positioned on the metaphase plate. When they divide, **anaphase II** is initiated, and the sister chromatids of each dyad are pulled to opposite poles. Because the number of dyads is equal to the haploid number, **telophase II** reveals one member of each pair of homologous chromosomes at each pole. Each chromosome is now a monad. Following cytokinesis in telophase II, four haploid gametes may result from a single meiotic event. At the conclusion of meiosis II, not only has the haploid state been achieved, but if crossing over has occurred, each monad is also a combination of maternal and paternal genetic information. As a result, the offspring produced by any gamete receives a mixture of genetic information originally present in his or her grandparents. Meiosis thus significantly increases the level of genetic variation in each ensuing generation.

NOW SOLVE THIS

2–2 An organism has a diploid number of 16 in a primary oocyte. (a) How many tetrads are present in prophase I? (b) How many dyads are present in prophase II? (c) How many monads migrate to each pole during anaphase II?

■ **HINT:** *This problem involves an understanding of what happens to the maternal and paternal members of each pair of homologous chromosomes during meiosis, asking you to extrapolate your understanding to chromosome behavior in an organism with a diploid number of 16. The major insight needed to solve this problem is to understand that maternal and paternal homologs synapse during meiosis. Once it is evident that each chromatid has duplicated, creating a tetrad in the early phases of meiosis, each original pair behaves as a unit and leads to two dyads during anaphase I.*

ESSENTIAL POINT

Meiosis converts a diploid cell into a haploid gamete or spore, making sexual reproduction possible. As a result of chromosome duplication and two subsequent meiotic divisions, each haploid cell receives one member of each homologous pair of chromosomes. ■

2.5 The Development of Gametes Varies in Spermatogenesis Compared to Oogenesis

Although events that occur during the meiotic divisions are similar in all cells participating in gametogenesis in most animal species, there are certain differences between the production of a male gamete (spermatogenesis) and a female gamete (oogenesis). **Figure 2–11** summarizes these processes.

Spermatogenesis takes place in the testes, the male reproductive organs. The process begins with the enlargement of an undifferentiated diploid germ cell called a **spermatogonium.** This cell grows to become a **primary spermatocyte,** which undergoes the first meiotic division. The products of this division, called **secondary spermatocytes,** contain a haploid number of dyads. The secondary spermatocytes then undergo meiosis II, and each of these cells produces two haploid **spermatids.** Spermatids go through a series of developmental changes, **spermiogenesis,** to become highly specialized, motile **spermatozoa,** or **sperm.** All sperm cells produced during spermatogenesis contain the haploid number of chromosomes and equal amounts of cytoplasm.

Spermatogenesis may be continuous or may occur periodically in mature male animals; its onset is determined by the species' reproductive cycles. Animals that reproduce year-round produce sperm continuously, whereas those whose breeding period is confined to a particular season produce sperm only during that time.

In animal **oogenesis,** the formation of **ova** (sing. **ovum**), or eggs, occurs in the ovaries, the female reproductive organs. The daughter cells resulting from the two meiotic divisions of this process receive equal amounts of genetic material, but they do *not* receive equal amounts of cytoplasm. Instead, during each division, almost all the cytoplasm of the **primary oocyte,** itself derived from the **oogonium,** is concentrated in one of the two daughter cells. The concentration of cytoplasm is necessary because a major function of the mature ovum is to nourish the developing embryo following fertilization.

During anaphase I in oogenesis, the tetrads of the primary oocyte separate, and the dyads move toward opposite poles. During telophase I, the dyads at one pole are pinched off with very little surrounding cytoplasm to form the **first polar body.** The first polar body may or may not divide

again to produce two small haploid cells. The other daughter cell produced by this first meiotic division contains most of the cytoplasm and is called the **secondary oocyte.** The mature ovum will be produced from the secondary oocyte during the second meiotic division. During this division, the cytoplasm of the secondary oocyte again divides unequally, producing an **ootid** and a **second polar body.** The ootid then differentiates into the mature ovum.

Unlike the divisions of spermatogenesis, the two meiotic divisions of oogenesis may not be continuous. In some animal species, the second division may directly follow the first. In others, including humans, the first division of all oocytes begins in the embryonic ovary but arrests in prophase I. Many years later, meiosis resumes in each oocyte just prior to its ovulation. The second division is completed only after fertilization.

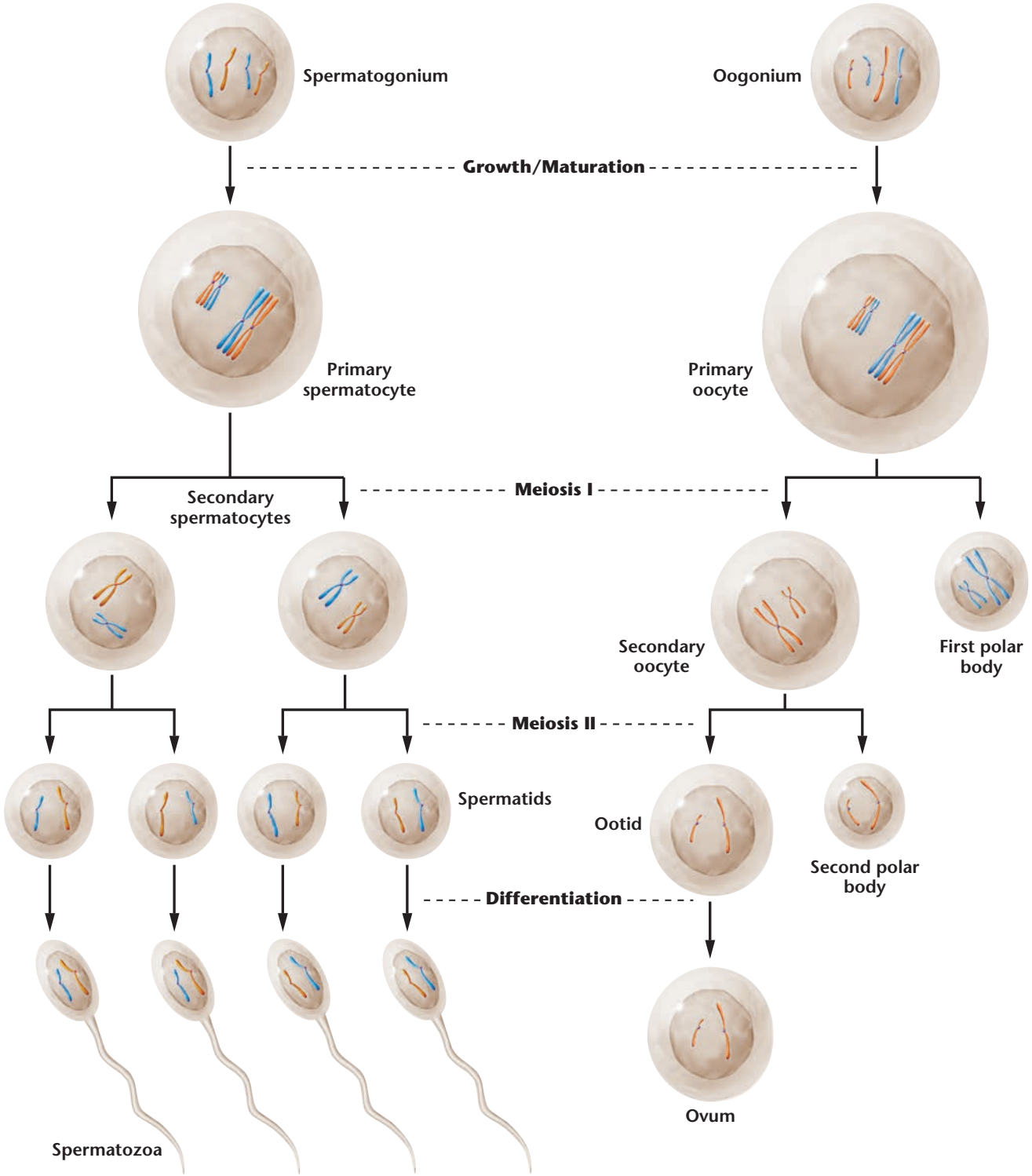


FIGURE 2–11 Spermatogenesis and oogenesis in animal cells.

NOW SOLVE THIS

2–3 Examine Figure 2–11, which shows oogenesis in animal cells. Will the genotype of the second polar body (derived from meiosis II) always be identical to that of the ootid? Why or why not?

■ **HINT:** *This problem involves an understanding of meiosis during oogenesis, asking you to demonstrate your knowledge of polar bodies. The key to its solution is to take into account that crossing over occurred between each pair of homologs during meiosis I.*

ESSENTIAL POINT

There is a major difference between meiosis in males and in females. On the one hand, spermatogenesis partitions the cytoplasmic volume equally and produces four haploid sperm cells. Oogenesis, on the other hand, collects the bulk of cytoplasm in one egg cell and reduces the other haploid products to polar bodies. The extra cytoplasm in the egg contributes to zygote development following fertilization. ■

2.6 Meiosis Is Critical to Sexual Reproduction in All Diploid Organisms

The process of meiosis is critical to the successful sexual reproduction of all diploid organisms. It is the mechanism by which the diploid amount of genetic information is reduced to the haploid amount. In animals, meiosis leads to the formation of gametes, whereas in plants haploid spores are produced, which in turn lead to the formation of haploid gametes.

Each diploid organism stores its genetic information in the form of homologous pairs of chromosomes. Each pair consists of one member derived from the maternal parent and one from the paternal parent. Following meiosis, haploid cells potentially contain either the paternal or the maternal representative of every homologous pair of chromosomes. However, the process of crossing over, which occurs in the first meiotic prophase, further reshuffles the alleles between the maternal and paternal members of each homologous pair, which then segregate and assort independently into gametes. These events result in the great amount of genetic variation present in gametes.

It is important to touch briefly on the significant role that meiosis plays in the life cycles of fungi and plants. In many fungi, the predominant stage of the life cycle consists of haploid vegetative cells. They arise through meiosis and proliferate by mitotic cell division. In multicellular plants, the life cycle alternates between the diploid **sporophyte stage** and the haploid **gametophyte stage.** While one or the other predominates in different plant groups during this "alternation of generations," the processes of meiosis and fertilization constitute the "bridges" between the sporophyte and gametophyte stages. Therefore, meiosis is an essential component of the life cycle of plants.

ESSENTIAL POINT

Meiosis results in extensive genetic variation by virtue of the exchange during crossing over between maternal and paternal chromatids and their random segregation into gametes. In addition, meiosis plays an important role in the life cycles of fungi and plants, serving as the bridge between alternating generations. ■

2.7 Electron Microscopy Has Revealed the Physical Structure of Mitotic and Meiotic Chromosomes

Thus far in this chapter, we have focused on mitotic and meiotic chromosomes, emphasizing their behavior during cell division and gamete formation. An interesting question is why chromosomes are invisible during interphase but visible during the various stages of mitosis and meiosis. Studies using electron microscopy clearly show why this is the case.

Recall that, during interphase, only dispersed chromatin fibers are present in the nucleus [**Figure 2–12(a)**]. Once mitosis begins, however, the fibers coil and fold, condensing into typical mitotic chromosomes [**Figure 2–12(b)**]. If the fibers comprising a mitotic chromosome are loosened, the areas of greatest spreading reveal individual fibers similar to those seen in interphase chromatin [**Figure 2–12(c)**].

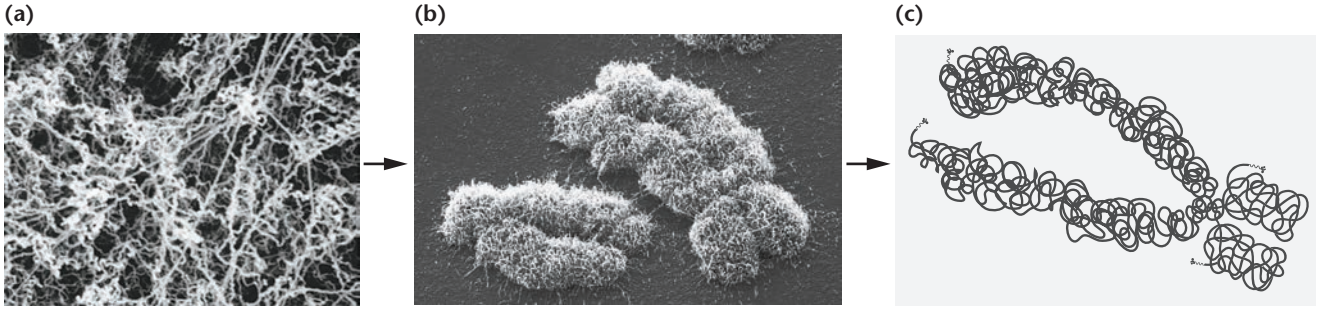


FIGURE 2–12 Comparison of (a) the chromatin fibers characteristic of the interphase nucleus with (b) metaphase chromosomes that are derived from chromatin during mitosis. Part (c) diagrams a mitotic chromosome, showing how chromatin is condensed to produce it. Part (a) is a transmission electron micrograph and part (b) is a scanning electron micrograph.

Very few fiber ends seem to be present, and in some cases, none can be seen. Instead, individual fibers always seem to loop back into the interior. Such fibers are obviously twisted and coiled around one another, forming the regular pattern of folding in the mitotic chromosome. Starting in late telophase of mitosis and continuing during G1 of interphase, chromosomes unwind to form the long fibers characteristic of chromatin, which consist of DNA and associated proteins, particularly proteins called *histones*. It is in this physical arrangement that DNA can most efficiently function during transcription and replication.

Electron microscopic observations of metaphase chromosomes in varying degrees of coiling led Ernest DuPraw to postulate the **folded-fiber model,** shown in Figure 2–12(c). During metaphase, each chromosome consists of two sister chromatids joined at the centromeric region. Each arm of the chromatid appears to be a single fiber wound much like a skein of yarn. The fiber is composed of tightly coiled double-stranded DNA and protein. An orderly coiling–twisting–condensing process appears to facilitate the transition of the interphase chromatin into the more condensed mitotic chromosomes. Geneticists believe that during the transition from interphase to prophase, a 5000-fold compaction occurs in the length of DNA within the chromatin fiber! This process must be extremely precise given the highly ordered and consistent appearance of mitotic chromosomes in all eukaryotes. Note particularly in the micrographs the clear distinction between the sister chromatids constituting each chromosome. They are joined only by the common centromere that they share prior to anaphase. We will return to this general topic later in the text when we consider chromosome structure in further detail (see Chapter 11).

ESSENTIAL POINT

Mitotic chromosomes are produced as a result of the coiling and condensation of chromatin fibers characteristic of interphase and are thus visible only during cell division. ■

PubMed: Exploring and Retrieving Biomedical Literature

MasteringGenetics™ Visit the Study Area: Exploring Genomics

PubMed is an Internet-based search system developed by the National Center of Biotechnology Information (NCBI) at the National Library of Medicine. Using PubMed, one can access over 23 million articles in over 5600 biomedical journals. The full text of many of the journals can be obtained electronically through college or university libraries, and some journals (such as *Proceedings of the National Academy of Sciences USA; Genome Biology;* and *Science*) provide free public access to articles within certain time frames.

In this exercise, we will explore PubMed to answer questions about relationships between tubulin, human cancers, and cancer therapies.

■ Exercise I – Tubulin, Cancer, and Mitosis

In this chapter we were introduced to tubulin and the dynamic behavior of microtubules during the cell cycle. Cancer cells are characterized by continuous and uncontrolled mitotic divisions.

Is it possible that tubulin and microtubules contribute to the development of cancer? Could these important structures be targets for cancer therapies?

1. To begin your search for the answers, access the PubMed site at http://www.ncbi.nlm.nih.gov/pubmed/.
2. In the search box, type "tubulin cancer" and then click the "Search" button to perform the search.
3. Select several research papers and read the abstracts.

To answer the question about tubulin's association with cancer, you may want to limit your search to fewer papers, perhaps those that are review articles. To do this, click the "Review" link under the Article Types category on the left side of the page.

Explore some of the articles, as abstracts or as full text, available in your library or by free public access. Prepare a brief report or verbally share your experiences with your class. Describe two of the most important things you learned during your exploration and identify the information sources you encountered during the search.

CASE STUDY | Triggering meiotic maturation of oocytes

A female athlete, who was training very hard, was concerned when her menstrual cycle stopped. Her doctor explained that her menstrual cycle is controlled by hormones that assess the energy status of her body before triggering ovulation. The availability of energy is measured by the concentration of leptin, an adipose-secreted hormone, which regulates the production of the luteinizing hormone (LH). The LH triggers meiotic maturation of one oocyte approximately every 28 days. As she was training very hard, her energy status had dropped below the level necessary to trigger meiotic maturation.

1. When does each meiotic division occur during oogenesis in human females?
2. Why is the meiotic maturation of oocytes linked to the energy status of a female's body?
3. Imagine that you are the athlete's doctor. What would you advise her to do to restore her menstrual cycle to normal?

INSIGHTS AND SOLUTIONS

This appearance of "Insights and Solutions" begins a feature that will have great value to you as a student. From this point on, "Insights and Solutions" precedes the "Problems and Discussion Questions" at each chapter's end to provide sample problems and solutions that demonstrate approaches you will find useful in genetic analysis. The insights you gain by working through the sample problems will improve your ability to solve the ensuing problems in each chapter.

1. In an organism with a diploid number of $2n = 6$, how many individual chromosomal structures will align on the metaphase plate during (a) mitosis, (b) meiosis I, and (c) meiosis II? Describe each configuration.

   Solution: (a) Remember that in mitosis, homologous chromosomes do not synapse, so there will be six double structures, each consisting of a pair of sister chromatids. In other words, the number of structures is equivalent to the diploid number.

   (b) In meiosis I, the homologs have synapsed, reducing the number of structures to three. Each is called a tetrad and consists of two pairs of sister chromatids.

   (c) In meiosis II, the same number of structures exist (three), but in this case they are called dyads. Each dyad is a pair of sister chromatids. When crossing over has occurred, each chromatid may contain parts of one of its nonsister chromatids, obtained during exchange in prophase I.

2. Disregarding crossing over, draw all possible alignment configurations that can occur during metaphase for the chromosomes shown in Figure 2–10.

   Solution: (a) As shown in the diagram below, four configurations are possible when $n = 2$.

3. For the chromosomes in the previous problem, assume that each of the larger chromosomes has a different allele for a given gene, *A* OR *a*, as shown. Also assume that each of the smaller chromosomes has a different allele for a second gene, *B* OR *b*. Calculate the probability of generating each possible combination of these alleles (*AB*, *Ab*, *aB*, *ab*) following meiosis I.

   Solution: (a) As shown in the accompanying diagram:

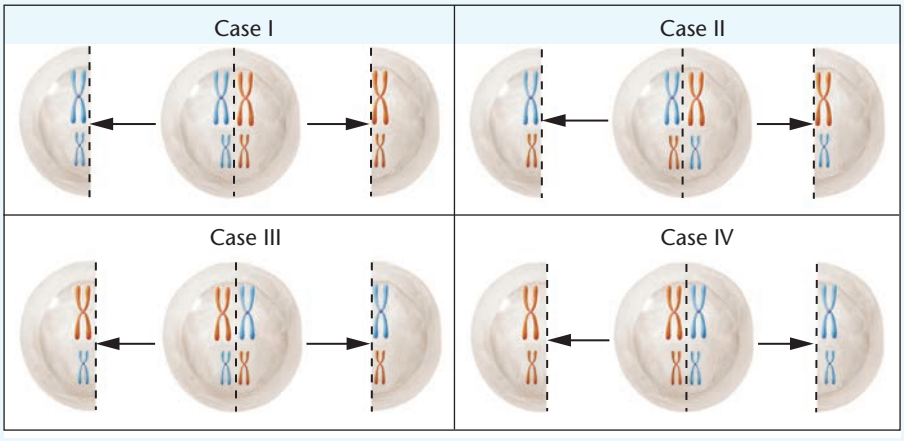
Case I	*AB* and *ab*
Case II	*Ab* and *aB*
Case III	*aB* and *Ab*
Case IV	*ab* and *AB*

   Total: $AB = 2\,(p = 1/4)$

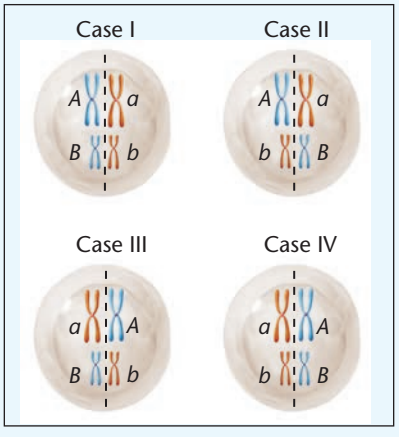
   $Ab = 2\,(p = 1/4)$

   $aB = 2\,(p = 1/4)$

   $ab = 2\,(p = 1/4)$



Solution for #2



Solution for #3

4. Describe the composition of a meiotic tetrad during prophase I, assuming no crossover event has occurred. What impact would a single crossover event have on this structure?

Solution: Such a tetrad contains four chromatids, existing as two pairs. Members of each pair are sister chromatids. They are held together by a common centromere. Members of one pair are maternally derived, whereas members of the other are paternally derived. Maternal and paternal members are called nonsister chromatids. A single crossover event has the effect of exchanging a portion of a maternal and a paternal chromatid, leading to a chiasma, where the two involved chromatids overlap physically in the tetrad. The process of exchange is referred to as crossing over.

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we focused on how chromosomes are distributed during cell division, both in dividing somatic cells (mitosis) and in gamete- and spore-forming cells (meiosis). We found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, answer the following questions.
(a) How do we know that chromosomes exist in homologous pairs?
(b) How do we know that DNA replication occurs during interphase, not early in mitosis?
(c) How do we know that mitotic chromosomes are derived from chromatin?

CONCEPT QUESTION

2. Review the Chapter Concepts list on page 28. All of these pertain to conceptual issues involving mitosis or meiosis. Based on these concepts, write a short essay that contrasts mitosis and meiosis, including their respective roles in organisms, the mechanisms by which they achieve their respective outcomes, and the consequences should either process fail to be executed with absolute fidelity. ■
3. What role do the following cellular components play in the storage, expression, or transmission of genetic information: (a) chromatin, (b) nucleolus, (c) ribosome, (d) mitochondrion, (e) centriole, (f) centromere?
4. Discuss the concepts of homologous chromosomes, diploidy, and haploidy. What characteristics do two homologous chromosomes share?
5. If two chromosomes of a species are the same length and have similar centromere placements and yet are not homologous, what is different about them?
6. Describe the events that characterize each stage of mitosis.
7. How do spindle fibers form and how do chromosomes separate in animal cells?
8. Compare chromosomal separation in plant and animal cells.
9. Why might different cells of the same organism have cell cycles of different durations?
10. Define and discuss these terms: (a) synapsis, (b) bivalents, (c) chiasmata, (d) crossing over, (e) chromomeres, (f) sister chromatids, (g) tetrads, (h) dyads, (i) monads.
11. Contrast the genetic content and the origin of sister versus nonsister chromatids during their earliest appearance in prophase I of meiosis. How might the genetic content of these change by the time tetrads have aligned at the equatorial plate during metaphase I?
12. Given the end results of the two types of division, why is it necessary for homologs to pair during meiosis and not desirable for them to pair during mitosis?
13. With increasing maternal age, the chances of observing trisomies increase significantly. Increasing paternal age is associated with *de novo* point mutations. Why?
14. How do the stages of mitosis and meiosis occur in a specific order and never alternate?
15. Trisomy 21 or Down syndrome occurs when there is a normal diploid chromosomal complement of 46 chromosomes plus one (extra) chromosome #21. Such individuals therefore have 47 chromosomes. Assume that a mating occurs between a female with Down syndrome and a normal 46-chromosome male. What proportion of the offspring would be expected to have Down syndrome? Justify your answer.
16. Considering the preceding problem, predict the number of different haploid cells that could be produced by meiosis if a fourth chromosome pair (W1 and W2) were added.
17. During oogenesis in an animal species with a haploid number of 6, one dyad undergoes nondisjunction during meiosis II. Following the second meiotic division, this dyad ends up intact in the ovum. How many chromosomes are present in (a) the mature ovum and (b) the second polar body? (c) Following fertilization by a normal sperm, what chromosome condition is created?
18. What is the probability that, in an organism with a haploid number of 10, a sperm will be formed that contains all 10 chromosomes whose centromeres were derived from maternal homologs?
19. During the first meiotic prophase, (a) when does crossing over occur; (b) when does synapsis occur; (c) during which stage are the chromosomes least condensed; and (d) when are chiasmata first visible?
20. Describe the role of meiosis in the life cycle of a vascular plant.
21. How many sister chromatids are seen in the metaphase for a single chromosome? How different are these structures from the interphase chromatin?
22. What is the significance of checkpoints in the cell cycle?
23. You are given a metaphase chromosome preparation (a slide) from an unknown organism that contains 12 chromosomes. Two that are clearly smaller than the rest appear identical in length and centromere placement. Describe all that you can about these chromosomes.
24. If one follows 50 primary oocytes in an animal through their various stages of oogenesis, how many secondary oocytes would be formed? How many first polar bodies would be formed? How many ootids would be formed? If one follows 50 primary spermatocytes in an animal through their various stages of spermatogenesis, how many secondary spermatocytes would be formed? How many spermatids would be formed?

25. Cell division cycle mutations render the mutants unable to continue the cell cycle. This phenotype creates a paradox where mutant cells must also be grown in the lab to further identify the gene and study the role of the protein. How do you think this problem can be solved?

For Problems 26–31, consider a diploid cell that contains three pairs of chromosomes designated AA, BB, and CC. Each pair contains a maternal and a paternal member (e.g., A^m and A^p). Using these designations, demonstrate your understanding of mitosis and meiosis by drawing chromatid combinations as requested. Be sure to indicate when chromatids are paired as a result of replication and/or synapsis. You may wish to use a large piece of brown manila wrapping paper or a cut-up paper grocery bag for this project and to work in partnership with another student. We recommend cooperative learning as an efficacious way to develop the skills you will need for solving the problems presented throughout this text.

26. In mitosis, what chromatid combination(s) will be present during metaphase? What combination(s) will be present at each pole at the completion of anaphase?
27. During meiosis I, assuming no crossing over, what chromatid combination(s) will be present at the completion of prophase? Draw all possible alignments of chromatids as migration begins during early anaphase.
28. Are any possible combinations present during prophase of meiosis II other than those that you drew in Problem 27? If so, draw them.
29. Draw all possible combinations of chromatids during the early phases of anaphase in meiosis II.
30. Assume that during meiosis I none of the C chromosomes disjoin at metaphase, but they separate into dyads (instead of monads) during meiosis II. How would this change the alignments that you constructed during the anaphase stages in meiosis I and II? Draw them.
31. Assume that each gamete resulting from Problem 30 fuses, in fertilization, with a normal haploid gamete. What combinations will result? What percentage of zygotes will be diploid, containing one paternal and one maternal member of each chromosome pair?

3 Mendelian Genetics

CHAPTER CONCEPTS

- Inheritance is governed by information stored in discrete factors called genes.
- Genes are transmitted from generation to generation on vehicles called chromosomes.
- Chromosomes, which exist in pairs in diploid organisms, provide the basis of biparental inheritance.
- During gamete formation, chromosomes are distributed according to postulates first described by Gregor Mendel, based on his nineteenth-century research with the garden pea.
- Mendelian postulates prescribe that homologous chromosomes segregate from one another and assort independently with other segregating homologs during gamete formation.
- Genetic ratios, expressed as probabilities, are subject to chance deviation and may be evaluated statistically.
- The analysis of pedigrees allows predictions concerning the genetic nature of human traits.

Gregor Johann Mendel, who in 1866 put forward the major postulates of transmission genetics as a result of experiments with the garden pea.

Although inheritance of biological traits has been recognized for thousands of years, the first significant insights into how it takes place only occurred about 150 years ago. In 1866, Gregor Johann Mendel published the results of a series of experiments that would lay the foundation for the formal discipline of genetics. Mendel's work went largely unnoticed until the turn of the twentieth century, but eventually, the concept of the gene as a distinct hereditary unit was established. Since then, the ways in which genes, as segments of chromosomes, are transmitted to offspring and control traits have been clarified. Research continued unabated throughout the twentieth century and into the present—indeed, studies in genetics, most recently at the molecular level, have remained at the forefront of biological research since the early 1900s.

When Mendel began his studies of inheritance using *Pisum sativum,* the garden pea, chromosomes and the role and mechanism of meiosis were totally unknown. Nevertheless, he determined that discrete *units of inheritance* exist and predicted their behavior in the formation of gametes. Subsequent investigators, with access to cytological data, were able to relate their own observations of chromosome behavior during meiosis and Mendel's principles of inheritance. Once this correlation was recognized, Mendel's postulates were accepted as the basis for the study of what is known as **transmission genetics**—how genes are transmitted from parents to offspring. These principles were derived directly from Mendel's experimentation.

3.1 Mendel Used a Model Experimental Approach to Study Patterns of Inheritance

Johann Mendel was born in 1822 to a peasant family in the Central European village of Heinzendorf. An excellent student in high school, he studied philosophy for several years afterward and in 1843, taking the name Gregor, was admitted to the Augustinian Monastery of St. Thomas in Brno, now part of the Czech Republic. In 1849, he was relieved of pastoral duties, and from 1851 to 1853, he attended the University of Vienna, where he studied physics and botany. He returned to Brno in 1854, where he taught physics and natural science for the next 16 years. Mendel received support from the monastery for his studies and research throughout his life.

In 1856, Mendel performed his first set of hybridization experiments with the garden pea, launching the research phase of his career. His experiments continued until 1868, when he was elected abbot of the monastery. Although he retained his interest in genetics, his new responsibilities demanded most of his time. In 1884, Mendel died of a kidney disorder. The local newspaper paid him the following tribute:

> "His death deprives the poor of a benefactor, and mankind at large of a man of the noblest character, one who was a warm friend, a promoter of the natural sciences, and an exemplary priest."

Mendel first reported the results of some simple genetic crosses between certain strains of the garden pea in 1865. Although his was not the first attempt to provide experimental evidence pertaining to inheritance, Mendel's success where others had failed can be attributed, at least in part, to his elegant experimental design and analysis.

Mendel showed remarkable insight into the methodology necessary for good experimental biology. First, he chose an organism that was easy to grow and to hybridize artificially. The pea plant is self-fertilizing in nature, but it is easy to cross-breed experimentally. It reproduces well and grows to maturity in a single season. Mendel followed seven visible features (we refer to them as characters, or characteristics), each represented by two contrasting forms, or **traits** (**Figure 3–1**). For the character stem height, for example, he experimented with the traits *tall* and *dwarf*. He selected six other visibly contrasting pairs of traits involving seed shape and color, pod shape and color, and flower color and position. From local seed merchants, Mendel obtained true-breeding strains, those in which each trait appeared unchanged generation after generation in self-fertilizing plants.

There were several other reasons for Mendel's success. In addition to his choice of a suitable organism, he restricted his examination to one or very few pairs of contrasting traits in each experiment. He also kept accurate quantitative records, a necessity in genetic experiments. From the analysis of his data, Mendel derived certain postulates that have become the principles of transmission genetics.

3.2 The Monohybrid Cross Reveals How One Trait Is Transmitted from Generation to Generation

Mendel's simplest crosses involved only one pair of contrasting traits. Each such experiment is called a **monohybrid cross.** A monohybrid cross is made by mating true-breeding individuals from two parent strains, each exhibiting one of the two contrasting forms of the character under study. Initially, we examine the first generation of offspring of such a cross, and then we consider the offspring of **selfing,** that is, of self-fertilization of individuals from this first generation. The original parents constitute the $\mathbf{P_1}$, or **parental generation;** their offspring are the $\mathbf{F_1}$, or **first filial generation;** the individuals resulting from the selfed $\mathbf{F_1}$ generation are the $\mathbf{F_2}$, or **second filial generation;** and so on.

The cross between true-breeding pea plants with tall stems and dwarf stems is representative of Mendel's monohybrid crosses. *Tall* and *dwarf* are contrasting traits of the character of stem height. Unless tall or dwarf plants are crossed together or with another strain, they will undergo self-fertilization and breed true, producing their respective traits generation after generation. However, when Mendel crossed tall plants with dwarf plants, the resulting F_1 generation consisted of only tall plants. When members of the F_1 generation were selfed, Mendel observed that 787 of 1064 F_2 plants were tall, while 277 of 1064 were dwarf. Note that in this cross (Figure 3–1), the dwarf trait disappeared in the F_1 generation, only to reappear in the F_2 generation.

Genetic data are usually expressed and analyzed as ratios. In this particular example, many identical P_1 crosses were made and many F_1 plants—all tall—were produced. As noted, of the 1064 F_2 offspring, 787 were tall and 277 were dwarf—a ratio of approximately 2.8:1.0, or about 3:1.

Mendel made similar crosses between pea plants exhibiting each of the other pairs of contrasting traits; the results of these crosses are shown in Figure 3–1. In every case, the outcome was similar to the tall/dwarf cross just described. For the character of interest, all F_1 offspring had

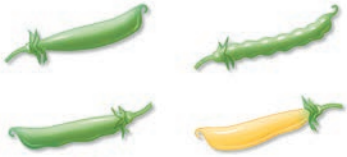
Character	Contrasting traits		F_1 results	F_2 results	F_2 ratio
Seed shape	round/wrinkled		all round	5474 round 1850 wrinkled	2.96:1
Seed color	yellow/green		all yellow	6022 yellow 2001 green	3.01:1
Pod shape	full/constricted		all full	882 full 299 constricted	2.95:1
Pod color	green/yellow		all green	428 green 152 yellow	2.82:1
Flower color	violet/white		all violet	705 violet 224 white	3.15:1
Flower position	axial/terminal		all axial	651 axial 207 terminal	3.14:1
Stem height	tall/dwarf		all tall	787 tall 277 dwarf	2.84:1

FIGURE 3–1 Seven pairs of contrasting traits and the results of Mendel's seven monohybrid crosses of the garden pea (*Pisum sativum*). In each case, pollen derived from plants exhibiting one trait was used to fertilize the ova of plants exhibiting the other trait. In the F_1 generation, one of the two traits was exhibited by all plants. The contrasting trait reappeared in approximately 1/4 of the F_2 plants.

the same trait exhibited by one of the parents, but in the F_2 offspring, an approximate ratio of 3:1 was obtained. That is, three-fourths looked like the F_1 plants, while one-fourth exhibited the contrasting trait, which had disappeared in the F_1 generation.

We note one further aspect of Mendel's monohybrid crosses. In each cross, the F_1 and F_2 patterns of inheritance were similar regardless of which P_1 plant served as the source of pollen (sperm) and which served as the source of the ovum (egg). The crosses could be made either way—pollination of dwarf plants by tall plants, or vice versa. Crosses made in both these ways are called **reciprocal crosses.** Therefore, the results of Mendel's monohybrid crosses were not sex-dependent.

To explain these results, Mendel proposed the existence of particulate *unit factors* for each trait. He suggested that these factors serve as the basic units of heredity and are passed unchanged from generation to generation, determining various traits expressed by each individual plant. Using these general ideas, Mendel proceeded to hypothesize precisely how such factors could account for the results of the monohybrid crosses.

Mendel's First Three Postulates

Using the consistent pattern of results in the monohybrid crosses, Mendel derived the following three postulates, or principles, of inheritance.

1. UNIT FACTORS IN PAIRS

Genetic characters are controlled by unit factors existing in pairs in individual organisms.

In the monohybrid cross involving tall and dwarf stems, a specific **unit factor** exists for each trait. Each diploid individual receives one factor from each parent. Because the factors occur in pairs, three combinations are possible: two factors for tall stems, two factors for dwarf stems, or one of each factor. Every individual possesses one of these three combinations, which determines stem height.

2. DOMINANCE/RECESSIVENESS

When two unlike unit factors responsible for a single character are present in a single individual, one unit factor is dominant to the other, which is said to be recessive.

In each monohybrid cross, the trait expressed in the F_1 generation is controlled by the dominant unit factor.

The trait not expressed is controlled by the recessive unit factor. The terms dominant and recessive are also used to designate traits. In this case, tall stems are said to be dominant over recessive dwarf stems.

3. SEGREGATION

During the formation of gametes, the paired unit factors separate, or segregate, randomly so that each gamete receives one or the other with equal likelihood.

If an individual contains a pair of like unit factors (e.g., both specific for tall), then all its gametes receive one of that same kind of unit factor (in this case, tall). If an individual contains unlike unit factors (e.g., one for tall and one for dwarf), then each gamete has a 50 percent probability of receiving either the tall or the dwarf unit factor.

These postulates provide a suitable explanation for the results of the monohybrid crosses. Let's use the tall/dwarf cross to illustrate. Mendel reasoned that P_1 tall plants contained identical paired unit factors, as did the P_1 dwarf plants. The gametes of tall plants all receive one tall unit factor as a result of **segregation.** Similarly, the gametes of dwarf plants all receive one dwarf unit factor. Following fertilization, all F_1 plants receive one unit factor from each parent—a tall factor from one and a dwarf factor from the other—reestablishing the paired relationship, but because tall is dominant to dwarf, all F_1 plants are tall.

When F_1 plants form gametes, the postulate of segregation demands that each gamete randomly receives either the tall *or* dwarf unit factor. Following random fertilization events during F_1 selfing, four F_2 combinations will result with equal frequency:

1. tall/tall
2. tall/dwarf
3. dwarf/tall
4. dwarf/dwarf

Combinations (1) and (4) will clearly result in tall and dwarf plants, respectively. According to the postulate of dominance/recessiveness, combinations (2) and (3) will both yield tall plants. Therefore, the F_2 is predicted to consist of 3/4 tall and 1/4 dwarf, or a ratio of 3:1. This is approximately what Mendel observed in his cross between tall and dwarf plants. A similar pattern was observed in each of the other monohybrid crosses (Figure 3–1).

ESSENTIAL POINT

Mendel's postulates help describe the basis for the inheritance of phenotypic traits. He hypothesized that unit factors exist in pairs and exhibit a dominant/recessive relationship in determining the expression of traits. He further postulated that unit factors segregate during gamete formation, such that each gamete receives one or the other factor, with equal probability. ■

Modern Genetic Terminology

To analyze the monohybrid cross and Mendel's first three postulates, we must first introduce several new terms as well as a symbol convention for the unit factors. Traits such as tall or dwarf are physical expressions of the information contained in unit factors. The physical expression of a trait is the **phenotype** of the individual. Mendel's unit factors represent units of inheritance called **genes** by modern geneticists. For any given character, such as plant height, the phenotype is determined by alternative forms of a single gene, called **alleles.** For example, the unit factors representing tall and dwarf are alleles determining the height of the pea plant.

Geneticists have several different systems for using symbols to represent genes. Later in the text (see Chapter 4), we will review a number of these conventions, but for now, we will adopt one to use consistently throughout this chapter. According to this convention, the first letter of the recessive trait symbolizes the character in question; in lowercase italic, it designates the allele for the recessive trait, and in uppercase italic, it designates the allele for the dominant trait. Thus for Mendel's pea plants, we use *d* for the *d*warf allele and *D* for the tall allele. When alleles are written in pairs to represent the two unit factors present in any individual (*DD*, *Dd*, or *dd*), the resulting symbol is called the **genotype.** The genotype designates the genetic makeup of an individual for the trait or traits it describes, whether the individual is haploid or diploid. By reading the genotype, we know the phenotype of the individual: *DD* and *Dd* are tall, and *dd* is dwarf. When both alleles are the same (*DD* or *dd*), the individual is **homozygous** for the trait, or a **homozygote;** when the alleles are different (*Dd*), we use the terms **heterozygous** and **heterozygote.** These symbols and terms are used in **Figure 3–2** to describe the monohybrid cross.

Punnett Squares

The genotypes and phenotypes resulting from combining gametes during fertilization can be easily visualized by constructing a diagram called a **Punnett square,** named after the person who first devised this approach, Reginald C. Punnett. **Figure 3–3** illustrates this method of analysis for our $F_1 \times F_1$ monohybrid cross. Each of the possible gametes is assigned a column or a row; the vertical columns represent those of the female parent, and the horizontal rows represent those of the male parent. After assigning the gametes to the rows and columns, we predict the new generation by entering the male and female gametic

information into each box and thus producing every possible resulting genotype. By filling out the Punnett square, we are listing all possible random fertilization events. The genotypes and phenotypes of all potential offspring are ascertained by reading the combinations in the boxes.

The Punnett square method is particularly useful when you are first learning about genetics and how to solve genetics problems. Note the ease with which the 3:1 phenotypic ratio and the 1:2:1 genotypic ratio may be derived for the F_2 generation in Figure 3–3.

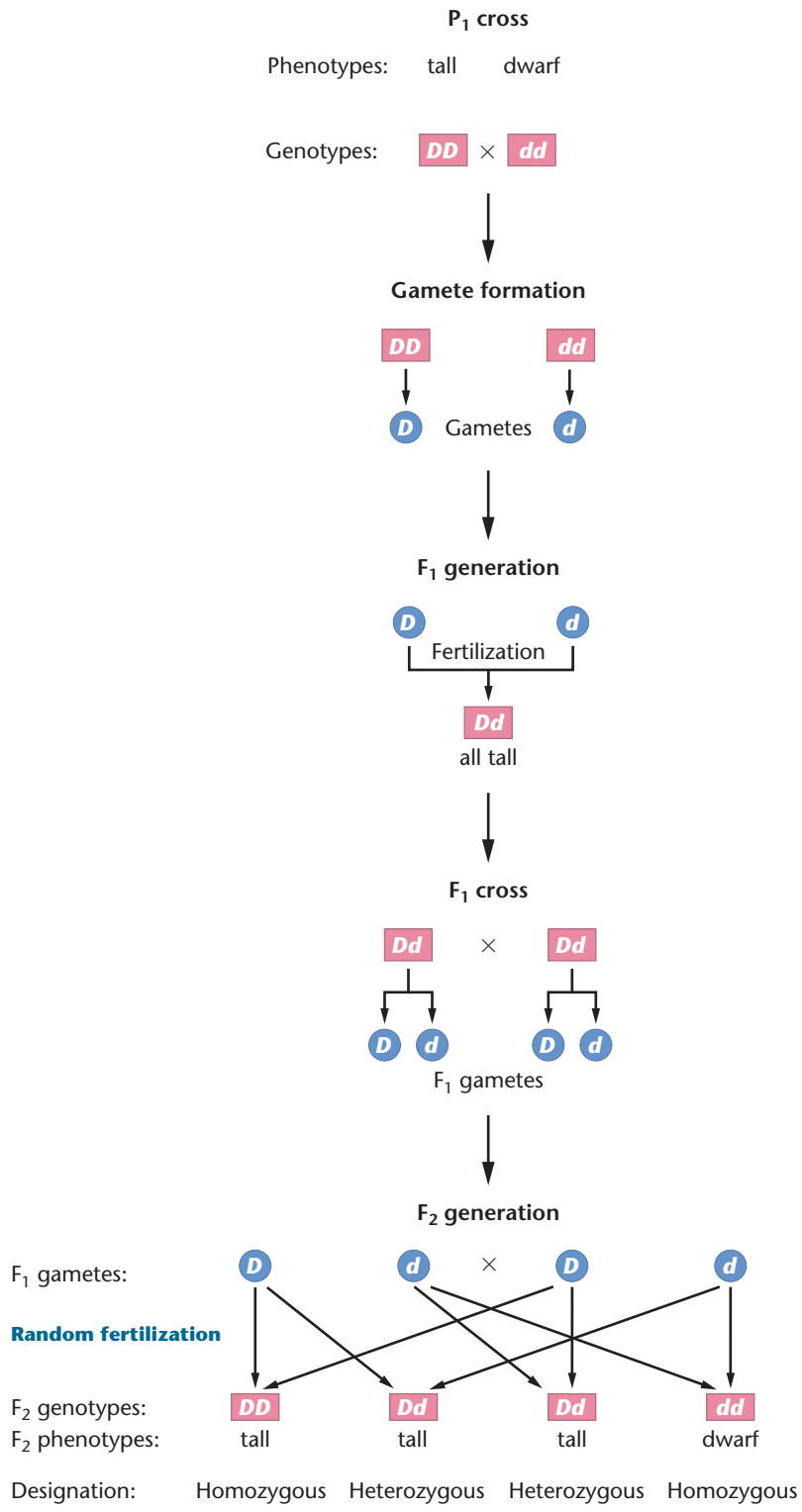


FIGURE 3–2 The monohybrid cross between tall (*D*) and dwarf (*d*) pea plants. Individuals are shown in rectangles, and gametes are shown in circles.

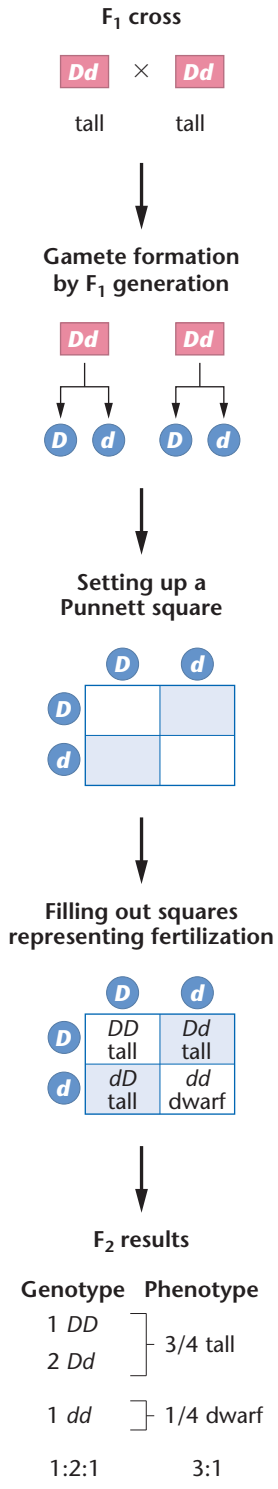


FIGURE 3–3 A Punnett square generating the F_2 ratio of the $F_1 \times F_1$ cross shown in Figure 3–2.

NOW SOLVE THIS

3–1 Pigeons may exhibit a checkered or plain color pattern. In a series of controlled matings, the following data were obtained.

	F_1 Progeny	
P_1 Cross	Checkered	Plain
(a) checkered × checkered	36	0
(b) checkered × plain	38	0
(c) plain × plain	0	35

Then F_1 offspring were selectively mated with the following results. (The P_1 cross giving rise to each F_1 pigeon is indicated in parentheses.)

	F_2 Progeny	
F_1 × F_1 Crosses	Checkered	Plain
(d) checkered (a) × plain (c)	34	0
(e) checkered (b) × plain (c)	17	14
(f) checkered (b) × checkered (b)	28	9
(g) checkered (a) × checkered (b)	39	0

How are the checkered and plain patterns inherited? Select and assign symbols for the genes involved, and determine the genotypes of the parents and offspring in each cross.

HINT: *This problem asks you to analyze the data produced from several crosses involving pigeons and to determine the mode of inheritance and the genotypes of the parents and offspring in a number of instances. The key to its solution is to first determine whether or not this is a monohybrid cross. To do so, convert the data to ratios that are characteristic of Mendelian crosses. In the case of this problem, ask first whether any of the F_2 ratios match Mendel's 3:1 monohybrid ratio. If so, the second step is to determine which trait is dominant and which is recessive.*

The Testcross: One Character

Tall plants produced in the F_2 generation are predicted to have either the *DD* or the *Dd* genotype. You might ask if there is a way to distinguish the genotype. Mendel devised a rather simple method that is still used today to discover the genotype of plants and animals: the **testcross.** The organism expressing the dominant phenotype but having an unknown genotype is crossed with a known *homozygous recessive individual.* For example, as shown in **Figure 3–4(a)**, if a tall plant of genotype *DD* is testcrossed with a dwarf plant, which must have the *dd* genotype, all offspring will be tall phenotypically and *Dd* genotypically. However, as shown in **Figure 3–4(b)**, if a tall plant is *Dd* and is crossed with a dwarf plant (*dd*), then one-half of the offspring will be tall (*Dd*) and the other half will be dwarf (*dd*).

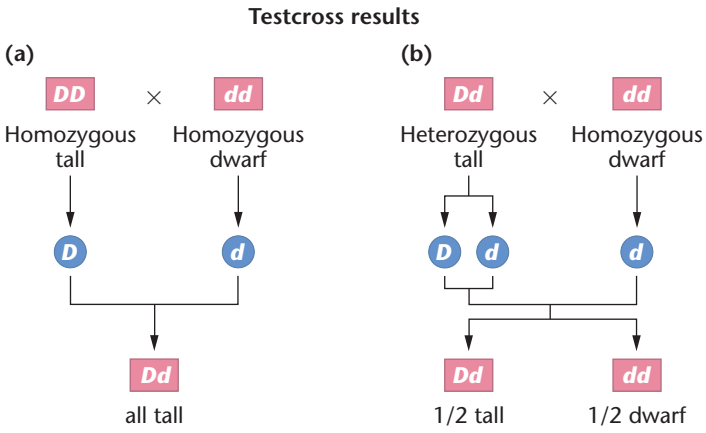


FIGURE 3–4 Testcross of a single character. In (a), the tall parent is homozygous, but in (b), the tall parent is heterozygous. The genotype of each tall P_1 plant can be determined by examining the offspring when each is crossed with the homozygous recessive dwarf plant.

Therefore, a 1:1 tall/dwarf ratio demonstrates the heterozygous nature of the tall plant of unknown genotype. The results of the testcross reinforced Mendel's conclusion that separate unit factors control traits.

3.3 Mendel's Dihybrid Cross Generated a Unique F_2 Ratio

As a natural extension of the monohybrid cross, Mendel also designed experiments in which he examined two characters simultaneously. Such a cross, involving two pairs of contrasting traits, is a **dihybrid cross,** or a *two-factor cross.* For example, if pea plants having yellow seeds that are round were bred with those having green seeds that are wrinkled, the results shown in **Figure 3–5** would occur: the F_1 offspring would all be yellow and round. It is therefore apparent that yellow is dominant to green and that round is dominant to wrinkled. When the F_1 individuals are selfed, approximately 9/16 of the F_2 plants express the yellow and round traits, 3/16 express yellow and wrinkled, 3/16 express green and round, and 1/16 express green and wrinkled.

A variation of this cross is also shown in Figure 3–5. Instead of crossing one P_1 parent with both dominant traits (yellow, round) to one with both recessive traits (green, wrinkled), plants with yellow, wrinkled seeds are crossed with those with green, round seeds. In spite of the change in the P_1 phenotypes, both the F_1 and F_2 results remain unchanged. Why this is so will become clear in the next section.

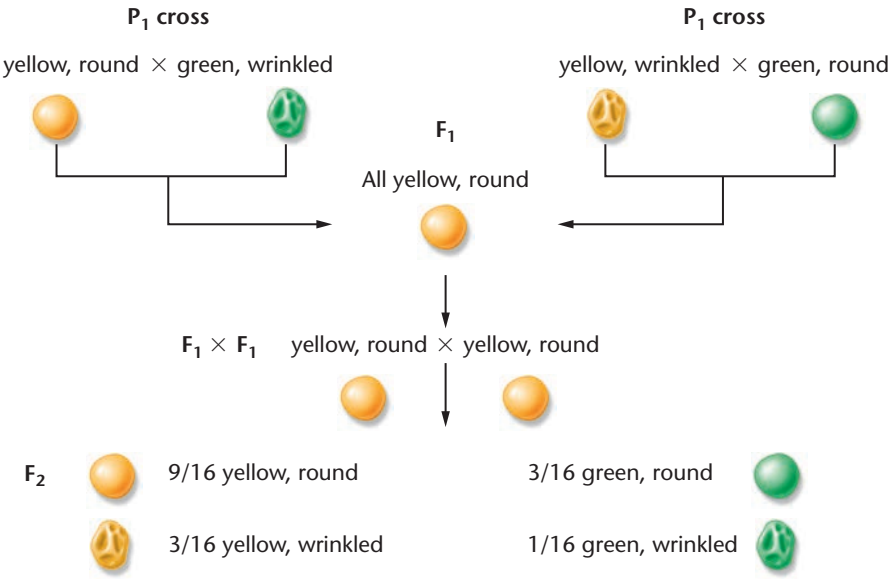


FIGURE 3–5 F_1 and F_2 results of Mendel's dihybrid crosses in which the plants on the top left with yellow, round seeds are crossed with plants having green, wrinkled seeds, and the plants on the top right with yellow, wrinkled seeds are crossed with plants having green, round seeds.

Mendel's Fourth Postulate: Independent Assortment

We can most easily understand the results of a dihybrid cross if we consider it theoretically as consisting of two monohybrid crosses conducted separately. Think of the two sets of traits as being inherited independently of each other; that is, the chance of any plant having yellow or green seeds is not at all influenced by the chance that this plant will have round or wrinkled seeds. Thus, because yellow is dominant to green, all F_1 plants in the first theoretical cross would have yellow seeds. In the second theoretical cross, all F_1 plants would have round seeds because round is dominant to wrinkled. When Mendel examined the F_1 plants of the dihybrid cross, all were yellow and round, as our theoretical crosses predict.

The predicted F_2 results of the first cross are 3/4 yellow and 1/4 green. Similarly, the second cross would yield 3/4 round and 1/4 wrinkled. Figure 3–5 shows that in the dihybrid cross, 12/16 F_2 plants are yellow, while 4/16 are green, exhibiting the expected 3:1 (3/4:1/4) ratio. Similarly, 12/16 of all F_2 plants have round seeds, while 4/16 have wrinkled seeds, again revealing the 3:1 ratio.

These numbers demonstrate that the two pairs of contrasting traits are inherited independently, so we can predict the frequencies of all possible F_2 phenotypes by applying the **product law** of probabilities: *When two independent events occur simultaneously, the probability of the two outcomes occurring in combination is equal to the product of their individual probabilities of occurrence.* For example, the probability of an F_2 plant having yellow and round seeds is (3/4)(3/4), or 9/16, because 3/4 of all F_2 plants should be yellow and 3/4 of all F_2 plants should be round.

In a like manner, the probabilities of the other three F_2 phenotypes can be calculated: yellow (3/4) and wrinkled (1/4) are predicted to be present together 3/16 of the time; green (1/4) and round (3/4) are predicted 3/16 of the time; and green (1/4) and wrinkled (1/4) are predicted 1/16 of the time. These calculations are shown in **Figure 3–6**.

It is now apparent why the F_1 and F_2 results are identical whether the initial cross is yellow, round plants bred with green, wrinkled plants, or whether yellow, wrinkled plants are bred with green, round plants. In both crosses,

F_1 yellow, round × yellow, round			
F_2	Of all offspring	Of all offspring	Combined probabilities
3/4 are yellow		3/4 are round and	(3/4)(3/4) = 9/16 yellow, round
		1/4 are wrinkled	(3/4)(1/4) = 3/16 yellow, wrinkled
1/4 are green		3/4 are round and	(1/4)(3/4) = 3/16 green, round
		1/4 are wrinkled	(1/4)(1/4) = 1/16 green, wrinkled

FIGURE 3–6 Computation of the combined probabilities of each F_2 phenotype for two independently inherited characters. The probability of each plant being yellow or green is independent of the probability of it bearing round or wrinkled seeds.

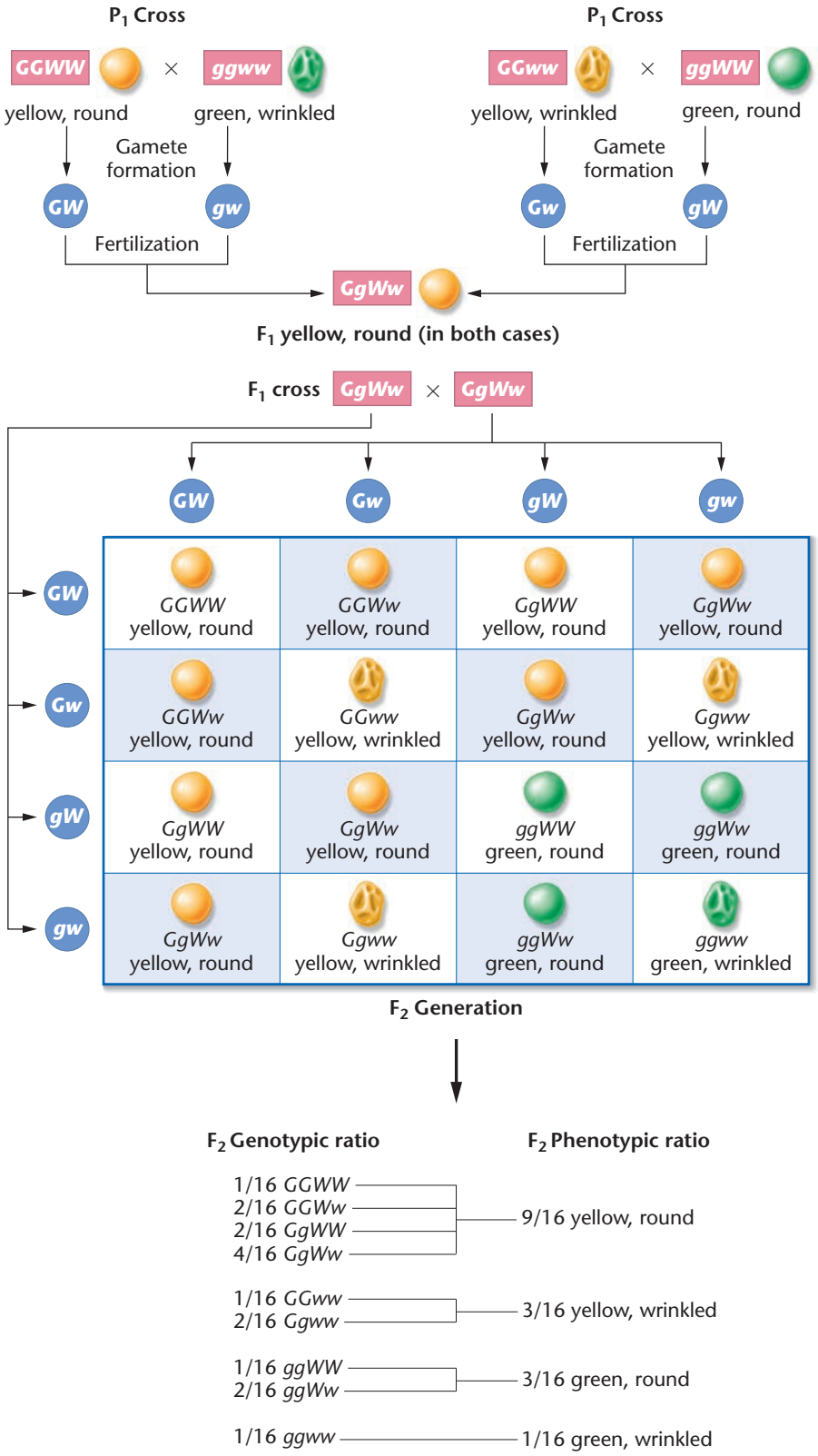


FIGURE 3–7 Analysis of the dihybrid crosses shown in Figure 3–5. The F_1 heterozygous plants are self-fertilized to produce an F_2 generation, which is computed using a Punnett square. Both the phenotypic and genotypic F_2 ratios are shown.

the F_1 genotype of all offspring is identical. As a result, the F_2 generation is also identical in both crosses.

On the basis of similar results in numerous dihybrid crosses, Mendel proposed a fourth postulate:

> **4. INDEPENDENT ASSORTMENT**
> *During gamete formation, segregating pairs of unit factors assort independently of each other.*
> This postulate stipulates that segregation of any pair of unit factors occurs independently of all others. As a result of random segregation, each gamete receives one member of every pair of unit factors. For one pair, whichever unit factor is received does not influence the outcome of segregation of any other pair. Thus, according to the postulate of independent assortment, all possible combinations of gametes should be formed in equal frequency.

The Punnett square in **Figure 3–7** shows how independent assortment works in the formation of the F_2 generation. Examine the formation of gametes by the F_1 plants; segregation prescribes that every gamete receives either a *G* or *g* allele and a *W* or *w* allele. Independent assortment stipulates that all four combinations (*GW*, *Gw*, *gW*, and *gw*) will be formed with equal probabilities.

In every $F_1 \times F_1$ fertilization event, each zygote has an equal probability of receiving one of the four combinations from each parent. If many offspring are produced, 9/16 have yellow, round seeds, 3/16 have yellow, wrinkled seeds, 3/16 have green, round seeds, and 1/16 have green, wrinkled seeds, yielding what is designated as **Mendel's 9:3:3:1 dihybrid ratio.** This is an ideal ratio based on probability events involving segregation, independent assortment, and random fertilization. Because of deviation due strictly to chance, particularly if small numbers of offspring are produced, actual results are highly unlikely to match the ideal ratio.

ESSENTIAL POINT

Mendel's postulate of independent assortment states that each pair of unit factors segregates independently of other such pairs. As a result, all possible combinations of gametes are formed with equal probability. ■

3.4 The Trihybrid Cross Demonstrates That Mendel's Principles Apply to Inheritance of Multiple Traits

Thus far, we have considered inheritance of up to two pairs of contrasting traits. Mendel demonstrated that the processes of segregation and independent assortment also apply to three pairs of contrasting traits, in what is called a **trihybrid cross,** or *three-factor cross.*

Although a trihybrid cross is somewhat more complex than a dihybrid cross, its results are easily calculated if the principles of segregation and independent assortment are followed. For example, consider the cross shown in **Figure 3–8** where the gene pairs of theoretical contrasting traits are represented by the symbols *A, a, B, b, C,* and *c.*

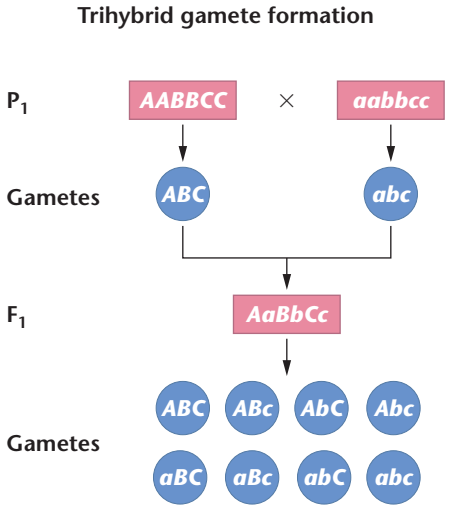


FIGURE 3–8 Formation of P_1 and F_1 gametes in a trihybrid cross.

NOW SOLVE THIS

3–2 Considering the Mendelian traits round versus wrinkled and yellow versus green, consider the crosses below and determine the genotypes of the parental plants by analyzing the phenotypes of their offspring.

Parental Plants	Offspring
(a) round, yellow × round, yellow	3/4 round, yellow 1/4 wrinkled, yellow
(b) wrinkled, yellow × round, yellow	6/16 wrinkled, yellow 2/16 wrinkled, green 6/16 round, yellow 2/16 round, green
(c) round, yellow × round, yellow	9/16 round, yellow 3/16 round, green 3/16 wrinkled, yellow 1/16 wrinkled, green
(d) round, yellow × wrinkled, green	1/4 round, yellow 1/4 round, green 1/4 wrinkled, yellow 1/4 wrinkled, green

■ **HINT:** *This problem involves a series of Mendelian dihybrid crosses where you are asked to determine the genotypes of the parents in a number of instances. The key to its solution is to write down everything that you know for certain. This reduces the problem to its bare essentials, clarifying what you need to determine. For example, the wrinkled, yellow plant in case (b) must be homozygous for the recessive wrinkled alleles and bear at least one dominant allele for the yellow trait. Having established this, you need only determine the remaining allele for cotyledon color.*

In the cross between *AABBCC* and *aabbcc* individuals, all F_1 individuals are heterozygous for all three gene pairs. Their genotype, *AaBbCc*, results in the phenotypic expression of the dominant *A, B*, and *C* traits. When F_1 individuals serve as parents, each produces eight different gametes in equal frequencies. At this point, we could construct a Punnett square with 64 separate boxes and read out the phenotypes—but such a method is cumbersome in a cross involving so many factors. Therefore, another method has been devised to calculate the predicted ratio.

The Forked-Line Method, or Branch Diagram

It is much less difficult to consider each contrasting pair of traits separately and then to combine these results by using the **forked-line method,** first shown in Figure 3–6. This method, also called a **branch diagram,** relies on the simple application of the laws of probability established for the dihybrid cross. Each gene pair is assumed to behave independently during gamete formation.

When the monohybrid cross AA × *aa* is made, we know that:

1. All F_1 individuals have the genotype *Aa* and express the phenotype represented by the *A* allele, which is called the *A* phenotype in the discussion that follows.
2. The F_2 generation consists of individuals with either the *A* phenotype or the *a* phenotype in the ratio of 3:1.

The same generalizations can be made for the *BB* × *bb* and *CC* × *cc* crosses. Thus, in the F_2 generation, 3/4 of all organisms will express phenotype *A*, 3/4 will express *B,* and 3/4 will express *C*. Similarly, 1/4 of all organisms will express *a*, 1/4 will express *b*, and 1/4 will express *c*. The proportions of organisms that express each phenotypic combination can be predicted by assuming that fertilization, following the independent assortment of these three gene pairs during gamete formation, is a random process. We apply the product law of probabilities once again. **Figure 3–9** uses the forked-line method to calculate the phenotypic proportions of the F_2 generation. They fall into the trihybrid ratio of 27:9:9:9:3:3:3:1. The same method can be used to solve crosses involving any number of gene pairs, *provided that all gene pairs assort independently from each other.* We shall see later that gene pairs do not always assort with complete independence. However, it appeared to be true for all of Mendel's characters.

ESSENTIAL POINT

The forked-line method is less complex than, but just as accurate as, the Punnett square in predicting the probabilities of phenotypes or genotypes from crosses involving two or more gene pairs. ■

NOW SOLVE THIS

3–3 Using the forked-line, or branch diagram, method, determine the genotypic and phenotypic ratios of these trihybrid crosses: (a) *AaBbCc* × *AaBBCC*, (b) *AaBBCc* × *aaBBCc*, and (c) *AaBbCc* × *AaBbCc*.

■ **HINT:** *This problem asks you to use the forked-line method to determine the outcome of a number of trihybrid crosses. The key to its solution is to realize that in using the forked-line method, you must consider each gene pair separately. For example, in this problem, first predict the outcome of each cross for the A /a genes, then for the B/b genes, and finally, for the C/c genes. Then you are prepared to pursue the outcome of each cross using the forked-line method.*

Generation of F_2 trihybrid phenotypes

A or *a*	*B* or *b*	*C* or *c*	Combined proportion
3/4 *A*	3/4 *B*	3/4 *C*	(3/4)(3/4)(3/4) *ABC* = 27/64 *ABC*
		1/4 *c*	(3/4)(3/4)(1/4) *ABc* = 9/64 *ABc*
	1/4 *b*	3/4 *C*	(3/4)(1/4)(3/4) *AbC* = 9/64 *AbC*
		1/4 *c*	(3/4)(1/4)(1/4) *Abc* = 3/64 *Abc*
1/4 *a*	3/4 *B*	3/4 *C*	(1/4)(3/4)(3/4) *aBC* = 9/64 *aBC*
		1/4 *c*	(1/4)(3/4)(1/4) *aBc* = 3/64 *aBc*
	1/4 *b*	3/4 *C*	(1/4)(1/4)(3/4) *abC* = 3/64 *abC*
		1/4 *c*	(1/4)(1/4)(1/4) *abc* = 1/64 *abc*

FIGURE 3–9 Generation of the F_2 trihybrid phenotypic ratio using the forked-line method. This method is based on the expected probability of occurrence of each phenotype.

3.5 Mendel's Work Was Rediscovered in the Early Twentieth Century

Mendel published his work in 1866. While his findings were often cited and discussed, their significance went unappreciated for about 35 years. Then, in the latter part of the nineteenth century, a remarkable observation set the scene for the recognition of Mendel's work: Walter Flemming's discovery of chromosomes in the nuclei of salamander cells. In 1879, Flemming described the behavior of these thread-like structures during cell division. As a result of his findings and the work of many other cytologists, the presence of discrete units within the nucleus soon became an integral part of scientists' ideas about inheritance.

In the early twentieth century, hybridization experiments similar to Mendel's were performed independently by three botanists, Hugo de Vries, Carl Correns, and Erich Tschermak. De Vries's work demonstrated the principle of segregation in several plant species. Apparently, he searched the existing literature and found that Mendel's work had anticipated his own conclusions! Correns and Tschermak also reached conclusions similar to those of Mendel.

About the same time, two cytologists, Walter Sutton and Theodor Boveri, independently published papers linking their discoveries of the behavior of chromosomes during meiosis to the Mendelian principles of segregation and independent assortment. They pointed out that the separation of chromosomes during meiosis could serve as the cytological basis of these two postulates. Although they thought that Mendel's unit factors were probably chromosomes rather than genes on chromosomes, their findings reestablished the importance of Mendel's work and led to many ensuing genetic investigations. Sutton and Boveri are credited with initiating the **chromosomal theory of inheritance,** the idea that the genetic material in living organisms is contained in chromosomes, which was developed during the next two decades. As we will see in subsequent chapters, work by Thomas H. Morgan, Alfred H. Sturtevant, Calvin Bridges, and others established beyond a reasonable doubt that Sutton's and Boveri's hypothesis was correct.

ESSENTIAL POINT

The discovery of chromosomes in the late 1800s, along with subsequent studies of their behavior during meiosis, led to the rebirth of Mendel's work, linking the behavior of his unit factors to that of chromosomes during meiosis. ■

Unit Factors, Genes, and Homologous Chromosomes

Because the correlation between Sutton's and Boveri's observations and Mendelian postulates serves as the foundation for the modern description of transmission genetics, we will examine this correlation in some depth before moving on to other topics.

As we know, each species possesses a specific number of chromosomes in each somatic cell nucleus. For diploid organisms, this number is called the **diploid number (2*n*)** and is characteristic of that species. During the formation of gametes (meiosis), the number is precisely halved (*n*), and when two gametes combine during fertilization, the diploid number is reestablished. During meiosis, however, the chromosome number is not reduced in a random manner. It was apparent to early cytologists that the diploid number of chromosomes is composed of homologous pairs identifiable by their morphological appearance and behavior. The gametes contain one member of each pair—thus the chromosome complement of a gamete is quite specific, and the number of chromosomes in each gamete is equal to the haploid number.

With this basic information, we can correlate the behavior of unit factors and chromosomes and genes. Unit factors are really genes located on homologous pairs of chromosomes. Members of each pair of homologs separate, or segregate, during gamete formation.

To illustrate the principle of independent assortment, it is important to distinguish between members of any given homologous pair of chromosomes. One member of each pair is derived from the **maternal parent,** whereas the other comes from the **paternal parent.** Following independent segregation of each pair of homologs, each gamete receives one member from each pair of chromosomes. All possible combinations are formed with equal probability. The independent behavior of Mendel's pairs of unit factors is due to their presence on separate pairs of homologous chromosomes.

Observations of the phenotypic diversity of living organisms make it logical to assume that there are many more genes than chromosomes. Therefore, each homolog must carry genetic information for more than one trait. The currently accepted concept is that a chromosome is composed of a large number of linearly ordered, information-containing genes. Mendel's paired unit factors (which determine tall or dwarf stems, for example) actually constitute a pair of genes located on one pair of homologous chromosomes. The location on a given chromosome where any particular gene occurs is called its **locus** (pl. loci). The different alleles of a given gene (for example, *G* and *g*) contain slightly different genetic information (green or yellow) that

determines the same character (seed color in this case). Although we have examined only genes with two alternative alleles, most genes have more than two allelic forms. We conclude this section by reviewing the criteria necessary to classify two chromosomes as a homologous pair:

1. During mitosis and meiosis, when chromosomes are visible in their characteristic shapes, both members of a homologous pair are the same size and exhibit identical centromere locations. The sex chromosomes (e.g., the X and the Y chromosomes in mammals) are an exception.
2. During early stages of meiosis, homologous chromosomes form pairs, or synapse.
3. Although it is not generally visible under the microscope, homologs contain the identical linear order of gene loci.

EVOLVING CONCEPT OF THE GENE

Based on the pioneering work of Gregor Mendel, the gene was viewed as a heritable unit factor that determines the expression of an observable trait, or phenotype. ■

3.6 Independent Assortment Leads to Extensive Genetic Variation

One consequence of independent assortment is the production by an individual of genetically dissimilar gametes. Genetic variation results because the two members of any homologous pair of chromosomes are rarely, if ever, genetically identical. As the maternal and paternal members of all pairs are distributed to gametes through independent assortment, all possible chromosome combinations are produced, leading to extensive genetic diversity.

We have seen that the number of possible gametes, each with different chromosome compositions, is 2^n, where n equals the haploid number. Thus, if a species has a haploid number of 4, then 2^4, or 16, different gamete combinations can be formed as a result of independent assortment. Although this number is not high, consider the human species, where $n = 23$. When 2^{23} is calculated, we find that in excess of 8×10^6, or over 8 million, different types of gametes are possible through independent assortment. Because fertilization represents an event involving only one of approximately 8×10^6 possible gametes from each of two parents, each offspring represents only one of $(8 \times 10^6)^2$ or one of only 64×10^{12} potential genetic combinations. Given that this probability is less than one in one trillion, it is no wonder that, except for identical twins, each member of the human species exhibits a distinctive set of traits—this number of combinations of chromosomes is far greater than the number of humans who have ever lived on Earth! Genetic variation resulting from independent assortment has been extremely important to the process of evolution in all sexually reproducing organisms.

3.7 Laws of Probability Help to Explain Genetic Events

Recall that genetic ratios—for example, 3/4 tall:1/4 dwarf—are most properly thought of as probabilities. These values predict the outcome of each fertilization event, such that the probability of each zygote having the genetic potential for becoming tall is 3/4, whereas the potential for its being a dwarf is 1/4. Probabilities range from 0.0, where an event *is certain not to occur,* to 1.0, where an event *is certain to occur.* In this section, we consider the relation of probability to genetics. When two or more events with known probabilities occur independently but at the same time, we can calculate the probability of their possible outcomes occurring together. This is accomplished by applying the **product law,** which states that *the probability of two or more events occurring simultaneously is equal to the product of their individual probabilities* (see Section 3.3). Two or more events are independent of one another if the outcome of each one does not affect the outcome of any of the others under consideration.

To illustrate the product law, consider the possible results if you toss a penny (P) and a nickel (N) at the same time and examine all combinations of heads (H) and tails (T) that can occur. There are four possible outcomes:

$$(P_H{:}N_H) = (1/2)(1/2) = 1/4$$
$$(P_T{:}N_H) = (1/2)(1/2) = 1/4$$
$$(P_H{:}N_T) = (1/2)(1/2) = 1/4$$
$$(P_T{:}N_T) = (1/2)(1/2) = 1/4$$

The probability of obtaining a head or a tail in the toss of either coin is 1/2 and is unrelated to the outcome for the other coin. Thus, all four possible combinations are predicted to occur with equal probability.

If we want to calculate the probability when the possible outcomes of two events are independent of one another but can be accomplished in more than one way, we can apply the **sum law.** For example, what is the probability of tossing our penny and nickel and obtaining one head and one tail? In such a case, we do not care whether it is the penny or the nickel that comes up heads, provided that the other coin has the alternative outcome. As we saw above, there are two ways in which the desired outcome can be accomplished, each with a probability of 1/4. The sum law states that *the probability of obtaining any single outcome, where that outcome can be achieved by two or more events, is equal to the sum of the individual probabilities of all such*

events. Thus, according to the sum law, the overall probability in our example is equal to

$$(1/4) + (1/4) = 1/2$$

One-half of all two-coin tosses are predicted to yield the desired outcome.

These simple probability laws will be useful throughout our discussions of transmission genetics and for solving genetics problems. In fact, we already applied the product law when we used the forked-line method to calculate the phenotypic results of Mendel's dihybrid and trihybrid crosses. When we wish to know the results of a cross, we need only calculate the probability of each possible outcome. The results of this calculation then allow us to predict the proportion of offspring expressing each phenotype or each genotype.

An important point to remember when you deal with probability is that predictions of possible outcomes are based on large sample sizes. If we predict that 9/16 of the offspring of a dihybrid cross will express both dominant traits, it is very unlikely that, in a small sample, exactly 9 of every 16 will express this phenotype. Instead, our prediction is that, of a large number of offspring, approximately 9/16 will do so. The deviation from the predicted ratio in smaller sample sizes is attributed to chance, a subject we examine in our discussion of statistics in the next section. As you shall see, the impact of deviation due strictly to chance diminishes as the sample size increases.

ESSENTIAL POINT

Since genetic ratios are expressed as probabilities, deriving outcomes of genetic crosses requires an understanding of the laws of probability. ■

3.8 Chi-Square Analysis Evaluates the Influence of Chance on Genetic Data

Mendel's 3:1 monohybrid and 9:3:3:1 dihybrid ratios are hypothetical predictions based on the following assumptions: (1) each allele is dominant or recessive, (2) segregation is unimpeded, (3) independent assortment occurs, and (4) fertilization is random. The final two assumptions are influenced by chance events and therefore are subject to random fluctuation. This concept of **chance deviation** is most easily illustrated by tossing a single coin numerous times and recording the number of heads and tails observed. In each toss, there is a probability of 1/2 that a head will occur and a probability of 1/2 that a tail will occur. Therefore, the expected ratio of many tosses is 1/2:1/2, or 1:1. If a coin is tossed 1000 times, usually *about* 500 heads and 500 tails will be observed. Any reasonable fluctuation from this hypothetical ratio (e.g., 486 heads and 514 tails) is attributed to chance.

As the total number of tosses is reduced, the impact of chance deviation increases. For example, if a coin is tossed only four times, you would not be too surprised if all four tosses resulted in only heads or only tails. For 1000 tosses, however, 1000 heads or 1000 tails would be most unexpected. In fact, you might believe that such a result would be impossible. Actually, all heads or all tails in 1000 tosses can be predicted to occur with a probability of $(1/2)^{1000}$. Since $(1/2)^{20}$ is less than one in a million times, an event occurring with a probability as small as $(1/2)^{1000}$ is virtually impossible. Two major points to keep in mind when predicting or analyzing genetic outcomes are:

1. The outcomes of independent assortment and fertilization, like coin tossing, are subject to random fluctuations from their predicted occurrences as a result of chance deviation.
2. As the sample size increases, the average deviation from the expected results decreases. Therefore, a larger sample size diminishes the impact of chance deviation on the final outcome.

Chi-Square Calculations and the Null Hypothesis

In genetics, being able to evaluate observed deviation is a crucial skill. When we assume that data will fit a given ratio such as 1:1, 3:1, or 9:3:3:1, we establish what is called the **null hypothesis (H_0).** It is so named because the hypothesis assumes that there is *no real difference* between the *measured values* (or ratio) and the *predicted values* (or ratio). Any apparent difference can be attributed purely to chance. The validity of the null hypothesis for a given set of data is measured using statistical analysis. Depending on the results of this analysis, the null hypothesis may either (1) *be rejected* or (2) *fail to be rejected*. If it is rejected, the observed deviation from the expected result is judged not to be attributable to chance alone. In this case, the null hypothesis and the underlying assumptions leading to it must be reexamined. If the null hypothesis fails to be rejected, any observed deviations are attributed to chance.

One of the simplest statistical tests for assessing the goodness of fit of the null hypothesis is **chi-square (χ^2) analysis.** This test takes into account the observed deviation in each component of a ratio (from what was expected) as well as the sample size and reduces them to a single numerical value. The value for χ^2 is then used to estimate how frequently the observed deviation can be expected to occur strictly as a result of chance. The formula used in chi-square analysis is

$$\chi^2 = \Sigma \frac{(o - e)^2}{e}$$

where o is the observed value for a given category, e is the expected value for that category, and Σ (the Greek letter sigma) represents the sum of the calculated values for each category in the ratio. Because $(o - e)$ is the deviation (d) in each case, the equation reduces to

$$\chi^2 = \Sigma \frac{d^2}{e}$$

Table 3.1(a) shows the steps in the χ^2 calculation for the F_2 results of a hypothetical monohybrid cross. To analyze the data obtained from this cross, work from left to right across the table, verifying the calculations as appropriate. Note that regardless of whether the deviation d is positive or negative, d^2 always becomes positive after the number is squared. In **Table 3.1(b)** F_2 results of a hypothetical dihybrid cross are analyzed. Make sure that you understand how each number was calculated in this example.

The final step in chi-square analysis is to interpret the χ^2 value. To do so, you must initially determine a value called the **degrees of freedom (*df*),** which is equal to $n - 1$, where n is the number of different categories into which the data are divided, in other words, the number of possible outcomes. For the 3:1 ratio, $n = 2$, so $df = 1$. For the 9:3:3:1 ratio, $n = 4$ and $df = 3$. Degrees of freedom must be taken into account because the greater the number of categories, the more deviation is expected as a result of chance.

Once you have determined the degrees of freedom, you can interpret the χ^2 value in terms of a corresponding **probability value (*p*)**. Since this calculation is complex, we usually take the p value from a standard table or graph. **Figure 3–10** shows a wide range of χ^2 values and the corresponding p values for various degrees of freedom in both a graph and a table. Let's use the graph to explain how to determine the p value. The caption for Figure 3–10(b) explains how to use the table.

To determine p using the graph, execute the following steps:

1. Locate the χ^2 value on the abscissa (the horizontal axis, or x-axis).
2. Draw a vertical line from this point up to the line on the graph representing the appropriate *df*.
3. From there, extend a horizontal line to the left until it intersects the ordinate (the vertical axis, or y-axis).
4. Estimate, by interpolation, the corresponding p value.

We used these steps for the monohybrid cross in Table 3.1(a) to estimate the p value of 0.48, as shown in Figure 3–10(a). Now try this method to see if you can determine the p value for the dihybrid cross [Table 3.1(b)]. Since the χ^2 value is 4.16 and $df = 3$, an approximate p value is 0.26. Checking this result in the table confirms that p values for both the monohybrid and dihybrid crosses are between 0.20 and 0.50.

Interpreting Probability Values

So far, we have been concerned with calculating χ^2 values and determining the corresponding p values. These steps bring us to the most important aspect of chi-square analysis: understanding the meaning of the p value. It is simplest to think of the p value as a percentage. Let's use the example of the dihybrid cross in Table 3.1(b) where $p = 0.26$, which can be thought of as 26 percent. In our example, the p value indicates that if we repeat the same experiment many times, 26 percent of the trials would be expected to exhibit chance deviation as great as or greater than that seen in the initial trial. Conversely, 74 percent of the repeats would show less deviation than initially observed as a result of chance. Thus, the p value reveals that a null

TABLE 3.1 Chi-Square Analysis

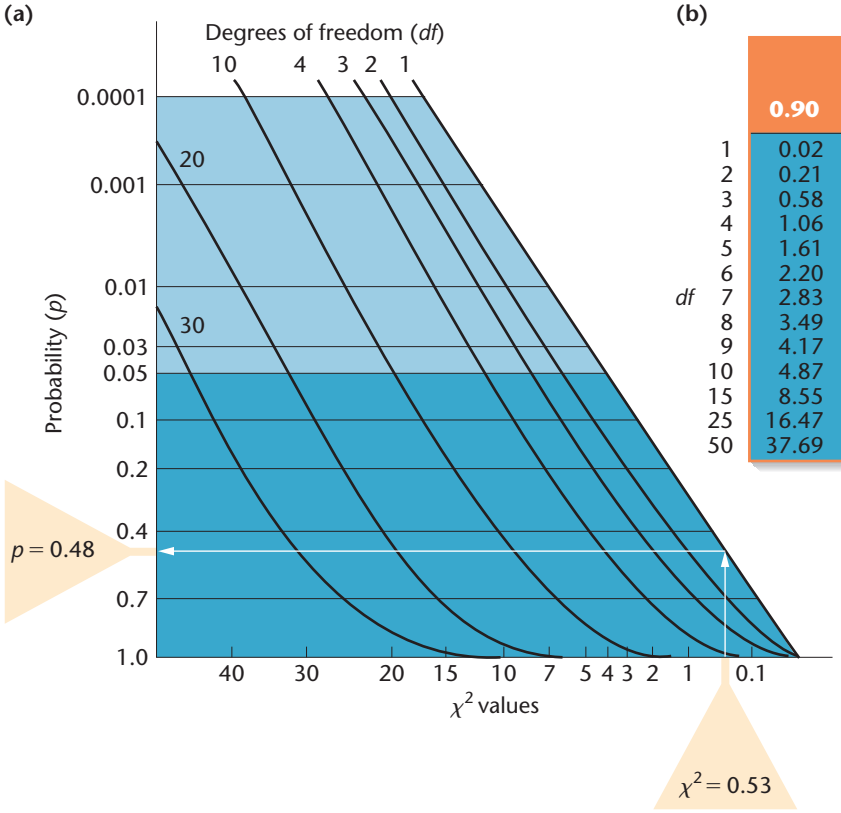
(a) Monohybrid

Cross Expected Ratio	Observed (*o*)	Expected (*e*)	Deviation (*o* – *e*)	Deviation (d^2)	d^2/e
3/4	740	3/4(1000) = 750	740–750 = −10	$(-10)^2 = 100$	100/750 = 0.13
1/4	260	1/4(1000) = 250	260–250 = +10	$(+10)^2 = 100$	100/250 = 0.40
	Total = 1000				$\chi^2 = 0.53$
					$p = 0.48$

(b) Dihybrid

Cross Expected Ratio	Observed (*o*)	Expected (*e*)	Deviation (*o* – *e*)	Deviation (d^2)	d^2/e
9/16	587	567	+20	400	0.71
3/16	197	189	+8	64	0.34
3/16	168	189	−21	441	2.33
1/16	56	63	−7	49	0.78
	Total = 1008				$\chi^2 = 4.16$
					$p = 0.26$

(a)



(b)

df	Probability (p) 0.90	0.50	0.20	0.05	0.01	0.001
1	0.02	0.46	1.64	3.84	6.64	10.83
2	0.21	1.39	3.22	5.99	9.21	13.82
3	0.58	2.37	4.64	7.82	11.35	16.27
4	1.06	3.36	5.99	9.49	13.28	18.47
5	1.61	4.35	7.29	11.07	15.09	20.52
6	2.20	5.35	8.56	12.59	16.81	22.46
7	2.83	6.35	9.80	14.07	18.48	24.32
8	3.49	7.34	11.03	15.51	20.09	26.13
9	4.17	8.34	12.24	16.92	21.67	27.88
10	4.87	9.34	13.44	18.31	23.21	29.59
15	8.55	14.34	19.31	25.00	30.58	37.30
25	16.47	24.34	30.68	37.65	44.31	52.62
50	37.69	49.34	58.16	67.51	76.15	86.60

χ^2 values

Fails to reject the null hypothesis
Rejects the null hypothesis

FIGURE 3–10 (a) Graph for converting χ^2 values to p values. (b) Table of χ^2 values for selected values of df and p. χ^2 values that lead to a p value of 0.05 or greater (darker blue areas) justify failure to reject the null hypothesis. Values leading to a p value of less than 0.05 (lighter blue areas) justify rejecting the null hypothesis. For example, the table in part (b) shows that for $\chi^2 = 0.53$ with 1 degree of freedom, the corresponding p value is between 0.20 and 0.50. The graph in (a) gives a more precise p value of 0.48 by interpolation. Thus, we fail to reject the null hypothesis.

hypothesis (concerning the 9:3:3:1 ratio, in this case) is never proved or disproved absolutely. Instead, a relative standard is set that we use to either *reject* or *fail to reject* the null hypothesis. This standard is most often a p value of 0.05. When applied to chi-square analysis, a p value less than 0.05 means that the observed deviation in the set of results will be obtained by chance alone less than 5 percent of the time. Such a p value indicates that the difference between the observed and predicted results is substantial and requires us to reject the null hypothesis.

On the other hand, p values of 0.05 or greater (0.05 to 1.0) indicate that the observed deviation will be obtained by chance alone 5 percent or more of the time. This conclusion allows us not to reject the null hypothesis (when we are using $p = 0.05$ as our standard). Thus, with its p value of 0.26, the null hypothesis that independent assortment accounts for the results fails to be rejected. Therefore, the observed deviation can be reasonably attributed to chance.

A final note is relevant here concerning the case where the null hypothesis is rejected, that is, where $p \leq 0.05$. Suppose we had tested a dataset to assess a possible 9:3:3:1 ratio, as in Table 3.1(b), but we rejected the null hypothesis based on our calculation. What are alternative interpretations of the data? Researchers will reassess the assumptions that underlie the null hypothesis. In our dyhibrid cross, we assumed that segregation operates faithfully for both gene pairs. We also assumed that fertilization is random and that the viability of all gametes is equal regardless of genotype—that is, all gametes are equally likely to participate in fertilization. Finally, we assumed that, following fertilization, all preadult stages and adult offspring are equally viable, regardless of their genotype. If any of these assumptions is incorrect, then the original hypothesis is not necessarily invalid.

An example will clarify this point. Suppose our null hypothesis is that a dihybrid cross between fruit flies will result in 3/16 mutant wingless flies. However, perhaps fewer of the mutant embryos are able to survive their preadult development or young adulthood compared to flies whose genotype gives rise to wings. As a result, when the data are gathered, there will be fewer than 3/16 wingless flies. Rejection of the null hypothesis is not in itself cause for us to reject the validity of the postulates of segregation and independent assortment, because other factors we are unaware of may also be affecting the outcome.

ESSENTIAL POINT

Chi-square analysis allows us to assess the null hypothesis, which states that there is no real difference between the expected and observed values. As such, it tests the probability of whether observed variations can be attributed to chance deviation. ■

NOW SOLVE THIS

3–4 In one of Mendel's dihybrid crosses, he observed 315 round, yellow, 108 round, green, 101 wrinkled, yellow, and 32 wrinkled, green F_2 plants. Analyze these data using the χ^2 test to see if

(a) they fit a 9:3:3:1 ratio.
(b) the round:wrinkled data fit a 3:1 ratio.
(c) the yellow:green data fit a 3:1 ratio.

■ **HINT:** *This problem asks you to apply χ^2 analysis to a set of data and to determine whether those data fit any of several ratios. The key to its solution is to first calculate χ^2 by initially determining the expected outcomes using the predicted ratios. Then follow a stepwise approach, determining the deviation in each case, and calculating d^2/e for each category. Once you have determined the χ^2 value, you must then determine and interpret the p value for each ratio.*

For more practice, see Problems 15, 16, and 17.

3.9 Pedigrees Reveal Patterns of Inheritance of Human Traits

We now explore how to determine the mode of inheritance of phenotypes in humans, where experimental matings are not made and where relatively few offspring are available for study. The traditional way to study inheritance has been to construct a family tree, indicating the presence or absence of the trait in question for each member of each generation. Such a family tree is called a **pedigree.** By analyzing a pedigree, we may be able to predict how the trait under study is inherited—for example, is it due to a dominant or recessive allele? When many pedigrees for the same trait are studied, we can often ascertain the mode of inheritance.

Pedigree Conventions

Figure 3–11 illustrates some of the conventions geneticists follow in constructing pedigrees. Circles represent females and squares designate males. If the sex of an individual is unknown, a diamond is used. Parents are generally connected to each other by a single horizontal line, and vertical lines lead to their offspring. If the parents are related—that is, **consanguineous**—such as first cousins, they are connected by a double line. Offspring are called **sibs** (short for **siblings**) and are connected by a horizontal **sibship line.** Sibs are placed in birth order from left to right and are labeled with Arabic numerals. Parents also receive an Arabic number designation. Each generation is indicated by a Roman numeral. When a pedigree traces only a single trait, the circles, squares, and diamonds are shaded if the phenotype being considered is expressed and unshaded if not. In some pedigrees, those individuals that fail to express a recessive trait but are known with certainty to be heterozygous carriers have a shaded dot within their unshaded circle or square. If an individual is deceased and the phenotype is unknown, a diagonal line is placed over the circle or square.

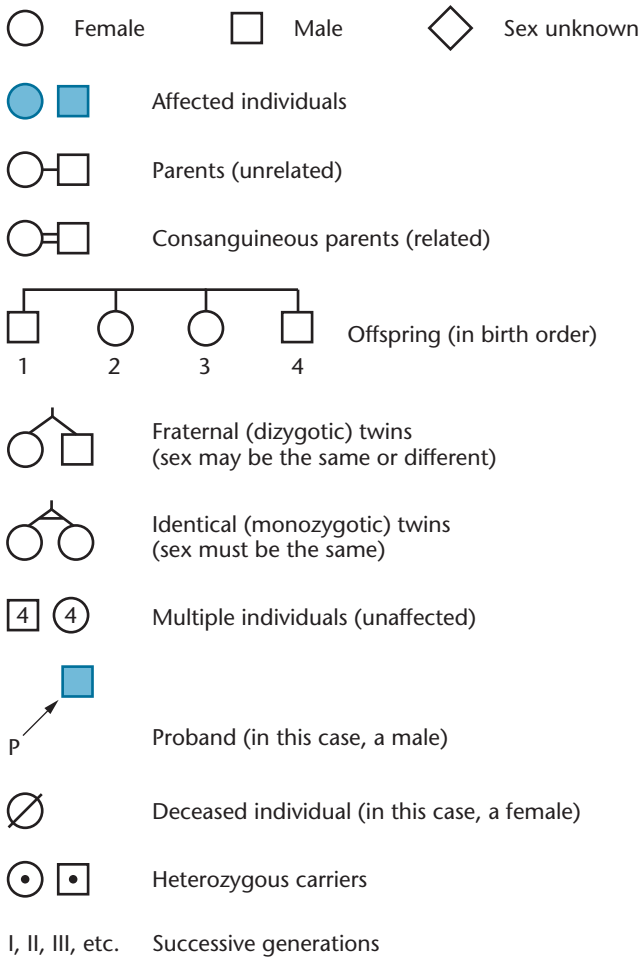


FIGURE 3–11 Conventions commonly encountered in human pedigrees.

Twins are indicated by diagonal lines stemming from a vertical line connected to the sibship line. For identical, or **monozygotic,** twins, the diagonal lines are linked by a horizontal line. Fraternal, or **dizygotic,** twins lack this connecting line. A number within one of the symbols represents that number of sibs of the same sex and of the same or unknown phenotypes. The individual whose phenotype first brought attention to the family is called the **proband** and is indicated by an arrow connected to the designation **p.** This term applies to either a male or a female.

Pedigree Analysis

In **Figure 3–12**, two pedigrees are shown. The first is a representative pedigree for a trait that demonstrates autosomal recessive inheritance, such as **albinism,** where synthesis of

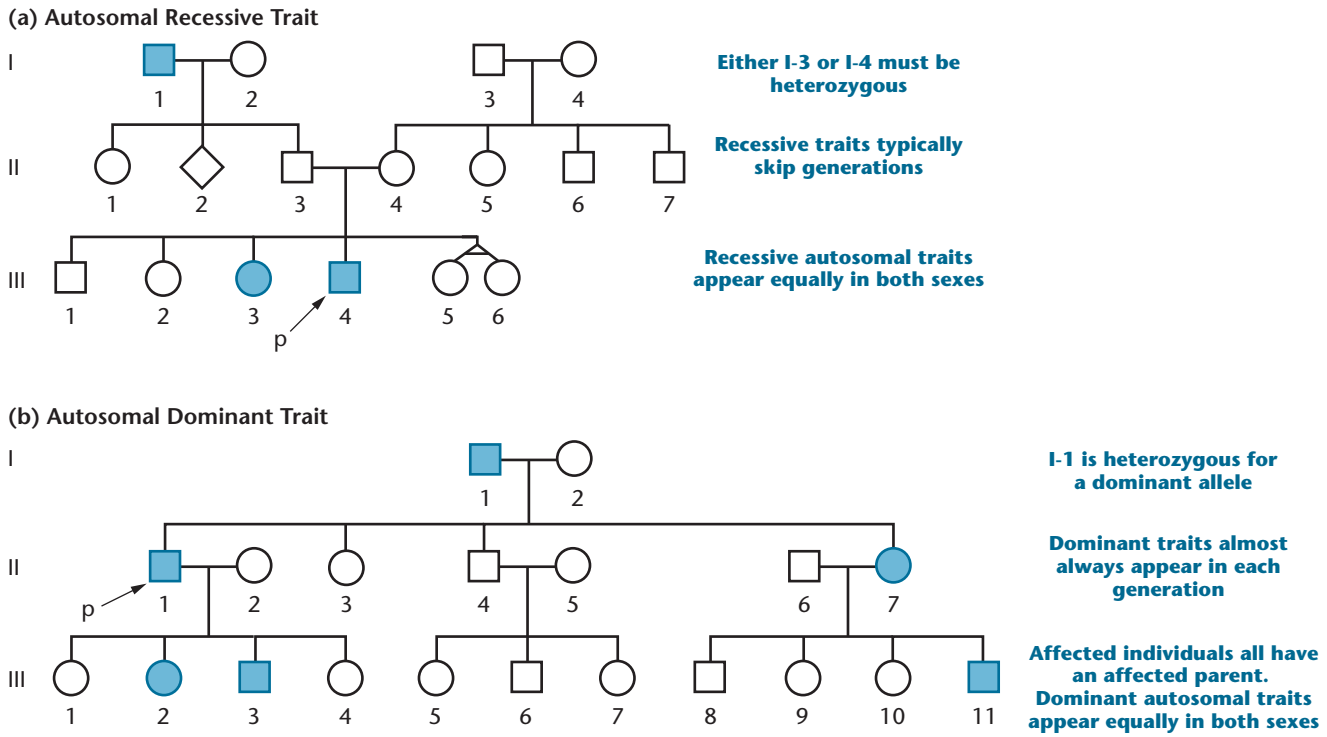


FIGURE 3–12 Representative pedigrees for two characteristics, each followed through three generations.

the pigment melanin is obstructed. The male parent of the first generation (I-1) is affected. Characteristic of a situation in which a parent has a rare recessive trait, the trait "disappears" in the offspring of the next generation. Assuming recessiveness, we might predict that the unaffected female parent (I-2) is a homozygous normal individual because none of the offspring show the disorder. Had she been heterozygous, one-half of the offspring would be expected to exhibit albinism, but none do. However, such a small sample (three offspring) prevents our knowing for certain.

Further evidence supports the prediction of a recessive trait. If albinism were inherited as a dominant trait, individual II-3 would have to express the disorder in order to pass it to his offspring (III-3 and III-4), but he does not. Inspection of the offspring constituting the third generation (row III) provides still further support for the hypothesis that albinism is a recessive trait. If it is, parents II-3 and II-4 are both heterozygous, and approximately one-fourth of their offspring should be affected. Two of the six offspring do show albinism. This deviation from the expected ratio is not unexpected in crosses with few offspring. Once we are confident that albinism is inherited as an autosomal recessive trait, we could portray the II-3 and II-4 individuals with a shaded dot within their larger square and circle. Finally, we can note that, characteristic of pedigrees for autosomal traits, both males and females are affected with equal probability. Later in the text (see Chapter 4), we will examine a pedigree representing a gene located on the sex-determining X chromosome. We will see certain patterns characteristic of the transmission of X-linked traits; for example, these traits are more prevalent in male offspring and are never passed from affected fathers to their sons.

The second pedigree illustrates the pattern of inheritance for a trait such as Huntington disease, which is caused by an autosomal dominant allele. The key to identifying a pedigree that reflects a dominant trait is that all affected offspring will have a parent who also expresses the trait. It is also possible, by chance, that none of the offspring will inherit the dominant allele. If so, the trait will cease to exist in future generations. Like recessive traits, provided that the gene is autosomal, both males and females are equally affected.

When a given autosomal dominant disease is rare within the population, and most are, then it is highly unlikely that affected individuals will inherit a copy of the mutant gene from both parents. Therefore, in most cases, affected individuals are heterozygous for the dominant allele. As a result, approximately one-half of the offspring inherit it. This is borne out in the second pedigree in Figure 3–12. Furthermore, if a mutation is dominant, and a single copy is sufficient to produce a mutant phenotype, homozygotes are likely to be even more severely affected, perhaps even failing to survive. An illustration of this is the dominant gene for **familial hypercholesterolemia.** Heterozygotes display a defect in their receptors for low-density lipoproteins, the so-called LDLs (known popularly as "bad cholesterol"). As a result, too little cholesterol is taken up by cells from the blood, and elevated plasma levels of LDLs result. Without intervention, such heterozygous individuals usually have heart attacks during the fourth decade of their life, or before. While heterozygotes have LDL levels

about double that of a normal individual, rare homozygotes have been detected. They lack LDL receptors altogether, and their LDL levels are nearly ten times above the normal range. They are very likely to have a heart attack very early in life, even before age 5, and almost inevitably before they reach the age of 20.

3.10 Tay–Sachs Disease: The Molecular Basis of a Recessive Disorder in Humans

We conclude this chapter by examining a case where the molecular basis of normal and mutant genes and their resultant phenotypes have now been revealed. This discussion expands your understanding of how genes control phenotypes.

Of particular interest are cases where a single mutant gene causes multiple effects associated with a severe disorder in humans. Let's consider the modern explanation of the gene that causes **Tay–Sachs disease (TSD),** a devastating recessive disorder involving unalterable destruction of the central nervous system. Infants with TSD are unaffected at birth and appear to develop normally until they are about 6 months old. Then, a progressive loss of mental and physical abilities occurs. Afflicted infants eventually become blind, deaf, mentally retarded, and paralyzed, often within only a year or two, seldom living beyond age 5. Typical of rare autosomal recessive disorders, two unaffected heterozygous parents, who most often have no family history of the disorder, have a probability of one in four of having a Tay–Sachs child.

We know that proteins are the end products of the expression of most all genes. The protein product involved in TSD has been identified, and we now have a clear understanding of the underlying molecular basis of the disorder. TSD results from the loss of activity of a single enzyme, **hexosaminidase A (Hex-A).** Hex-A, normally found in lysosomes within cells, is needed to break down the ganglioside GM2, a lipid component of nerve cell membranes. Without functional Hex-A, gangliosides accumulate within neurons in the brain and cause deterioration of the nervous system. Heterozygous carriers of TSD with one normal copy of the gene produce only about 50 percent of the normal amount of Hex-A, but they show no symptoms of the disorder. The observation that the activity of only one gene (one wild-type allele) is sufficient for the normal development and function of the nervous system explains and illustrates the molecular basis of recessive mutations. Only when both genes are disrupted by mutation is the mutant phenotype evident. The responsible gene is located on chromosome 15 and codes for the alpha subunit of the Hex-A enzyme. More than 50 different mutations within the gene have been identified that lead to TSD phenotypes.

Online Mendelian Inheritance in Man

MasteringGenetics™ Visit the Study Area: Exploring Genomics

The Online Mendelian Inheritance in Man (OMIM) database is a catalog of human genes and human disorders that are inherited in a Mendelian manner. Genetic disorders that arise from major chromosomal aberrations, such as monosomy or trisomy (the loss of a chromosome or the presence of a superfluous chromosome, respectively), are not included. The OMIM database, updated daily, is a version of the book *Mendelian Inheritance in Man,* conceived and edited by Dr. Victor McKusick of Johns Hopkins University, until he passed in 2008.

The OMIM entries provide links to a wealth of information, including DNA and protein sequences, chromosomal maps, disease descriptions, and relevant scientific publications. In this exercise, you will explore OMIM to answer questions about the recessive human disease sickle-cell anemia and other Mendelian inherited disorders.

Exercise I – Sickle-cell Anemia

In this chapter, you were introduced to recessive and dominant human traits. You will now discover more about sickle-cell anemia as an autosomal recessive disease by exploring the OMIM database.

1. To begin the search, access the OMIM site at: www.omim.org.
2. In the "SEARCH" box, type "sickle-cell anemia" and click on the "Search" button to perform the search.
3. Click on the link for the entry #603903.
4. Review the text that appears to learn about sickle-cell anemia. Examine the list of subject headings in the right-hand column and explore these links for more information about sickle-cell anemia.
5. Select one or two references at the bottom of the page and follow them to their abstracts in PubMed.

6. Using the information in this entry, answer the following questions:
   a. Which gene is mutated in individuals with sickle-cell anemia?
   b. What are the major symptoms of this disorder?
   c. What was the first published scientific description of sickle-cell anemia?
   d. Describe two other features of this disorder that you learned from the OMIM database and state where in the database you found this information.

■ **Exercise II – Other Recessive or Dominant Disorders**

Select another human disorder that is inherited as either a dominant or recessive trait and investigate its features, following the general procedure presented above. Follow links from OMIM to other databases if you choose.

Describe several interesting pieces of information you acquired during your exploration and cite the information sources you encountered during the search.

CASE STUDY | To test or not to test

Thomas first discovered a potentially devastating piece of family history when he learned the medical diagnosis for his brother's increasing dementia, muscular rigidity, and frequency of seizures. His brother, at age 49, was diagnosed with Huntington disease (HD), a dominantly inherited condition that typically begins with such symptoms around the age of 45 and leads to death in one's early 60s. As depressing as the news was to Thomas, it helped explain his father's suicide. Thomas, 38, now wonders what his chances are of carrying the gene for HD, leading him and his wife to discuss the pros and cons of him undergoing genetic testing. Thomas and his wife have two teenage children, a boy and a girl.

1. What role might a genetic counselor play in this real-life scenario?
2. How might the preparation and analysis of a pedigree help explain the dilemma facing Thomas and his family?
3. If Thomas decides to go ahead with the genetic test, what should be the role of the health insurance industry in such cases?
4. If Thomas tests positive for HD, and you were one of his children, would you want to be tested?

INSIGHTS AND SOLUTIONS

Genetics problems are in many ways similar to word problems in algebra. The approach to solving them is identical: (1) analyze the problem carefully; (2) translate words into symbols and define each symbol precisely; and (3) choose and apply a specific technique to solve the problem. The first two steps are the most critical. The third step is largely mechanical. Keep this in mind as you analyze the following problems.

1. Mendel found that full pea pods are dominant over constricted pods, while round seeds are dominant over wrinkled seeds. One of his crosses was between full, round plants and constricted, wrinkled plants. From this cross, he obtained an F_1 generation that was all full and round. In the F_2 generation, Mendel obtained his classic 9:3:3:1 ratio. Using this information, determine the expected F_1 and F_2 results of a cross between homozygous constricted, round and full, wrinkled plants.

Solution: First, assign gene symbols to each pair of contrasting traits. Use the lowercase first letter of each recessive trait to designate that trait, and use the same letter in uppercase to designate the dominant trait. Thus, *C* and *c* indicate full and constricted pods, respectively, and *W* and *w* indicate the round and wrinkled phenotypes, respectively.

Determine the genotypes of the P_1 generation, form the gametes, combine them in the F_1 generation, and read off the phenotype(s):

P_1:	*ccWW* constricted, round	×	*CCww* full, wrinkled
	↓		↓
Gametes:	*cW*		Cw
F_1:		*CcWw* full, round	

You can immediately see that the F_1 generation expresses both dominant phenotypes and is heterozygous for both gene pairs. Thus, you expect that the F_2 generation will yield the classic Mendelian ratio of 9:3:3:1. Let's work it out anyway, just to confirm this expectation, using the forked-line method. Both gene pairs are heterozygous and can be expected to assort independently, so we can predict the F_2 outcomes from each gene pair separately and then proceed with the forked-line method.

The F_2 offspring should exhibit the individual traits in the following proportions:

Cc × *Cc*		*Ww* × *Ww*	
↓		↓	
CC *Cc* *cC*	full	*WW* *Ww* *wW*	round
cc	constricted	*ww*	wrinkled

(continued)

Insights and Solutions—continued

Using these proportions to complete a forked-line diagram confirms the 9:3:3:1 phenotypic ratio. (Remember that this ratio represents proportions of 9/16:3/16:3/16:1/16.) Note that we are applying the product law as we compute the final probabilities:

3/4 full
- 3/4 round — (3/4)(3/4) → 9/16 full, round
- 1/4 wrinkled — (3/4)(1/4) → 3/16 full, wrinkled

1/4 constricted
- 3/4 round — (1/4)(3/4) → 3/16 constricted, round
- 1/4 wrinkled — (1/4)(1/4) → 1/16 constricted, wrinkled

2. In another cross, involving parent plants of unknown genotype and phenotype, the following offspring were obtained.

   3/8 full, round
   3/8 full, wrinkled
   1/8 constricted, round
   1/8 constricted, wrinkled

   Determine the genotypes and phenotypes of the parents.

   Solution: This problem is more difficult and requires keener insight because you must work backward to arrive at the answer. The best approach is to consider the outcomes of pod shape separately from those of seed texture.

   Of all the plants, 3/8 + 3/8 = 3/4 are full and 1/8 + 1/8 = 1/4 are constricted. Of the various genotypic combinations that can serve as parents, which will give rise to a ratio of 3/4:1/4? This ratio is identical to Mendel's monohybrid F_2 results, and we can propose that both unknown parents share the same genetic characteristic as the monohybrid F_1 parents: they must both be heterozygous for the genes controlling pod shape and thus are *Cc*.

   Before we accept this hypothesis, let's consider the possible genotypic combinations that control seed texture. If we consider this characteristic alone, we can see that the traits are expressed in a ratio of 3/8 + 1/8 = 1/2 round: 3/8 + 1/8 = 1/2 wrinkled. To generate such a ratio, the parents cannot both be heterozygous or their offspring would yield a 3/4:1/4 phenotypic ratio. They cannot both be homozygous or all offspring would express a single phenotype. Thus, we are left with testing the hypothesis that one parent is homozygous and one is heterozygous for the alleles controlling texture. The potential case of *WW* × *Ww* does not work because it would also yield only a single phenotype. This leaves us with the potential case of *ww* × *Ww*. Offspring in such a mating will yield 1/2 *Ww* (round): 1/2 *ww* (wrinkled), exactly the outcome we are seeking.

   Now, let's combine our hypotheses and predict the outcome of the cross. In our solution, we use a dash (–) to indicate that the second allele may be dominant or recessive, since we are only predicting phenotypes.

3/4 *C*–
- 1/2 *Ww* → 3/8 *C*–*Ww* full, round
- 1/2 *ww* → 3/8 *C*–*ww* full, wrinkled

1/4 *cc*
- 1/2 *Ww* → 1/8 *ccWw* constricted, round
- 1/2 *ww* → 1/8 *ccww* constricted, wrinkled

As you can see, this cross produces offspring in proportions that match our initial information, and we have solved the problem. Note that, in the solution, we have used genotypes in the forked-line method, in contrast to the use of phenotypes in Solution 1.

3. In the laboratory, a genetics student crossed flies with normal long wings with flies expressing the *dumpy* mutation (truncated wings), which she believed was a recessive trait. In the F_1 generation, all flies had long wings. The following results were obtained in the F_2 generation:
   792 long-winged flies
   208 dumpy-winged flies
   The student tested the hypothesis that the dumpy wing is inherited as a recessive trait using χ^2 analysis of the F_2 data.

   (a) What ratio was hypothesized?

   (b) Did the analysis support the hypothesis?

   (c) What do the data suggest about the *dumpy* mutation?

   Solution:

   (a) The student hypothesized that the F_2 data (792:208) fit Mendel's 3:1 monohybrid ratio for recessive genes.

   (b) The initial step in χ^2 analysis is to calculate the expected results (*e*) for a ratio of 3:1. Then we can compute deviation *o* – *e* (*d*) and the remaining numbers.

Ratio	*o*	*e*	*d*	d^2	d^2/e
3/4	792	750	42	1764	2.35
1/4	208	250	−42	1764	7.06
	Total = 1000				

$$\chi^2 = \Sigma\frac{d^2}{e}$$
$$= 2.35 + 7.06$$
$$= 9.41$$

We consult Figure 3–10 to determine the probability (*p*) and to decide whether the deviations can be attributed to chance. There are two possible outcomes ($n = 2$), so the degrees of freedom (df) = $n - 1$, or 1. The table in Figure 3–10(b) shows that *p* is a value between 0.01 and 0.001; the graph in Figure 3–10(a) gives an estimate of about 0.001. Since $p < 0.05$, we reject the null hypothesis. The data do not fit a 3:1 ratio.

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

When working out genetics problems in this and succeeding chapters, always assume that members of the P_1 generation are homozygous, unless the information or data you are given require you to do otherwise.

HOW DO WE KNOW?

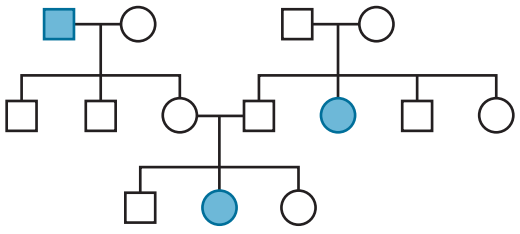
1. In this chapter, we focused on the Mendelian postulates, probability, and pedigree analysis. We also considered some of the methods and reasoning by which these ideas, concepts, and techniques were developed. On the basis of these discussions, what answers would you propose to the following questions:
   (a) How was Mendel able to derive postulates concerning the behavior of "unit factors" during gamete formation, when he could not directly observe them?
   (b) How do we know whether an organism expressing a dominant trait is homozygous or heterozygous?
   (c) In analyzing genetic data, how do we know whether deviation from the expected ratio is due to chance rather than to another, independent factor?
   (d) Since experimental crosses are not performed in humans, how do we know how traits are inherited?

CONCEPT QUESTION

2. Review the Chapter Concepts list on p. 47. The first five concepts provide a modern interpretation of Mendelian postulates. Based on these concepts, write a short essay that correlates Mendel's four postulates with what is now known about genes, alleles, and homologous chromosomes. ■
3. In a cross between a black and a white guinea pig, all members of the F_1 generation are black. The F_2 generation is made up of approximately 3/4 black and 1/4 white guinea pigs. Diagram this cross, and show the genotypes and phenotypes.
4. Early-onset myopia in humans is inherited as a simple dominant trait. Determine the genotypes of the parents and offspring for the following families. Mention alternate genotypes wherever applicable.
   (a) One normal (without early-onset myopia) parent and one abnormal (with early-onset myopia) parent produce six children, out of which only one is normal.
   (b) An abnormal male and a normal female produce five normal children.
5. In a problem involving albinism (see Problem 4), which of Mendel's postulates are demonstrated?
6. Why was the garden pea a good choice as an experimental organism in Mendel's work?
7. Mendel crossed peas having round seeds and yellow cotyledons with peas having wrinkled seeds and green cotyledons. All the F_1 plants had round seeds with yellow cotyledons. Diagram this cross through the F_2 generation, using both the Punnett square and forked-line methods.
8. Refer to the Now Solve This Problem 3-2 on p. 55. Are any of the crosses in this problem testcrosses? If so, which one(s)?
9. Which of Mendel's postulates can be demonstrated in the Now Solve This Problem 3-2 on p. 55 but not in Problem 3 above? Define this postulate.
10. Correlate Mendel's four postulates with what is now known about homologous chromosomes, genes, alleles, and the process of meiosis.
11. Distinguish between Mendel's postulates of segregation and independent assortment.
12. Two organisms, *AABBCCDDEE* and *aabbccddee*, are mated to produce an F_1 that is self-fertilized. If the capital letters represent dominant, independently assorting alleles:
    (a) How many different genotypes will occur in the F_2?
    (b) What proportion of the F_2 genotypes will be recessive for all five loci?
    (c) Would you change your answers to (a) and/or (b) if the initial cross occurred between *AAbbCCddee* × *aaBBccDDEE* parents?
    (d) Would you change your answers to (a) and/or (b) if the initial cross occurred between *AABBCCDDEE* × *aabbccddEE* parents?
13. Albinism, lack of pigmentation in humans, results from an autosomal recessive gene (a). Two parents with normal pigmentation have an albino child.
    (a) What is the probability that their next child will be albino?
    (b) What is the probability that their next child will be an albino girl?
    (c) What is the probability that their next three children will be albino?
14. Mendel crossed peas with round, green seeds with peas having wrinkled, yellow seeds. All F_1 plants had seeds that were round and yellow. Predict the results of testcrossing these F_1 plants.
15. Shown are F_2 results of two of Mendel's monohybrid crosses. State a null hypothesis that you will test using chi-square analysis. Calculate the χ^2 value and determine the p value for both crosses, then interpret the p values. Which cross shows a greater amount of deviation?

(a) Full pods	882
Constricted pods	299
(b) Violet flowers	705
White flowers	224

16. A plant breeder observed that for a certain leaf trait of maize that shows two phenotypes (phenotype 1 and phenotype 2), the F_1 generation exhibits 200 plants with phenotype 1 and 160 with phenotype 2. Using two different null hypotheses and chi-square analysis, compute if the data fits (a) a 3:1 ratio, and (b) a 1:1 ratio.
17. Define critical p value. Explain what significance this value has for predicting the reproducibility of an experiment involving crosses. Explain why the null hypothesis is generally rejected for p values lower than 0.05.
18. Consider three independently assorting gene pairs, *A*/*a*, *B*/*b*, and *C*/*c*, where each demonstrates typical dominance (*A*–, *B*–, *C*–) and recessiveness (*aa, bb, cc*). What is the probability of obtaining an offspring that is *AABbCc* from parents that are *AaBbCC* and *AABbCc*?
19. What is the probability of obtaining a triply recessive individual from the parents shown in Problem 18?
20. Of all offspring of the parents in Problem 18, what proportion will express all three dominant traits?
21. For the following pedigree, predict the mode of inheritance and the resulting genotypes of each individual. Assume that the alleles *A* and *a* control the expression of the trait.

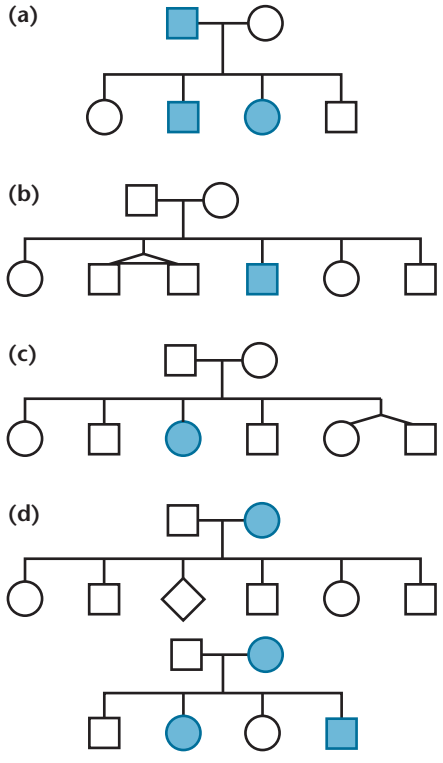


22. Which of Mendel's postulates are demonstrated by the pedigree in Problem 21? List and define these postulates.
23. Among dogs, short hair is dominant to long hair and dark coat color is dominant to white (albino) coat color. Assume that these two coat traits are caused by independently segregating gene pairs. For each of the crosses given below, write the most probable genotype (or genotypes if more than one answer is possible) for the parents. It is important that you select a realistic symbol set and define each symbol below.

Parental Phenotypes		*Phenotypes of Offspring*			
Short Dark	*Long Dark*	*Short Albino*	*Long Albino*		
(a) dark, short × dark, long		26	24	0	0
(b) albino, short × albino, short		0	0	102	33
(c) dark, short × albino, short		16	0	16	0
(d) dark, short × dark, short		175	67	61	21

Assume that for cross (d), you were interested in determining whether fur color follows a 3:1 ratio. Set up (but do not complete the calculations) a Chi-square test for these data [fur color in cross (d)].

24. Draw all possible conclusions concerning the mode of inheritance of the trait expressed in each of the following limited pedigrees. (Each case is based on a different trait.)



25. Two true-breeding pea plants are crossed. One parent is round, terminal, violet, constricted, while the other expresses the contrasting phenotypes of wrinkled, axial, white, full. The four pairs of contrasting traits are controlled by four genes, each located on a separate chromosome. In the F_1 generation, only round, axial, violet, and full are expressed. In the F_2 generation, all possible combinations of these traits are expressed in ratios consistent with Mendelian inheritance.

(a) What conclusion can you draw about the inheritance of these traits based on the F_1 results?
(b) Which phenotype appears most frequently in the F_2 results? Write a mathematical expression that predicts the frequency of occurrence of this phenotype.
(c) Which F_2 phenotype is expected to occur least frequently? Write a mathematical expression that predicts this frequency.
(d) How often is either P_1 phenotype likely to occur in the F_2 generation?
(e) If the F_1 plant is testcrossed, how many different phenotypes will be produced?

26. The wild-type (normal) fruit fly, *Drosophila melanogaster,* has straight wings and long bristles. Mutant strains have been isolated with either curled wings or short bristles. The genes representing these two mutant traits are located on separate chromosomes. Carefully examine the data from the five crosses below. (a) For each mutation, determine whether it is dominant or recessive. In each case, identify which crosses support your answer; and (b) define gene symbols and determine the genotypes of the parents for each cross.

	Number of Progeny			
Cross	straight wings, long bristles	straight wings, short bristles	curled wings, long bristles	curled wings, short bristles
1. straight, short × straight, short	30	90	10	30
2. straight, long × straight, long	120	0	40	0
3. curled, long × straight, short	40	40	40	40
4. straight, short × straight, short	40	120	0	0
5. curled, short × straight, short	20	60	20	60

27. In an intra-species cross performed in mustard plants of two different species (*Brassica juncea* and *Brassica oleracea*), a tall plant (*TT*) was crossed with a dwarf (*tt*) variety in each of the two species. The members of the F_1 generation were crossed to produce the F_2 generation. Of the F_2 plants, *Brassica juncea* had 60 tall and 20 dwarf plants, while *Brassica oleracea* had 100 tall and 20 dwarf plants. Use chi-square analysis to analyze these results.

4 Modification of Mendelian Ratios

CHAPTER CONCEPTS

- While alleles are transmitted from parent to offspring according to Mendelian principles, they sometimes fail to display the clear-cut dominant/recessive relationship observed by Mendel.
- In many cases, in contrast to Mendelian genetics, two or more genes are known to influence the phenotype of a single characteristic.
- Still another exception to Mendelian inheritance is the presence of genes on sex chromosomes, whereby one of the sexes contains only a single member of that chromosome.
- Phenotypes are often the combined result of both genetics and the environment within which genes are expressed.
- The result of the various exceptions to Mendelian principles is the occurrence of phenotypic ratios that differ from those resulting from standard monohybrid, dihybrid, and trihybrid crosses.
- Extranuclear inheritance, resulting from the expression of genes present in the DNA found in mitochondria and chloroplasts, modifies Mendelian inheritance patterns. Such genes are most often transmitted through the female gamete.

Labrador retriever puppies, which may display brown (also called chocolate), golden (also called yellow), or black coats.

In Chapter 3, we discussed the fundamental principles of transmission genetics. We saw that genes are present on homologous chromosomes and that these chromosomes segregate from each other and assort independently with other segregating chromosomes during gamete formation. These two postulates are the basic principles of gene transmission from parent to offspring. However, when gene expression does not adhere to a simple dominant/recessive mode or when more than one pair of genes influences the expression of a single character, the classic 3:1 and 9:3:3:1 ratios are usually modified. In this and the next several chapters, we consider more complex modes of inheritance. In spite of the greater complexity of these situations, the fundamental principles set down by Mendel still hold.

In this chapter, we restrict our initial discussion to the inheritance of traits controlled by only one set of genes. In diploid organisms, which have homologous pairs of chromosomes, two copies of each gene influence such traits. The copies need not be identical because alternative forms of genes (alleles) occur within populations. How alleles influence phenotypes is our primary focus. We will then consider gene interaction, a situation in which a single phenotype is affected by more than one set of genes. Numerous examples will be presented to illustrate a variety of heritable patterns observed in such situations.

Thus far, we have restricted our discussion to chromosomes other than the X and Y pair. By examining cases where genes are present on the X chromosome, illustrating X-linkage, we will see yet another modification

of Mendelian ratios. Our discussion of modified ratios also includes the consideration of sex-limited and sex-influenced inheritance, cases where the sex of the individual, but not necessarily genes on the X chromosome, influences the phenotype. We will also consider how a given phenotype often varies depending on the overall environment in which a gene, a cell, or an organism finds itself. This discussion points out that phenotypic expression depends on more than just the genotype of an organism. Finally, we conclude with a discussion of extranuclear inheritance, cases where DNA within organelles influences an organism's phenotype.

4.1 Alleles Alter Phenotypes in Different Ways

After Mendel's work was rediscovered in the early 1900s, researchers focused on the many ways in which genes influence an individual's phenotype. Each type of inheritance was more thoroughly investigated when observations of genetic data did not conform precisely to the expected Mendelian ratios, and hypotheses that modified and extended the Mendelian principles were proposed and tested with specifically designed crosses. The explanations were in accord with the principle that a phenotype is under the control of one or more genes located at specific loci on one or more pairs of homologous chromosomes.

To understand the various modes of inheritance, we must first examine the potential function of alleles. Alleles are alternative forms of the same gene. The allele that occurs most frequently in a population, the one that we arbitrarily designate as normal, is called the **wild-type allele.** This is often, but not always, dominant. Wild-type alleles are responsible for the corresponding wild-type phenotype and are the standards against which all other mutations occurring at a particular locus are compared.

A mutant allele contains modified genetic information and often specifies an altered gene product. For example, in human populations, there are many known alleles of the gene that encodes the β chain of human hemoglobin. All such alleles store information necessary for the synthesis of the β-chain polypeptide, but each allele specifies a slightly different form of the same molecule. Once the allele's product has been manufactured, the function of the product may or may not be altered.

The process of mutation is the source of alleles. For a new allele to be recognized when observing an organism, it must cause a change in the phenotype. A new phenotype results from a change in functional activity of the cellular product specified by that gene. Often, the mutation causes the diminution or the loss of the specific wild-type function. For example, if a gene is responsible for the synthesis of a specific enzyme, a mutation in that gene may ultimately change the conformation of this enzyme and reduce or eliminate its affinity for the substrate. Such a case is designated as a **loss-of-function mutation.** If the loss is complete, the mutation has resulted in what is called a **null allele.**

Conversely, other mutations may enhance the function of the wild-type product. Most often when this occurs, it is the result of increasing the quantity of the gene product. In such cases, the mutation may be affecting the regulation of transcription of the gene under consideration. Such cases are designated **gain-of-function mutations,** which generally result in dominant alleles since one copy in a diploid organism is sufficient to alter the normal phenotype. Examples of gain-of-function mutations include the genetic conversion of proto-oncogenes, which regulate the cell cycle, to oncogenes, where regulation is overridden by excess gene product. The result is the creation of a cancerous cell.

Having introduced the concept of gain- or loss-of-function mutations, it is important to note the possibility that a mutation will create an allele where no change in function can be detected. In this case, the mutation would not be immediately apparent since no phenotypic variation would be evident. However, such a mutation could be detected if the DNA sequence of the gene was examined directly. These are sometimes referred to as **neutral mutations** because the gene product presents no change to either the phenotype or the evolutionary fitness of the organism.

Finally, we note here that while a phenotypic trait may be affected by a single mutation in one gene, traits are often influenced by more than one gene. For example, enzymatic reactions are most often part of complex metabolic pathways leading to the synthesis of an end product, such as an amino acid. Mutations in any of the various reactions have a common effect—the failure to synthesize the end product. Therefore, phenotypic traits related to the end product are often influenced by more than one gene.

In each of the many crosses discussed in the next few chapters, only one or a few gene pairs are involved. Keep in mind that in each cross discussed, all genes that are not under consideration are assumed to have no effect on the inheritance patterns described.

4.2 Geneticists Use a Variety of Symbols for Alleles

In Chapter 3, we learned a standard convention that is used to symbolize alleles for very simple Mendelian traits. The initial letter of the name of a recessive trait, lowercased

and italicized, denotes the recessive allele, and the same letter in uppercase refers to the dominant allele. Thus, in the case of *tall* and *dwarf,* where *dwarf* is recessive, *D* and *d* represent the alleles responsible for these respective traits. Mendel used upper- and lowercase letters such as these to symbolize his unit factors.

Another useful system was developed in genetic studies of the fruit fly *Drosophila melanogaster* to discriminate between wild-type and mutant traits. This system uses the initial letter, or a combination of two or three letters, of the name of the mutant trait. If the trait is recessive, lowercase is used; if it is dominant, uppercase is used. The contrasting wild-type trait is denoted by the same letter, but with a superscript +. For example, *ebony* is a recessive body color mutation in *Drosophila*. The normal wild-type body color is gray. Using this system, we denote *ebony* by the symbol *e*, and we denote gray by e^+. The responsible locus may be occupied by either the wild-type allele (e^+) or the mutant allele (*e*). A diploid fly may thus exhibit one of three possible genotypes:

e^+/e^+	gray homozygote (wild type)
e^+/e	gray heterozygote (wild type)
e/e	ebony homozygote (mutant)

The slash between the letters indicates that the two allele designations represent the same locus on two homologous chromosomes. If we instead consider a dominant wing mutation such as *Wrinkled* (*Wr*) wing in *Drosophila*, the three possible designations are Wr^+/Wr^+, Wr^+/Wr, and *Wr*/*Wr*. The latter two genotypes express the wrinkled-wing phenotype.

One advantage of this system is that further abbreviation can be used when convenient: the wild-type allele may simply be denoted by the + symbol. With *ebony* as an example, the designations of the three possible genotypes become

+/+	gray homozygote (wild type)
+/*e*	gray heterozygote (wild type)
e/*e*	ebony homozygote (mutant)

Another variation is utilized when no dominance exists between alleles. We simply use uppercase italic letters and superscripts to denote alternative alleles (e.g., R^1 and R^2, L^M and L^N, I^A and I^B). Their use will become apparent later in this chapter.

Many diverse systems of genetic nomenclature are used to identify genes in various organisms. Usually, the symbol selected reflects the function of the gene or even a disorder caused by a mutant gene. For example, the yeast *cdk* is the abbreviation for the *c*yclin *d*ependent *k*inase gene, whose product is involved in cell-cycle regulation. In bacteria, leu^- refers to a mutation that interrupts the biosynthesis of the amino acid leucine, where the wild-type gene is designated leu^+. The symbol *dnaA* represents a bacterial gene involved in DNA replication (and DnaA is the protein made by that gene). In humans, capital letters are used to name genes: *BRCA*1 represents the first gene associated with susceptibility to *br*east *ca*ncer. Although these different systems may seem complex, they are useful ways to symbolize genes.

4.3 Neither Allele Is Dominant in Incomplete, or Partial, Dominance

A cross between parents with contrasting traits may generate offspring with an intermediate phenotype. For example, if plants such as four-o'clocks or snapdragons with red flowers are crossed with white-flowered plants, the offspring have pink flowers. Some red pigment is produced in the F_1 intermediate pink-colored flowers. Therefore, neither red nor white flower color is dominant. This situation is known as **incomplete,** or **partial, dominance.**

If this phenotype is under the control of a single gene and two alleles where neither is dominant, the results of the F_1 (pink) × F_1 (pink) cross can be predicted. The resulting F_2 generation shown in **Figure 4–1** confirms the hypothesis that only one pair of alleles determines these phenotypes. The *genotypic ratio* (1:2:1) of the F_2 generation is identical to that of Mendel's monohybrid cross. However, because neither allele is dominant, the *phenotypic ratio* is identical to the *genotypic ratio.* Note that because neither allele is recessive, we have chosen not to use upper and lowercase letters as symbols. Instead, we denoted the red and white alleles as R^1 and R^2, respectively. We could have used W^1 and W^2 or still other designations such as C^W and C^R, where *C* indicates "color" and the *W* and *R* superscripts indicate white and red.

Clear-cut cases of incomplete dominance, which result in intermediate expression of the overt phenotype, are relatively rare. However, even when complete dominance seems apparent, careful examination of the gene product, rather than the phenotype, often reveals an intermediate level of gene expression. An example is the human biochemical disorder **Tay–Sachs disease,** in which homozygous recessive individuals are severely affected with a fatal lipid storage disorder (see Chapter 3, page 64). There is almost no activity of the enzyme **hexosaminidase** in afflicted individuals. Heterozygotes, with only a single copy of the mutant gene, are phenotypically normal but express only about 50 percent of the enzyme activity found in homozygous normal individuals. Fortunately, this level of enzyme activity is adequate to achieve normal biochemical function—a situation not uncommon in enzyme disorders.

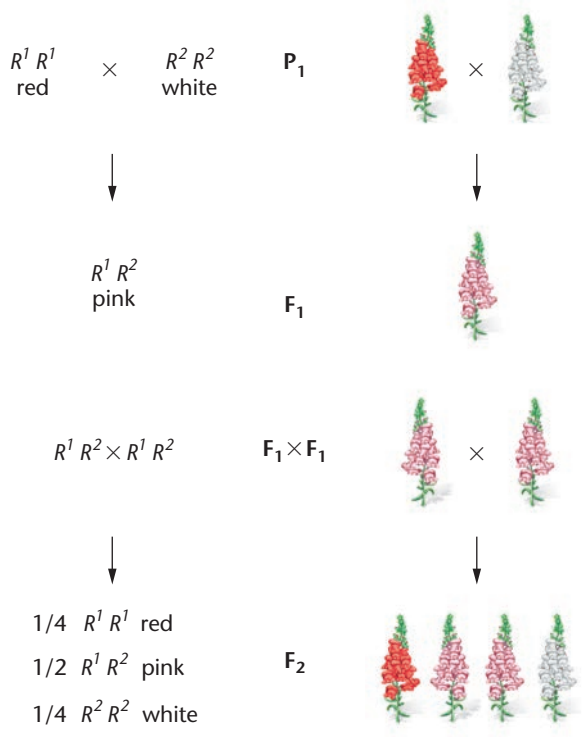


FIGURE 4–1 Incomplete dominance shown in the flower color of snapdragons.

4.4 In Codominance, the Influence of Both Alleles in a Heterozygote Is Clearly Evident

If two alleles of a single gene are responsible for producing two distinct, detectable gene products, a situation different from incomplete dominance or dominance/recessiveness arises. In this case, *the joint expression of both alleles in a heterozygote* is called **codominance.** The **MN blood group** in humans illustrates this phenomenon and is characterized by an antigen called a glycoprotein, found on the surface of red blood cells. In the human population, two forms of this glycoprotein exist, designated M and N; an individual may exhibit either one or both of them.

The MN system is under the control of an autosomal locus found on chromosome 4 and two alleles designated L^M and L^N. Humans are diploid, so three combinations are possible, each resulting in a distinct blood type:

Genotype	Phenotype
$L^M L^M$	M
$L^M L^N$	MN
$L^N L^N$	N

As predicted, a mating between two heterozygous MN parents may produce children of all three blood types, as follows:

$$L^M L^N \times L^M L^N$$
$$\downarrow$$
$$1/4\ L^M L^M$$
$$1/2\ L^M L^N$$
$$1/4\ L^N L^N$$

Once again the genotypic ratio, 1:2:1, is upheld.

Codominant inheritance is characterized by *distinct expression of the gene products of both alleles*. This characteristic distinguishes it from incomplete dominance, where heterozygotes express an intermediate, blended phenotype. We shall see another example of codominance when we examine the ABO blood-type system in the following section.

4.5 Multiple Alleles of a Gene May Exist in a Population

The information stored in any gene is extensive, and mutations can modify this information in many ways. Each change produces a different allele. Therefore, for any specific gene, the number of alleles within members of a population need not be restricted to two. When three or more alleles of the same gene are found, **multiple alleles** are present that create a unique mode of inheritance. It is important to realize that *multiple alleles can be studied only in populations.* An individual diploid organism has, at most, two homologous gene loci that may be occupied by different alleles of the same gene. However, among many members of a species, numerous alternative forms of the same gene can exist.

The ABO Blood Group

The simplest case of multiple alleles is that in which three alternative alleles of one gene exist. This situation is illustrated by the **ABO blood group** in humans, discovered by Karl Landsteiner in the early 1900s. The ABO system, like the MN blood group, is characterized by the presence of antigens on the surface of red blood cells. The A and B antigens are distinct from MN antigens and are under the

control of a different gene, located on chromosome 9. As in the MN system, one combination of alleles in the ABO system exhibits a codominant mode of inheritance.

When individuals are tested using antisera that contain antibodies against the A or B antigen, four phenotypes are revealed. Each individual has either the A antigen (A phenotype), the B antigen (B phenotype), the A and B antigens (AB phenotype), or neither antigen (O phenotype). In 1924, it was hypothesized that these phenotypes were inherited as the result of three alleles of a single gene. This hypothesis was based on studies of the blood types of many different families.

Although different designations can be used, we use the symbols I^A, I^B, and i to distinguish these three alleles; the i designation stands for *isoagglutinogen*, another term for antigen. If we assume that the I^A and I^B alleles are responsible for the production of their respective A and B antigens and that i is an allele that does not produce any detectable A or B antigens, we can list the various genotypic possibilities and assign the appropriate phenotype to each:

Genotype	Antigen	Phenotype
$I^A I^A$	A	A
$I^A i$	A	
$I^B I^B$	B	B
$I^B i$	B	
$I^A I^B$	A, B	AB
$i i$	Neither	O

In these assignments the I^A and I^B alleles are dominant to the i allele but are codominant to each other. Our knowledge of human blood types has several practical applications, the most important of which are compatible blood transfusions and organ transplantations.

The Bombay Phenotype

The biochemical basis of the ABO blood-type system has been carefully worked out. The A and B antigens are actually carbohydrate groups (sugars) that are bound to lipid molecules (fatty acids) protruding from the membrane of the red blood cell. The specificity of the A and B antigens is based on the terminal sugar of the carbohydrate group. Both the A and B antigens are derived from a precursor molecule called the **H substance,** to which one or two terminal sugars are added.

In extremely rare instances, first recognized in a woman in Bombay in 1952, the H substance is incompletely formed. As a result, it is an inadequate substrate for the enzyme that normally adds the terminal sugar. This condition results in the expression of blood type O and is called the **Bombay phenotype.** Research has revealed that this condition is due to a rare recessive mutation at a locus separate from that controlling the A and B antigens. The gene is now designated *FUT1* (encoding an enzyme, fucosyl transferase),

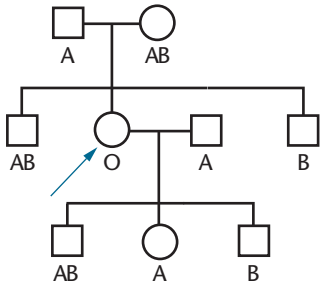


FIGURE 4–2 A partial pedigree of a woman with the Bombay phenotype. Functionally, her ABO blood group behaves as type O. Genetically, she is type B.

and individuals that are homozygous for the mutation cannot synthesize the complete H substance. Thus, even though they may have the I^A and/or I^B alleles, neither the A nor B antigen can be added to the cell surface. This information explains why the woman in Bombay expressed blood type O, even though one of her parents was type AB (thus she should not have been type O), and why she was able to pass the I^B allele to her children (**Figure 4–2**).

The *white* Locus in *Drosophila*

Many other phenotypes in plants and animals are known to be controlled by multiple allelic inheritance. In *Drosophila,* many alleles are known at practically every locus.

NOW SOLVE THIS

4–1 In the guinea pig, one locus involved in the control of coat color may be occupied by any of four alleles: C (full color), c^k (sepia), c^d (cream), or c^a (albino), with an order of dominance of: $C > c^k > c^d > c^a$. (C is dominant to all others, c^k is dominant to c^d and c^a, but not C, etc.) In the following crosses, determine the parental genotypes and predict the phenotypic ratios that would result:

(a) sepia × cream, where both guinea pigs had an albino parent

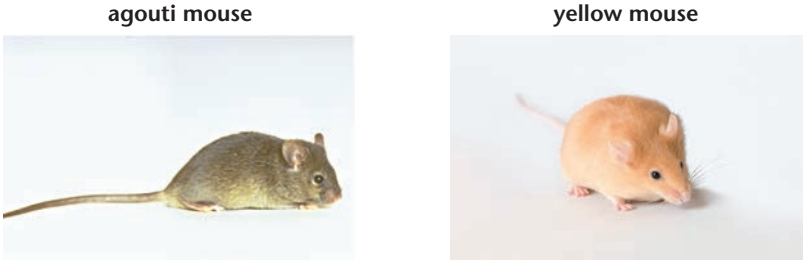
(b) sepia × cream, where the sepia guinea pig had an albino parent and the cream guinea pig had two sepia parents

(c) sepia × cream, where the sepia guinea pig had two full-color parents and the cream guinea pig had two sepia parents

(d) sepia × cream, where the sepia guinea pig had a full-color parent and an albino parent and the cream guinea pig had two full-color parents

HINT: *This problem involves an understanding of multiple alleles. The key to its solution is to note particularly the hierarchy of dominance of the various alleles. Remember also that even though there can be more than two alleles in a population, an individual can have at most two of these. Thus, the allelic distribution into gametes adheres to the principle of segregation.*

FIGURE 4–3 An agouti and a yellow mouse.



The recessive mutation that causes white eyes, discovered by Thomas H. Morgan and Calvin Bridges in 1912, is one of over 100 alleles that can occupy this locus. In this allelic series, eye colors range from complete absence of pigment in the *white* allele to deep ruby in the *white-satsuma* allele, to orange in the *white-apricot* allele, to a buff color in the *white-buff* allele. These alleles are designated w, w^{sat}, w^{a}, and w^{bf}, respectively. In each case, the total amount of pigment in these mutant eyes is reduced to less than 20 percent of that found in the brick-red, wild-type eye.

4.6 Lethal Alleles Represent Essential Genes

Many gene products are essential to an organism's survival. Mutations resulting in the synthesis of a gene product that is nonfunctional can often be tolerated in the heterozygous state; that is, one wild-type allele may be sufficient to produce enough of the essential product to allow survival. However, such a mutation behaves as a *recessive lethal allele,* and homozygous recessive individuals will not survive. The time of death will depend on when the product is essential. In mammals, for example, this might occur during development, early childhood, or even adulthood.

In some cases, the allele responsible for a lethal effect when homozygous may also result in a distinctive mutant phenotype when present heterozygously. It is behaving as a recessive lethal allele but is dominant with respect to the phenotype. For example, a mutation that causes a yellow coat in mice was discovered in the early part of this century. The yellow coat varies from the normal agouti (wild-type) coat phenotype, as shown in **Figure 4–3**. Crosses between the various combinations of the two strains yield unusual results:

		Crosses		
(A) agouti	×	agouti	⟶	all agouti
(B) yellow	×	yellow	⟶	2/3 yellow: 1/3 agouti
(C) agouti	×	yellow	⟶	1/2 yellow: 1/2 agouti

These results are explained on the basis of a single pair of alleles. With regard to coat color, the mutant *yellow* allele (A^{Y}) is dominant to the wild-type *agouti* allele (A), so heterozygous mice will have yellow coats. However, the *yellow* allele is also a homozygous recessive lethal. When present in two copies, the mice die before birth. Thus, there are no homozygous yellow mice. The genetic basis for these three crosses is shown in Figure 4–3.

In other cases, a mutation may behave as a *dominant lethal allele.* In such cases, the presence of just one copy of the allele results in the death of the individual. In humans, a disorder called **Huntington disease** (previously referred to as Huntington's chorea) is due to a dominant autosomal allele H, where the onset of the disease in heterozygotes (Hh) is delayed, usually well into adulthood. Affected individuals then undergo gradual nervous and motor degeneration until they die. This lethal disorder is particularly tragic because it has such a late onset, typically at about age 40. By that time, the affected individual may have produced a family, and each of the children has a 50 percent probability of inheriting the lethal allele, transmitting the allele to his or her offspring, and eventually developing the disorder. The American folk singer and composer Woody Guthrie (father of modern-day folk singer Arlo Guthrie) died from this disease at age 39.

Dominant lethal alleles are rarely observed. For these alleles to exist in a population, the affected individuals must reproduce before the lethal allele is expressed, as can occur in Huntington disease. If all affected individuals die before reaching reproductive age, the mutant gene will not

EVOLVING CONCEPT OF THE GENE

Based on the work of many geneticists following the rediscovery of Mendel's work in the very early part of the twentieth century, the chromosome theory of inheritance was put forward, which hypothesized that chromosomes are the carriers of genes and that meiosis is the physical basis of Mendel's postulates. In the ensuing 40 years, the concept of a gene evolved to reflect the idea that this hereditary unit can exist in multiple forms, or alleles, each of which can have an impact on the phenotype in different ways, leading to incomplete dominance, codominance, and even lethality. It became clear that the process of mutation was the source of new alleles. ■

be passed to future generations, and the mutation will disappear from the population unless it arises again as a result of a new mutation.

ESSENTIAL POINT

Since Mendel's work was rediscovered, transmission genetics has been expanded to include many alternative modes of inheritance, including the study of incomplete dominance, codominance, multiple alleles, and lethal alleles. ■

4.7 Combinations of Two Gene Pairs with Two Modes of Inheritance Modify the 9:3:3:1 Ratio

Each example discussed so far modifies Mendel's 3:1 F_2 monohybrid ratio. Therefore, combining any two of these modes of inheritance in a dihybrid cross will likewise modify the classical 9:3:3:1 ratio. Having established the foundation for the modes of inheritance of incomplete dominance, codominance, multiple alleles, and lethal alleles, we can now deal with the situation of two modes of inheritance occurring simultaneously. Mendel's principle of independent assortment applies to these situations, provided that the genes controlling each character are not linked on the same chromosome—in other words, that they do not demonstrate what is called *genetic linkage*.

Consider, for example, a mating that occurs between two humans who are both heterozygous for the autosomal recessive gene that causes albinism and who are both of blood type AB. What is the probability of a particular phenotypic combination occurring in each of their children? Albinism is inherited in the simple Mendelian fashion, and the blood types are determined by the series of three multiple alleles, I^A, I^B, and I. The solution to this problem is diagrammed in **Figure 4–4**, using the forked-line method. This dihybrid cross does not

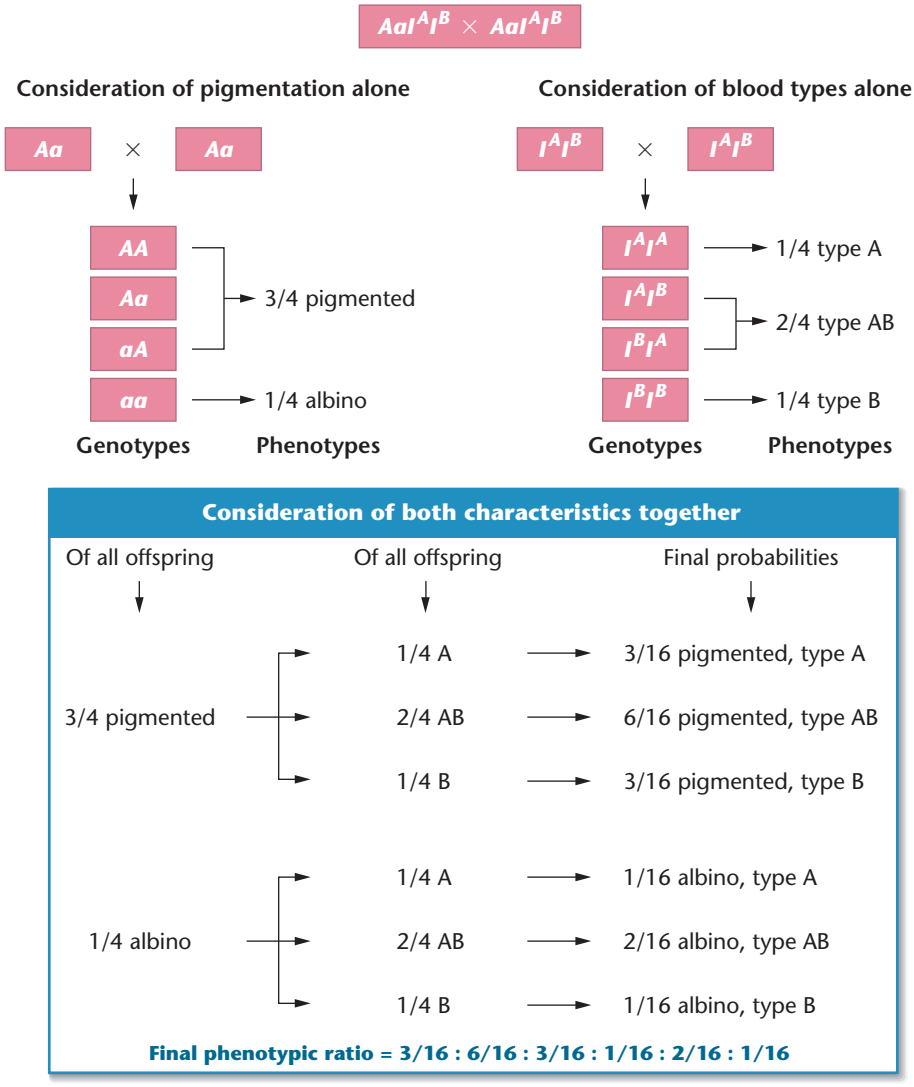


FIGURE 4–4 Calculation of the mating probabilities involving the ABO blood type and albinism in humans, using the forked-line method.

yield the classical four phenotypes in a 9:3:3:1 ratio. Instead, six phenotypes occur in a 3:6:3:1:2:1 ratio, establishing the expected probability for each phenotype. This is just one of the many variants of modified ratios that are possible when different modes of inheritance are combined.

4.8 Phenotypes Are Often Affected by More Than One Gene

Soon after Mendel's work was rediscovered, experimentation revealed that individual characteristics displaying discrete phenotypes are often under the control of more than one gene. This was a significant discovery because it revealed that genetic influence on the phenotype is often much more complex than Mendel had envisioned. Instead of single genes controlling the development of individual parts of the plant or animal body, it soon became clear that phenotypic characters can be influenced by the interactions of many different genes and their products.

The term **gene interaction** is often used to describe the idea that several genes influence a particular characteristic. This does not mean, however, that two or more genes, or their products, necessarily interact directly with one another to influence a particular phenotype. Rather, the cellular function of numerous gene products contributes to the development of a common phenotype. For example, the development of an organ such as the compound eye of an insect is exceedingly complex and leads to a structure with multiple phenotypic manifestations—such as specific size, shape, texture, and color. The formation of the eye results from a complex cascade of events during its development. This process exemplifies the developmental concept of **epigenesis,** whereby each step of development increases the complexity of this sensory organ and is under the control and influence of one or more genes.

An enlightening example of epigenesis and multiple gene interaction involves the formation of the inner ear in mammals. The inner ear consists of distinctive anatomical features to capture, funnel, and transmit external sound waves and to convert them into nerve impulses. During the formation of the ear, a cascade of intricate developmental events occur, influenced by many genes. Mutations that interrupt many of the steps of ear development lead to a common phenotype: **hereditary deafness.** In a sense, these many genes "interact" to produce a common phenotype. In such situations, the mutant phenotype is described as a **heterogeneous trait,** reflecting the many genes involved. In humans, while a few common alleles are responsible for the vast majority of cases of hereditary deafness, over 50 genes are involved in development of the ability to discern sound.

Epistasis

Some of the best examples of gene interaction are those that reveal the phenomenon of **epistasis** where the expression of one gene or gene pair masks or modifies the expression of another gene or gene pair. Sometimes the genes involved control the expression of the same general phenotypic characteristic in an antagonistic manner, as when masking occurs. In other cases, however, the genes involved exert their influence on one another in a complementary, or cooperative, fashion.

For example, the homozygous presence of a recessive allele prevents or overrides the expression of other alleles at a second locus (or several other loci). In this case, the alleles at the first locus are said to be **epistatic** to those at the second locus, and the alleles at the second locus are **hypostatic** to those at the first locus. In another example, a single dominant allele at the first locus influences the expression of the alleles at a second gene locus. In a third example, two gene pairs are said *to complement one another* such that at least one dominant allele at each locus is required to express a particular phenotype.

The Bombay phenotype discussed earlier is an example of the homozygous recessive condition at one locus masking the expression of a second locus. There, we established that the homozygous presence of the mutant form of the *FUT1* gene masks the expression of the I^A and I^B alleles. Only individuals containing at least one wild-type *FUT1* allele can form the A or B antigen. As a result, individuals whose genotypes include the I^A or I^B allele and who lack a wild-type allele are of the type O phenotype, regardless of their potential to make either antigen. An example of the outcome of matings between individuals heterozygous at both loci is illustrated in **Figure 4–5**. If many such individuals have children, the phenotypic ratio of 3 A: 6 AB: 3 B: 4 O is expected in their offspring.

It is important to note the following points when examining this cross and the predicted phenotypic ratio:

1. A key distinction exists in this cross compared to the modified dihybrid cross shown in Figure 4–4: *only one characteristic—blood type—is being followed*. In the modified dihybrid cross of Figure 4–4, blood type *and* skin pigmentation are followed as separate phenotypic characteristics.

2. Even though only a single character was followed, the phenotypic ratio is expressed in sixteenths. If we knew nothing about the H substance and the genes controlling it, we could still be confident that a second gene pair, other than that controlling the A and B antigens, is involved in the phenotypic expression. *When studying a single character, a ratio that is expressed in 16 parts (e.g., 3:6:3:4) suggests that two gene pairs are "interacting" during the expression of the phenotype under consideration.*

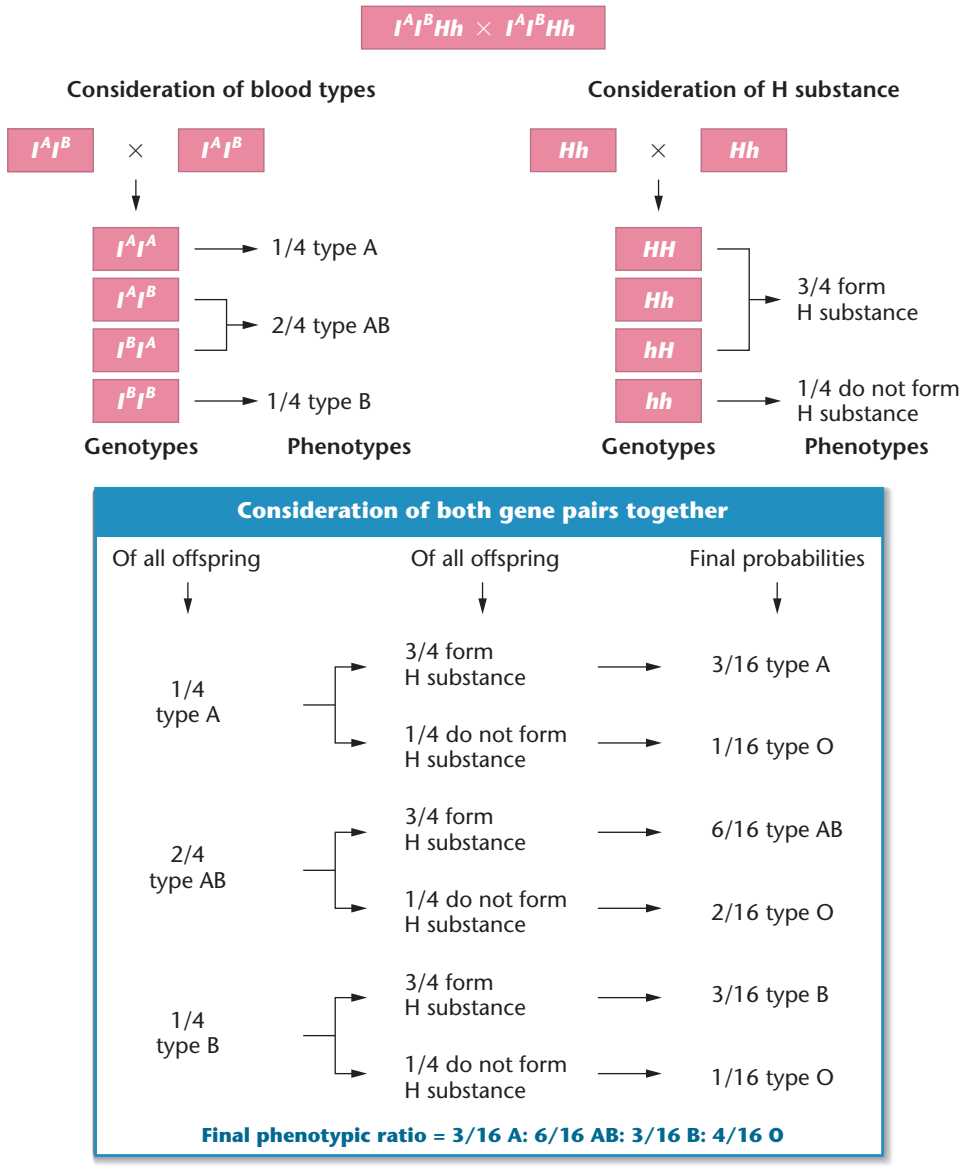
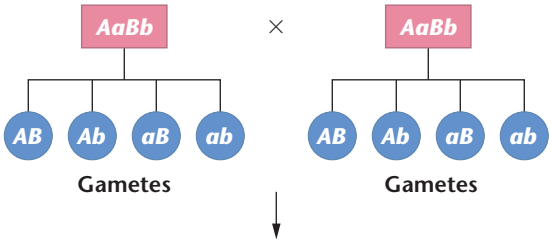


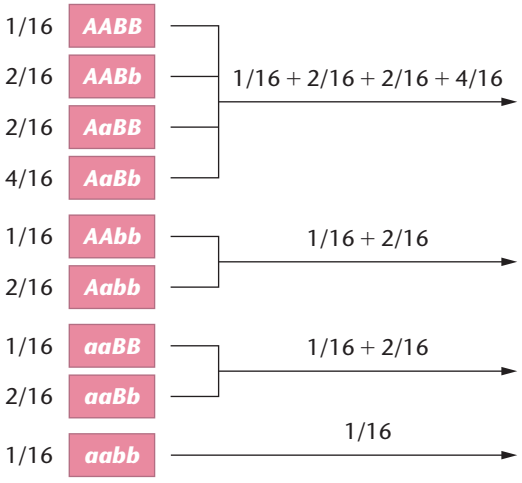
FIGURE 4–5 The outcome of a mating between individuals who are heterozygous at two genes determining their ABO blood type. Final phenotypes are calculated by considering both genes separately and then combining the results using the forked-line method.

The study of gene interaction reveals inheritance patterns that modify the classical Mendelian dihybrid F_2 ratio (9:3:3:1) in other ways as well. In these examples, epistasis combines one or more of the four phenotypic categories in various ways. The generation of these four groups is reviewed in Figure 4–6, along with several modified ratios.

As we discuss these and other examples, we will make several assumptions and adopt certain conventions:

1. In each case, distinct phenotypic classes are produced, each clearly discernible from all others. Such traits illustrate *discontinuous variation,* where phenotypic categories are discrete and qualitatively different from one another.
2. The genes considered in each cross are not linked and therefore assort independently of one another during gamete formation. To allow you to easily compare the results of different crosses, we designated alleles as *A*, *a* and *B*, *b* in each case.
3. When we assume that complete dominance exists between the alleles of any gene pair, such that *AA* and *Aa* or *BB* and *Bb* are equivalent in their genetic effects, we use the designations *A*− or *B*− for both combinations, where the dash (−) indicates that either allele may be present, without consequence to the phenotype.





Dihybrid ratio	Modified ratios				
9/16 *A–B–*	9/16	12/16	9/16	9/16	15/16
3/16 *A–bb*	3/16		7/16	6/16	
3/16 *aaB–*	4/16	3/16			
1/16 *aabb*		1/16		1/16	1/16

FIGURE 4–6 Generation of the various modified dihybrid ratios from the nine unique genotypes produced in a cross between individuals who are heterozygous at two genes.

4. All P_1 crosses involve homozygous individuals (e.g., *AABB* × *aabb, AAbb* × *aaBB,* or *aaBB* × *AAbb*). Therefore, each F_1 generation consists of only heterozygotes of genotype *AaBb*.

5. In each example, the F_2 generation produced from these heterozygous parents is our main focus of analysis. When two genes are involved (as in Figure 4–6), the F_2 genotypes fall into four categories: 9/16 *A–B–*, 3/16 *A–bb*, 3/16 *aaB–*, and 1/6 *aabb*. Because of dominance, all genotypes in each category have an equivalent effect on the phenotype.

Case 1 is the inheritance of coat color in mice (**Figure 4–7**). Normal wild-type coat color is agouti, a grayish pattern formed by alternating bands of pigment on each hair. Agouti is dominant to black (non-agouti) hair, which is caused by a recessive mutation, *a*. Thus, *A–* results in agouti, while *aa* yields black coat color. When it is homozygous, a recessive mutation, *b*, at a separate locus, eliminates pigmentation altogether, yielding albino mice (*bb*), regardless of the genotype at the other locus. The presence of at least one *B* allele allows pigmentation to occur in much the same way that the *H* allele in humans allows the expression of the ABO blood types. In a cross between agouti (*AABB*) and albino (*aabb*), members of the F_1 are all *AaBb* and have agouti coat color. In the F_2 progeny of a cross between two F_1 heterozygotes, the following genotypes and phenotypes are observed:

F_1: *AaBb* × *AaBb*

↓

F_2 Ratio	Genotype	Phenotype	Final Phenotypic Ratio
9/16	*A–B–*	agouti	9/16 agouti
3/16	*A–bb*	albino	3/16 black
3/16	*aaB–*	black	4/16 albino
1/16	*aabb*	albino	

We can envision gene interaction yielding the observed 9:3:4 F_2 ratio as a two-step process:

	Gene *B*		Gene *A*	
Precursor Molecule (colorless)	↓ → *B–*	Black Pigment	↓ → *A–*	Agouti Pattern

In the presence of a *B* allele, black pigment can be made from a colorless substance. In the presence of an *A* allele, the black pigment is deposited during the development of hair in a pattern that produces the agouti phenotype. If the *aa* genotype occurs, all of the hair remains black. If the *bb* genotype occurs, no black pigment is produced, regardless of the presence of the *A* or *a* alleles, and the mouse is albino.

Case	Organism	Character	F_2 Phenotypes				Modified ratio
			9/16	3/16	3/16	1/16	
1	Mouse	Coat color	agouti	albino	black	albino	9:3:4
2	Squash	Color	white		yellow	green	12:3:1
3	Pea	Flower color	purple	white			9:7
4	Squash	Fruit shape	disc	sphere		long	9:6:1
5	Chicken	Color	white		colored	white	13:3
6	Mouse	Color	white-spotted	white	colored	white-spotted	10:3:3
7	Shepherd's purse	Seed capsule	triangular			ovoid	15:1
8	Flour beetle	Color	6/16 sooty and 3/16 red	black	jet	black	6:3:3:4

FIGURE 4–7 The basis of modified dihybrid F_2 phenotypic ratios, resulting from crosses between doubly heterozygous F_1 individuals. The four groupings of the F_2 genotypes shown in Figure 4–6 and across the top of this figure are combined in various ways to produce these ratios.

Therefore, the *bb* genotype masks or suppresses the expression of the *A* gene. As a result, this is referred to as *recessive epistasis*.

A second type of epistasis, called *dominant epistasis,* occurs when a dominant allele at one genetic locus masks the expression of the alleles at a second locus. For instance, Case 2 of Figure 4–7 deals with the inheritance of fruit color in summer squash. Here, the dominant allele *A* results in white fruit color regardless of the genotype at a second locus, *B*. In the absence of the dominant *A* allele (the *aa* genotype), *BB* or *Bb* results in yellow color, while *bb* results in green color. Therefore, if two white-colored double heterozygotes (*AaBb*) are crossed, this type of epistasis generates an interesting phenotypic ratio:

F_1: *AaBb* × *AaBb*

↓

F_2 Ratio	Genotype	Phenotype	Final Phenotypic Ratio
9/16	*A—B—*	white	12/16 white
3/16	*A—bb*	white	
3/16	*aaB—*	yellow	13/16 yellow
1/16	*aabb*	green	1/16 green

Of the offspring, 9/16 are *A—B—* and are thus white. The 3/16 bearing the genotypes *A—bb* are also white. Finally, 3/16 are yellow (*aaB—*) while 1/16 are green (*aabb*); and we obtain the modified ratio of 12:3:1.

Our third type of gene interaction (Case 3 of Figure 4–7) was first discovered by William Bateson and Reginald Punnett (of Punnett square fame). It is demonstrated in a cross between two true-breeding strains of white-flowered sweet peas. Unexpectedly, the results of this cross yield all purple F_1 plants, and the F_2 plants occur in a ratio of 9/16 purple to 7/16 white. The proposed explanation suggests that the presence of at least one dominant allele of each of two gene pairs is essential for flowers to be purple. Thus, this cross represents a case of *complementary gene interaction*. All other genotype combinations yield white flowers because the homozygous condition of *either* recessive allele masks the expression of the dominant allele at the other locus. The cross is shown as follows:

P_1: *AAbb* × *aaBB*

white white

↓

F_1: All *AaBb* purple

↓

F_2 Ratio	Genotype	Phenotype	Final Phenotypic Ratio
9/16	*A—B—*	purple	
3/16	*A—bb*	white	9/16 purple
3/16	*aa B—*	white	7/16 white
1/16	*aabb*	white	

We can now see how two gene pairs might yield such results:

	Gene *A*		Gene *B*	
Precursor	↓	Intermediate	↓	Final
Substance	⟶	Product	⟶	Product
(colorless)	*A—*	(colorless)	*B—*	(purple)

At least one dominant allele from each pair of genes is necessary to ensure both biochemical conversions to the final product, yielding purple flowers. In our cross, this will occur in 9/16 of the F_2 offspring. All other plants (7/16) have flowers that remain white.

The preceding examples illustrate how the products of two genes "interact" to influence the development of a common phenotype. In other instances, more than two genes and their products are involved in controlling phenotypic expression.

Novel Phenotypes

Other cases of gene interaction yield novel, or new, phenotypes in the F_2 generation, in addition to producing modified dihybrid ratios. Case 4 in Figure 4–7 depicts the inheritance of fruit shape in the summer squash *Cucurbita pepo*. When plants with disc-shaped fruit (*AABB*) are crossed to plants with long fruit (*aabb*), the F_1 generation all have disc fruit. However, in the F_2 progeny, fruit with a novel shape—sphere—appear, along with fruit exhibiting the parental phenotypes. A variety of fruit shapes are shown in **Figure 4–8**.

The F_2 generation, with a modified 9:6:1 ratio, is generated as follows:

F_1: *AaBb* × *AaBb*

disc disc

↓

F_2 Ratio	Genotype	Phenotype	Final Phenotypic Ratio
9/16	*A—B—*	disc	9/16 disc
3/16	*A—bb*	sphere	6/16 sphere
3/16	*aaB—*	sphere	
1/16	*aabb*	long	1/16 long

In this example of gene interaction, both gene pairs influence fruit shape equally. A dominant allele at either locus ensures a sphere-shaped fruit. In the absence of dominant alleles, the fruit is long. However, if both dominant alleles (*A* and *B*) are present, the fruit displays a flattened, disc shape.

FIGURE 4–8 Summer squash exhibiting the fruit-shape phenotypes disc, long, and sphere.

Other Modified Dihybrid Ratios

The remaining cases (5–8) in Figure 4–7 show additional modifications of the dihybrid ratio and provide still other examples of gene interactions. However, all eight cases have two things in common. First, we have not violated the principles of segregation and independent assortment to explain the inheritance pattern of each case. Therefore, the added complexity of inheritance in these examples does not detract from the validity of Mendel's conclusions. Second, the F_2 phenotypic ratio in each example has been expressed in sixteenths. When similar observations are made in crosses where the inheritance pattern is unknown, it suggests to geneticists that two gene pairs are controlling the observed phenotypes. You should make the same inference in your analysis of genetics problems.

NOW SOLVE THIS

4–2 In some plants a red flower pigment, cyanidin, is synthesized from a colorless precursor. The addition of a hydroxyl group (OH^-) to the cyanidin molecule causes it to become purple. In a cross between two randomly selected purple varieties, the following results were obtained:

94 purple
31 red
43 white

How many genes are involved in the determination of these flower colors? Which genotypic combinations produce which phenotypes? Diagram the purple × purple cross.

HINT: *This problem describes a plant in which flower color, a single characteristic, can take on one of three variations. The key to its solution is to first analyze the raw data and convert the numbers to a meaningful ratio. This will guide you in determining how many gene pairs are involved. Then you can group the genotypes in a way that corresponds to the phenotypic ratio.*

ESSENTIAL POINT

Mendel's classic F_2 ratio is often modified in instances when gene interaction controls phenotypic variation. Such instances can be identified when the final ratio is divided into eighths or sixteenths.

4.9 Complementation Analysis Can Determine if Two Mutations Causing a Similar Phenotype Are Alleles of the Same Gene

An interesting situation arises when two mutations, both of which produce a similar phenotype, are isolated independently. Suppose that two investigators independently

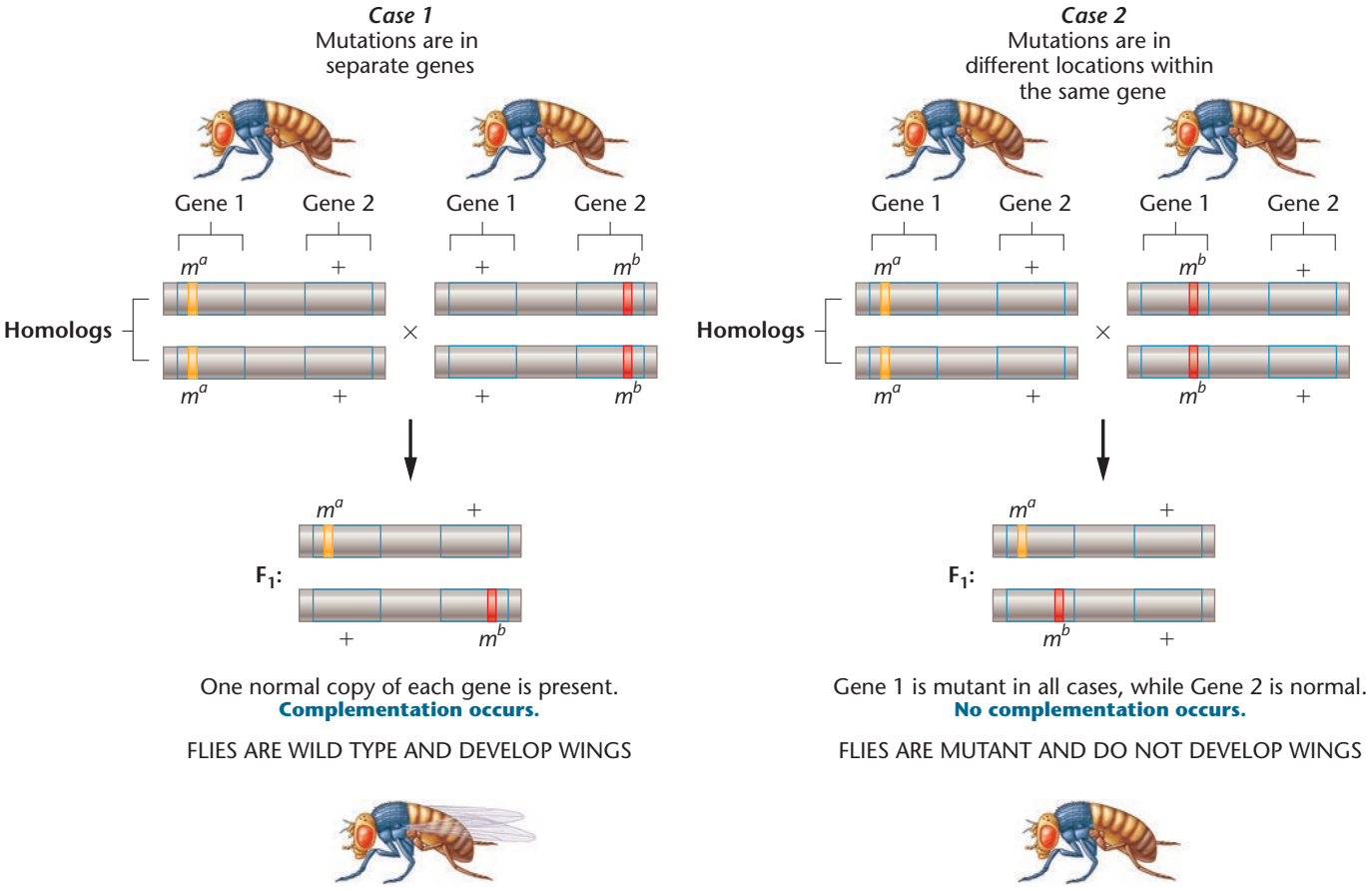


FIGURE 4–9 Complementation analysis of alternative outcomes of two wingless mutations in *Drosophila* (m^a and m^b). In Case 1, the mutations are not alleles of the same gene, whereas in Case 2, the mutations are alleles of the same gene.

isolate and establish a true-breeding strain of wingless *Drosophila* and demonstrate that each mutant phenotype is due to a recessive mutation. We might assume that both strains contain mutations in the same gene. However, since we know that many genes are involved in the formation of wings, mutations in any one of them might inhibit wing formation during development. The experimental approach called **complementation analysis** allows us to determine whether two such mutations are in the same gene—that is, whether they are alleles of the same gene or whether they represent mutations in separate genes.

To repeat, our analysis seeks to answer this simple question: *Are two mutations that yield similar phenotypes present in the same gene or in two different genes*? To find the answer, we cross the two mutant strains and analyze the F_1 generation. Two alternative outcomes and interpretations of this cross are shown in **Figure 4–9**. We discuss both cases, using the designations m^a for one of the mutations and m^b for the other one. Now we will determine experimentally whether or not m^a and m^b are alleles of the same gene.

Case 1. *All offspring develop normal wings.*
Interpretation: The two recessive mutations are in separate genes and are not alleles of one another. Following the cross, all F_1 flies are heterozygous for both genes. *Complementation* is said to occur. Since each mutation is in a separate gene and each F_1 fly is heterozygous at both loci, the normal products of both genes are produced (by the one normal copy of each gene), and wings develop.

Case 2. *All offspring fail to develop wings.*
Interpretation: The two mutations affect the same gene and are alleles of one another. Complementation does *not* occur. Since the two mutations affect the same gene, the F_1 flies are homozygous for the two mutant alleles (the m^a allele and the m^b allele). No normal product of the gene is produced, and in the absence of this essential product, wings do not form.

Complementation analysis, as originally devised by the Nobel Prize-winning *Drosophila* geneticist Edward B. Lewis, may be used to screen any number of individual mutations that result in the same phenotype. Such an analysis may reveal that only a single gene is involved or that two or more genes are involved. All mutations determined to be present in any single gene are said to fall into the same **complementation group,** and they will complement

mutations in all other groups. When large numbers of mutations affecting the same trait are available and studied using complementation analysis, it is possible to predict the total number of genes involved in the determination of that trait.

ESSENTIAL POINT

Complementation analysis determines whether independently isolated mutations producing similar phenotypes are alleles of one another or whether they represent separate genes. ■

4.10 Expression of a Single Gene May Have Multiple Effects

While the previous sections have focused on the effects of two or more genes on a single characteristic, the converse situation, where expression of a single gene has multiple phenotypic effects, is also quite common. This phenomenon, which often becomes apparent when phenotypes are examined carefully, is referred to as **pleiotropy.** We will review two such cases involving human genetic disorders to illustrate this point.

Marfan syndrome is a human malady resulting from an autosomal dominant mutation in the gene encoding the connective tissue protein fibrillin. Because this protein is widespread in many tissues in the body, one would expect multiple effects of such a defect. In fact, fibrillin is important to the structural integrity of the lens of the eye, to the lining of vessels such as the aorta, and to bones, among other tissues. As a result, the phenotype associated with Marfan syndrome includes lens dislocation, increased risk of aortic aneurysm, and lengthened long bones in limbs. This disorder is of historical interest in that speculation abounds that Abraham Lincoln was afflicted.

Our second example involves another human autosomal dominant disorder, **porphyria variegata.** Afflicted individuals cannot adequately metabolize the porphyrin component of hemoglobin when this respiratory pigment is broken down as red blood cells are replaced. The accumulation of excess porphyrins is immediately evident in the urine, which takes on a deep red color. The severe features of the disorder are due to the toxicity of the buildup of porphyrins in the body, particularly in the brain. Complete phenotypic characterization includes abdominal pain, muscular weakness, fever, a racing pulse, insomnia, headaches, vision problems (that can lead to blindness), delirium, and ultimately convulsions. As you can see, deciding which phenotypic trait best characterizes the disorder is impossible.

Like Marfan syndrome, porphyria variegata is also of historical significance. George III, king of England during the American Revolution, is believed to have suffered from episodes involving all of the above symptoms. He ultimately became blind and senile prior to his death. We could cite many other examples to illustrate pleiotropy, but suffice it to say that if one looks carefully, most mutations display more than a single manifestation when expressed.

ESSENTIAL POINT

Pleiotropy refers to multiple phenotypic effects caused by a single mutation. ■

4.11 X-Linkage Describes Genes on the X Chromosome

In many animal and some plant species, one of the sexes contains a pair of unlike chromosomes that are involved in sex determination. In many cases, these are designated as the X and Y. For example, in both *Drosophila* and humans, males contain an X and a Y chromosome, whereas females contain two X chromosomes. While the Y chromosome must contain a region of pairing homology with the X chromosome if the two are to synapse and segregate during meiosis, much of the remainder of the Y chromosome in humans and other species is considered to be relatively inert genetically. Thus, it lacks most genes that are present on the X chromosome. As a result, genes present on the X chromosome exhibit unique patterns of inheritance in comparison with autosomal genes. The term **X-linkage** is used to describe these situations.

In the following discussion, we will focus on inheritance patterns resulting from genes present on the X but absent from the Y chromosome. This situation results in a modification of Mendelian ratios, the central theme of this chapter.

X-Linkage in *Drosophila*

One of the first cases of X-linkage was documented by Thomas H. Morgan around 1920 during his studies of the *white* mutation in the eyes of *Drosophila*. The normal wild-type red eye color is dominant to white. We will use this case to illustrate X-linkage.

Morgan's work established that the inheritance pattern of the white-eye trait is clearly related to the sex of the parent carrying the mutant allele. Unlike the outcome of the typical monohybrid cross, reciprocal crosses between white- and red-eyed flies did not yield identical results. In contrast, in all of Mendel's monohybrid crosses, F_1 and F_2 data were similar regardless of which P_1 parent exhibited the recessive mutant trait. Morgan's analysis led to the conclusion that the *white* locus is present on the X chromosome rather than on one of the autosomes. As such, both the gene and the trait are said to be X-linked.

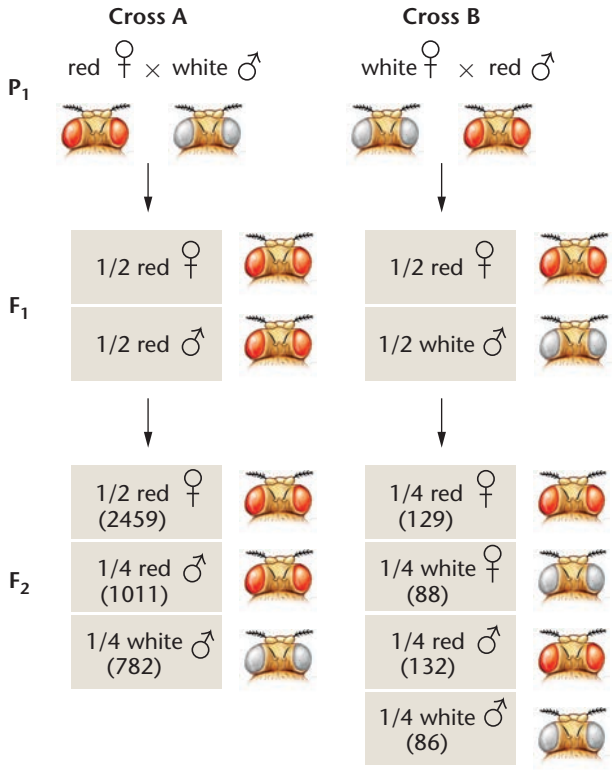


FIGURE 4–10 The F_1 and F_2 results of T. H. Morgan's reciprocal crosses involving the X-linked *white* mutation in *Drosophila melanogaster*. The actual F_2 data are shown in parentheses. Photographs of red and white eyes are shown in Chapter 1, Figure 1–3.

Results of reciprocal crosses between white-eyed and red-eyed flies are shown in **Figure 4–10**. The obvious differences in phenotypic ratios in both the F_1 and F_2 generations are dependent on whether or not the P_1 white-eyed parent was male or female.

Morgan was able to correlate these observations with the difference found in the sex-chromosome composition between male and female *Drosophila*. He hypothesized that the recessive allele for white eyes is found on the X chromosome, but its corresponding locus is absent from the Y chromosome. Females thus have two available gene sites, one on each X chromosome, whereas males have only one available gene site on their single X chromosome.

Morgan's interpretation of X-linked inheritance, shown in **Figure 4–11**, provides a suitable theoretical explanation for his results. Since the Y chromosome lacks homology with most genes on the X chromosome, whatever alleles are present on the X chromosome of the males will be expressed directly in their phenotype. Males cannot be homozygous or heterozygous for X-linked genes, and this condition is referred to as being **hemizygous.**

One result of X-linkage is the **crisscross pattern of inheritance,** whereby phenotypic traits controlled by recessive X-linked genes are passed from homozygous mothers to all sons. This pattern occurs because females exhibiting a recessive trait carry the mutant allele on both X chromosomes. Because male offspring receive one of

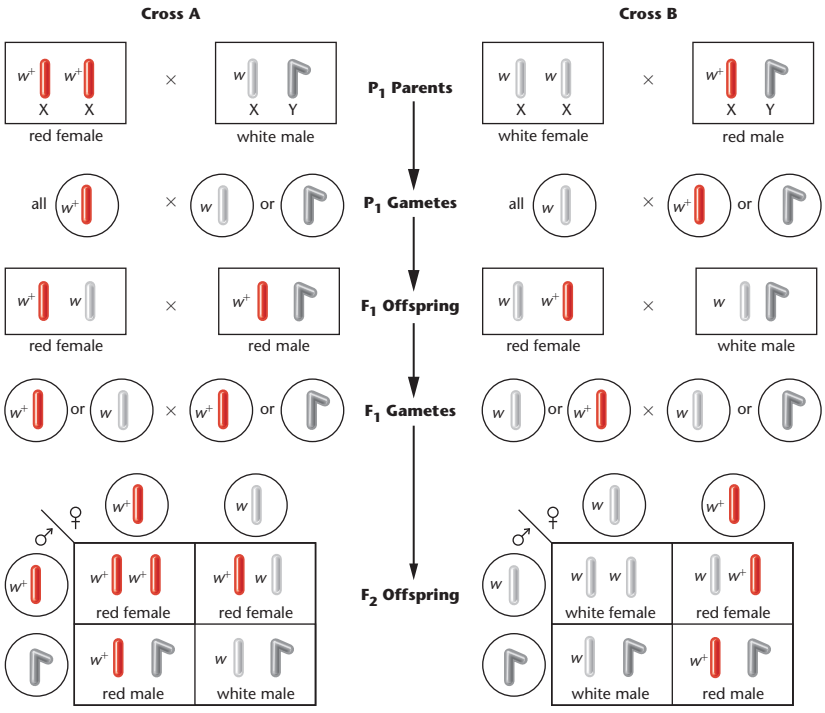


FIGURE 4–11 The chromosomal explanation of the results of the X-linked crosses shown in Figure 4–10.

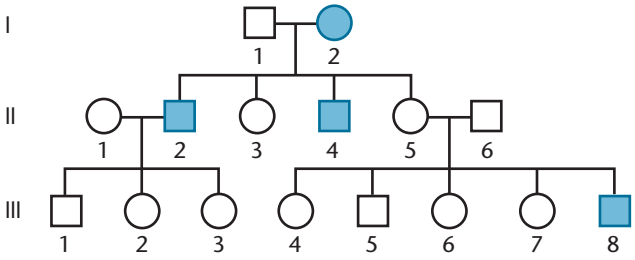


FIGURE 4–12 A human pedigree of the X-linked color-blindness trait. The photograph is of an Ishihara color-blindness chart. Those with normal vision will see the number 15, while those with red-green color blindness will see the number 17.

their mother's two X chromosomes and are hemizygous for all alleles present on that X, all sons will express the same recessive X-linked traits as their mother.

Morgan's work has taken on great historical significance. By 1910, the correlation between Mendel's work and the behavior of chromosomes during meiosis had provided the basis for the **chromosome theory of inheritance,** first introduced in Chapter 3 . Work involving the X chromosome around 1920 is considered to be the first solid experimental evidence in support of this theory. In the ensuing two decades, these findings inspired further research, which provided indisputable evidence in support of this theory.

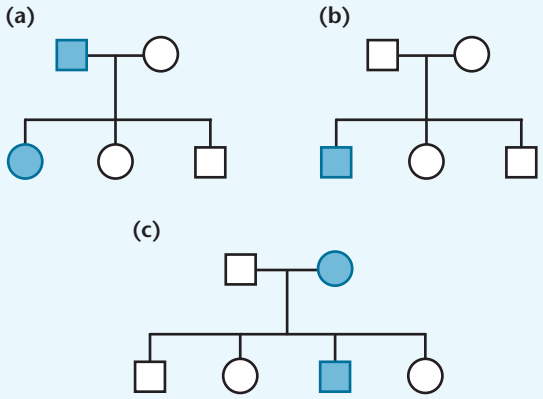
X-Linkage in Humans

In humans, many genes and the respective traits they control are recognized as being linked to the X chromosome. These X-linked traits can be easily identified in a pedigree because of the crisscross pattern of inheritance. A pedigree for one form of human **color blindness** is shown in Figure 4–12. The mother in generation I passes the trait to all her sons but to none of her daughters. If the offspring in generation II marry normal individuals, the color-blind sons will produce all normal male and female offspring (III-1, 2, and 3); the normal-visioned daughters will produce normal-visioned female offspring (III-4, 6, and 7), as well as color-blind (III-8) and normal-visioned (III-5) male offspring.

The way in which X-linked genes are transmitted causes unusual circumstances associated with recessive X-linked disorders, in comparison to recessive autosomal disorders. For example, if an X-linked disorder debilitates or is lethal to the affected individual prior to reproductive maturation, the disorder occurs exclusively in males. This is so because the only sources of the lethal allele in the population are in heterozygous females who are "carriers" and do not express the disorder. They pass the allele to one-half of their sons, who develop the disorder because they are hemizygous but rarely, if ever, reproduce. Heterozygous females also pass the allele to one-half of their daughters, who become carriers but do not develop the disorder. An example of such an X-linked disorder is Duchenne muscular dystrophy. The disease has an onset prior to age 6 and is often lethal around age 20. It normally occurs only in males.

NOW SOLVE THIS

4–3 Below are three pedigrees. For each trait, consider whether it is or is not consistent with X-linked recessive inheritance. In a sentence or two, indicate why or why not.



■ **HINT:** *This problem involves potential X-linked recessive traits as analyzed in pedigrees. The key to its solution is to focus on hemizygosity, where an X-linked recessive allele is always expressed in males, but never passed from a father to his sons. Homozygous females, on the other hand, pass the trait to all sons, but not to their daughters unless the father is also affected.*

ESSENTIAL POINT

Genes located on the X chromosome result in a characteristic mode of genetic transmission referred to as X-linkage, displaying so-called crisscross inheritance, whereby affected mothers pass X-linked traits to all of their sons. ■

4.12 In Sex-Limited and Sex-Influenced Inheritance, an Individual's Sex Influences the Phenotype

In contrast to X-linked inheritance, patterns of gene expression may be affected by the sex of an individual even when the genes are not on the X chromosome. In numerous examples in different organisms, the sex of the individual plays a determining role in the expression of certain phenotypes. In some cases, the expression of a specific phenotype is absolutely limited to one sex; in others, the sex of an individual influences the expression of a phenotype

that is not limited to one sex or the other. This distinction differentiates sex-limited inheritance from sex-influenced inheritance. In both types of inheritance, autosomal genes are responsible for the existence of contrasting phenotypes, but the expression of these genes is dependent on the hormone constitution of the individual. Thus, the heterozygous genotype may exhibit one phenotype in males and the contrasting one in females. In domestic fowl, for example, tail and neck plumage is often distinctly different in males and females (**Figure 4–13**), demonstrating **sex-limited inheritance.** Cock feathering is longer, more curved, and pointed, whereas hen feathering is shorter and less curved. Inheritance of these feather phenotypes is controlled by a single pair of autosomal alleles whose expression is modified by the individual's sex hormones.

As shown in the following chart, hen feathering is due to a dominant allele, *H*, but regardless of the homozygous presence of the recessive *h* allele, all females remain hen-feathered. Only in males does the *hh* genotype result in cock feathering.

Genotype	Phenotype	
	Females	Males
HH	Hen-feathered	Hen-feathered
Hh	Hen-feathered	Hen-feathered
hh	Hen-feathered	Cock-feathered

In certain breeds of fowl, the hen-feathering or cock-feathering allele has become fixed in the population. In the Leghorn breed, all individuals are of the *hh* genotype; as a result, males always differ from females in their plumage. Sebright bantams are all *HH*, resulting in no sexual distinction in feathering phenotypes.

Another example of sex-limited inheritance involves the autosomal genes responsible for milk yield in dairy cattle. Regardless of the overall genotype that influences the quantity of milk production, those genes are obviously expressed only in females.

FIGURE 4–13 Hen feathering (left) and cock feathering (right) in domestic fowl. Note that the hen's feathers are shorter and less curved.

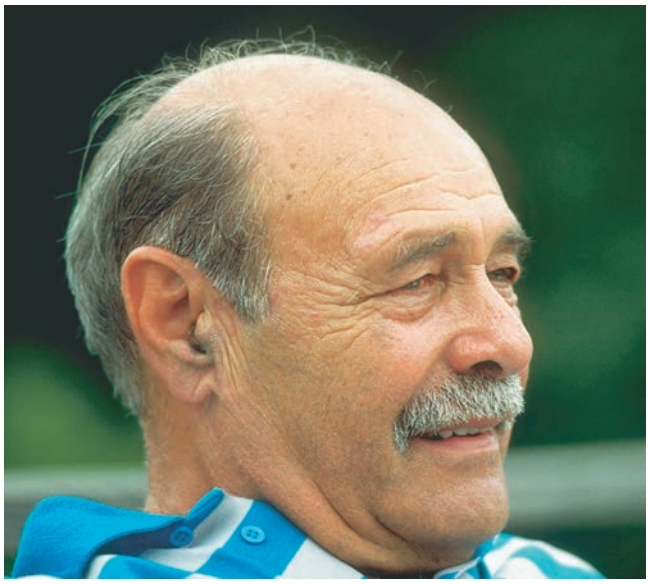


FIGURE 4–14 Pattern baldness, a sex-influenced autosomal trait in humans.

Cases of **sex-influenced inheritance** include pattern baldness in humans, horn formation in certain breeds of sheep (e.g., Dorset Horn sheep), and certain coat-color patterns in cattle. In such cases, autosomal genes are responsible for the contrasting phenotypes displayed by both males and females, but the expression of these genes is dependent on the hormonal constitution of the individual. Thus, the heterozygous genotype exhibits one phenotype in one sex and the contrasting one in the other. For example, **pattern baldness** in humans, where the hair is very thin on the top of the head (**Figure 4–14**), is inherited in this way:

Genotype	Phenotype	
	Females	Males
BB	Bald	Bald
Bb	Not bald	Bald
bb	Not bald	Not bald

Females can display pattern baldness, but this phenotype is much more prevalent in males. When females do inherit the *BB* genotype, the phenotype is less pronounced than in males and is expressed later in life.

ESSENTIAL POINT

Sex-limited and sex-influenced inheritance occurs when the sex of the organism affects the phenotype controlled by a gene located on an autosome. ■

4.13 Genetic Background and the Environment Affect Phenotypic Expression

We now focus on **phenotypic expression.** In previous discussions, we assumed that the genotype of an organism is always directly expressed in its phenotype. For example, pea plants homozygous for the recessive *d* allele (*dd*) will always be dwarf. We discussed gene expression as though the genes operate in a closed system in which the presence or absence of functional products directly determines the collective phenotype of an individual. The situation is actually much more complex. Most gene products function within the internal milieu of the cell, and cells interact with one another in various ways. Furthermore, the organism exists under diverse environmental influences. Thus, gene expression and the resultant phenotype are often modified through the interaction between an individual's particular genotype and the external environment. Here, we deal with several important variables that are known to modify gene expression.

Penetrance and Expressivity

Some mutant genotypes are always expressed as a distinct phenotype, whereas others produce a proportion of individuals whose phenotypes cannot be distinguished from normal (wild type). The degree of expression of a particular trait can be studied quantitatively by determining the *penetrance* and *expressivity* of the genotype under investigation. The percentage of individuals who show at least some degree of expression of a mutant genotype defines the **penetrance** of the mutation. For example, the phenotypic expression of many mutant alleles in *Drosophila* can overlap with wild type. If 15 percent of mutant flies show the wild-type appearance, the mutant gene is said to have a penetrance of 85 percent.

By contrast, **expressivity** reflects the *range of expression* of the mutant genotype. Flies homozygous for the recessive mutant *eyeless* gene yield phenotypes that range from the presence of normal eyes to a partial reduction in size to the complete absence of one or both eyes (**Figure 4–15**). Although the average reduction of eye size is one-fourth to one-half, expressivity ranges from complete loss of both eyes to completely normal eyes.

Examples such as the expression of the *eyeless* gene provide the basis for experiments to determine the causes of phenotypic variation. If, on one hand, a laboratory environment is held constant and extensive phenotypic variation is still observed, other genes may be influencing or modifying the *eyeless* phenotype. On the other hand, if the genetic background is not the cause of the phenotypic variation, environmental factors such as temperature, humidity, and nutrition may be involved. In the case of the *eyeless* phenotype, experiments have shown that both genetic background and environmental factors influence its expression.

Genetic Background: Position Effects

Although it is difficult to assess the specific effect of the **genetic background** and the expression of a gene responsible for determining a potential phenotype, one effect of genetic background has been well characterized, the **position effect.** In such instances, the physical location of a gene in relation to other genetic material may influence its expression. For example, if a region of a chromosome is relocated or rearranged (called a translocation or an inversion event), normal expression of genes in that chromosomal region may be modified. This is particularly true if the gene is relocated to or near certain areas of the chromosome that are prematurely condensed and genetically inert, referred to as **heterochromatin.** An example of a position effect involves female *Drosophila* heterozygous for the X-linked recessive eye color mutant *white* (*w*). The w^+/w genotype normally results in a wild-type brick-red eye color. However, if the region of the X chromosome containing the wild-type w^+ allele is translocated so that it is close to a heterochromatic region, expression of the w^+ allele is modified. Instead of having a red color, the eyes are variegated, or mottled with red and white patches. Apparently, heterochromatic regions inhibit the expression of adjacent genes.

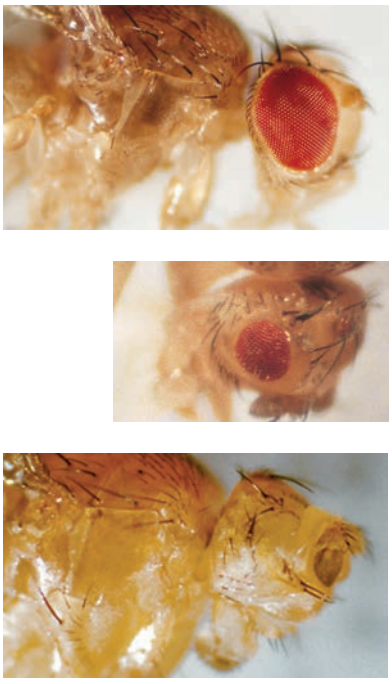


FIGURE 4–15 Variable expressivity, as shown in flies homozygous for the *eyeless* mutation in *Drosophila*. Gradations in phenotype range from wild type to partial reduction to eyeless.

Temperature Effects—An Introduction to Conditional Mutations

Chemical activity depends on the kinetic energy of the reacting substances, which in turn depends on the surrounding temperature. We can thus expect temperature to influence phenotypes. An example is seen in the evening primrose, which produces red flowers when grown at 23°C and white flowers when grown at 18°C. An even more striking example is seen in Siamese cats and Himalayan rabbits, which exhibit dark fur in certain regions where their body temperature is slightly cooler, particularly the nose, ears, and paws (**Figure 4–16**). In these cases, it appears that the enzyme normally responsible for pigment production is functional only at the lower temperatures present in the extremities, but it loses its catalytic function at the slightly higher temperatures found throughout the rest of the body.

Mutations whose expression is affected by temperature, called **temperature-sensitive mutations,** are examples of **conditional mutations,** whereby phenotypic expression is determined by environmental conditions. Examples of temperature-sensitive mutations are known in viruses and a variety of organisms, including bacteria, fungi, and *Drosophila*. In extreme cases, an organism carrying a mutant allele may express a mutant phenotype when grown at one temperature but express the wild-type phenotype when reared at another temperature. This type of temperature effect is useful in studying mutations that interrupt essential processes during development and are thus normally detrimental or lethal. For example, if bacterial viruses are cultured under *permissive conditions* of 25°C, the mutant gene product is functional, infection proceeds normally, and new viruses are produced and can be studied. However, if bacterial viruses carrying temperature-sensitive mutations infect bacteria cultured at 42°C—the *restrictive condition*—infection progresses up to the point where the essential gene product is required (e.g., for viral assembly) and then arrests. Temperature-sensitive mutations are easily induced and isolated in viruses, and have added immensely to the study of viral genetics.

Onset of Genetic Expression

Not all genetic traits become apparent at the same time during an organism's life span. In most cases, the age at which a mutant gene exerts a noticeable phenotype depends on events during the normal sequence of growth and development. In humans, the prenatal, infant, preadult, and adult phases require different genetic information. As a result, many severe inherited disorders are not manifested until after birth. For example, as we saw in Chapter 3, **Tay–Sachs disease,** inherited as an autosomal recessive, is a lethal lipid metabolism disease involving an abnormal enzyme, hexosaminidase A. Newborns appear to be phenotypically normal for the first few months. Then developmental retardation, paralysis, and blindness ensue, and most affected children die around the age of 3.

Lesch–Nyhan syndrome, inherited as an X-linked recessive disease, is characterized by abnormal nucleic acid metabolism (biochemical salvage of nitrogenous purine bases), leading to the accumulation of uric acid in blood and tissues, mental retardation, palsy, and self-mutilation of the lips and fingers. The disorder is due to a mutation in the gene encoding hypoxanthine-guanine phosphoribosyl transferase (HGPRT). Newborns are normal for six to eight months prior to the onset of the first symptoms.

Still another example involves **Duchenne muscular dystrophy (DMD),** an X-linked recessive disorder associated with progressive muscular wasting. It is not usually diagnosed until the child is 3 to 5 years old. Even with modern medical intervention, the disease is often fatal in the early 20s.

(a)

(b)

FIGURE 4–16 (a) A Himalayan rabbit. (b) A Siamese cat. Both species show dark fur color on the snout, ears, and paws. The patches are due to the temperature-sensitive allele responsible for pigment production.

Perhaps the most age-variable of all inherited human disorders is **Huntington disease.** Inherited as an autosomal dominant, Huntington disease affects the frontal lobes of the cerebral cortex, where progressive cell death occurs over a period of more than a decade. Brain deterioration is accompanied by spastic uncontrolled movements, intellectual and emotional deterioration, and ultimately death. Onset of this disease has been reported at all ages, but it most frequently occurs between ages 30 and 50, with a mean onset age of 38 years.

These examples support the concept that the critical expression of genes varies throughout the life cycle of all organisms, including humans. Gene products may play more essential roles at certain life stages, and it is likely that the internal physiological environment of an organism changes with age.

Genetic Anticipation

Interest in studying the genetic onset of phenotypic expression has intensified with the discovery of heritable disorders that exhibit a progressively earlier age of onset and an increased severity of the disorder in each successive generation. This phenomenon is called **genetic anticipation.**

Myotonic dystrophy (DM), the most common type of adult muscular dystrophy, clearly illustrates genetic anticipation. Individuals afflicted with this autosomal dominant disorder exhibit extreme variation in the severity of symptoms. Mildly affected individuals develop cataracts as adults but have little or no muscular weakness. Severely affected individuals demonstrate more extensive myopathy and may be mentally retarded. In its most extreme form, the disease is fatal just after birth. In 1989, C. J. Howeler and colleagues confirmed the correlation of increased severity and earlier onset with successive generations. They studied 61 parent–child pairs, and in 60 cases, age of onset was earlier and more severe in the child than in his or her affected parent.

In 1992, an explanation was put forward for the molecular cause of the mutation responsible for DM, as well as the basis of genetic anticipation. A particular region of the DM gene—a short trinucleotide DNA sequence—is repeated a variable number of times and is unstable. Normal individuals average about five copies of this region; minimally affected individuals have about 50 copies; and severely affected individuals have over 1000 copies. The most remarkable observation was that in successive generations, the size of the repeated segment increases. Although it is not yet clear how this expansion in size affects onset and phenotypic expression, the correlation is extremely strong. Several other inherited human disorders, including the fragile-X syndrome, Kennedy disease, and Huntington disease, also reveal an association between the size of specific regions of the responsible gene and disease severity.

4.14 Genomic (Parental) Imprinting and Gene Silencing

A final example involving genetic background involves what is called **genomic,** or **parental, imprinting,** whereby the process of selective *gene silencing* occurs during early development, impacting subsequent phenotypic expression. Examples involve cases where genes or regions of a chromosome are imprinted on one homolog but not the other. The impact of silencing depends on the parental origin of the genes or regions that are involved. Such silencing leads to the direct phenotypic expression of the allele(s) on the homolog that is not silenced. Thus, the imprinting step, the critical issue in understanding this phenomenon, is thought to occur before or during gamete formation, leading to differentially marked genes (or chromosome regions) in sperm-forming versus egg-forming tissues.

The first example of genomic imprinting was discovered in 1991, in three specific mouse genes. One is the gene encoding insulin-like growth factor II (*Igf2*). A mouse that carries two normal alleles of this gene is normal in size, whereas a mouse that carries two mutant alleles lacks the growth factor and is dwarf. The size of a heterozygous mouse—one allele normal and one mutant—depends on the parental origin of the wild-type allele. The mouse is normal in size if the normal allele comes from the father, but it is dwarf if the normal allele came from the mother. From this, we can deduce that the normal *Igf2* gene is imprinted and thus silenced during egg production, but it functions normally when it has passed through sperm-producing tissue in males. The imprint is inherited in the sense that the *Igf2* gene in all progeny cells formed during development remain silenced. Imprinting in the next generation then depends on whether the gene passes through sperm-producing or egg-forming tissue.

An example in humans involves two distinct genetic disorders thought to be caused by differential imprinting of the same region of the long arm of chromosome 15 (15q1). In both cases, the disorders are due to an identical deletion of this region in one member of the chromosome 15 pair. The first disorder, **Prader–Willi syndrome (PWS),** results when the paternal segment is deleted and an undeleted maternal chromosome remains. If the maternal segment is deleted and an undeleted paternal chromosome remains, an entirely different disorder, **Angelman syndrome (AS),** results.

These two conditions exhibit different phenotypes. PWS entails mental retardation, a severe eating disorder marked by an uncontrollable appetite, obesity, diabetes, and growth retardation. Angelman syndrome also involves mental retardation, but involuntary muscle contractions (chorea) and seizures characterize the disorder. We can

conclude that the involved region of chromosome 15 is imprinted differently in male and female gametes and that both an undeleted maternal and a paternal region are required for normal development.

Although numerous questions remain unanswered regarding genomic imprinting, it is now clear that many genes are subject to this process. More than 50 have been identified in mammals thus far. It appears that regions of chromosomes rather than specific genes are imprinted. This phenomenon is an example of the more general topic of **epigenetics,** where genetic expression is *not* the direct result of the information stored in the nucleotide sequence of DNA. Instead, the DNA is altered in a way that affects its expression. These changes are stable in the sense that they are transmitted during cell division to progeny cells, and often through gametes to future generations.

The precise molecular mechanism of imprinting and other epigenetic events is still a matter for conjecture, but it seems certain that **DNA methylation** is involved. In most eukaryotes, methyl groups can be added to the carbon atom at position 5 in cytosine (see Chapter 9) as a result of the activity of the enzyme DNA methyltransferase. Methyl groups are added when the dinucleotide CpG or groups of CpG units (called CpG islands) are present along a DNA chain.

DNA methylation is a reasonable mechanism for establishing a molecular imprint, since there is evidence that a high level of methylation can inhibit gene activity and that active genes (or their regulatory sequences) are often undermethylated. We will encounter other examples throughout the text, and return to more comprehensive coverage of epigenetics in Special Topics in Modern Genetics later in this book.

ESSENTIAL POINT

Phenotypic expression is not always the direct reflection of the genotype. A percentage of organisms may not express the expected phenotype at all, the basis of the penetrance of a mutant gene. In addition, the phenotype can be modified by genetic background, temperature, and nutrition. The onset of expression of a gene may vary during the lifetime of an organism, and it may even be imprinted so that it is expressed differently depending on parental origin. ■

4.15 Extranuclear Inheritance Modifies Mendelian Patterns

Throughout the history of genetics, occasional reports have challenged the basic tenet of Mendelian transmission genetics—that the phenotype is determined solely by nuclear genes located on the chromosomes of both parents. In this final section of the chapter, we consider several examples of inheritance patterns that vary from those predicted by the traditional biparental inheritance of nuclear genes, phenomena that are designated as **extranuclear inheritance.** In the following cases, we will focus on two broad categories. In the first, an organism's phenotype is affected by the expression of genes contained in the DNA of mitochondria or chloroplasts rather than the nucleus, generally referred to as organelle heredity. In the second category, referred to as a maternal effect, an organism's phenotype is determined by genetic information expressed in the gamete of the mother—such that, following fertilization, the developing zygote's phenotype is influenced not by the individual's genotype, but by gene products directed by the genotype of the mother.

Initially, such observations met with skepticism. However, with increasing knowledge of molecular genetics and the discovery of DNA in mitochondria and chloroplasts, the phenomenon of extranuclear inheritance came to be recognized as an important aspect of genetics.

Organelle Heredity: DNA in Chloroplasts and Mitochondria

We begin by examining examples of inheritance patterns related to chloroplast and mitochondrial function. Before DNA was discovered in these organelles, the exact mechanism of transmission of the traits was not clear, except that their inheritance appeared to be linked to something in the cytoplasm rather than to genes in the nucleus. Furthermore, transmission was most often from the maternal parent through the ooplasm, causing the results of reciprocal crosses to vary. Such an extranuclear pattern of inheritance is now appropriately called **organelle heredity.**

Analysis of the inheritance patterns resulting from mutant alleles in chloroplasts and mitochondria has been difficult for two reasons. First, the function of these organelles is dependent on gene products from both nuclear and organelle DNA, making the discovery of the genetic origin of mutations affecting organelle function difficult. Second, many mitochondria and chloroplasts are contributed to each progeny. Thus, if only one or a few of the organelles contain a mutant gene in a cell among a population of mostly normal mitochondria, the corresponding mutant phenotype may not be revealed. This condition, referred to as **heteroplasmy,** may lead to normal cells since the organelles lacking the mutation provide the basis of wild-type function. Analysis is therefore much more complex than for Mendelian characters.

Chloroplasts: Variegation in Four-o'clock Plants

In 1908, Karl Correns (one of the rediscoverers of Mendel's work) provided the earliest example of inheritance linked to chloroplast transmission. Correns discovered a variant of the four-o'clock plant, *Mirabilis jalapa*, that had branches with either white, green, or variegated white-and-green leaves. The white areas in variegated leaves and in the completely white leaves lack chlorophyll that provides

Source of Pollen	Location of Ovule		
	White branch	Green branch	Variegated branch
White branch	White	Green	White, green, or variegated
Green branch	White	Green	White, green, or variegated
Variegated branch	White	Green	White, green, or variegated

FIGURE 4–17 Offspring from crosses between flowers from various branches of four-o'clock plants. The photograph illustrates variegation in leaves of the madagascar spur.

the green color to normal leaves. Chlorophyll is the light-absorbing pigment made within chloroplasts.

Correns was curious about how inheritance of this phenotypic trait occurred. As shown in **Figure 4–17**, inheritance in all possible combinations of crosses is strictly determined by the phenotype of the ovule source. For example, if the seeds (representing the progeny) were derived from ovules on branches with green leaves, all progeny plants bore only green leaves, regardless of the phenotype of the source of pollen. Correns concluded that inheritance was transmitted through the cytoplasm of the maternal parent because the pollen, which contributes little or no cytoplasm to the zygote, had no apparent influence on the progeny phenotypes.

Since leaf coloration is related to the chloroplast, genetic information contained either in that organelle or somehow present in the cytoplasm and influencing the chloroplast must be responsible for the inheritance pattern. It now seems certain that the genetic "defect" that eliminates the green chlorophyll in the white patches on leaves is a mutation in the DNA housed in the chloroplast.

Mitochondrial Mutations: *poky* in *Neurospora* and *petite* in *Saccharomyces*

Mutations affecting mitochondrial function have been discovered and studied, revealing that they too contain a distinctive genetic system. As with chloroplasts, mitochondrial mutations are transmitted through the cytoplasm. In our current discussion, we will emphasize the link between mitochondrial mutations and the resultant extranuclear inheritance patterns.

In 1952, Mary B. Mitchell and Hershel K. Mitchell studied the bread mold *Neurospora crassa*. They discovered a slow-growing mutant strain and named it *poky*. Slow growth is associated with impaired mitochondrial function, specifically in relation to certain cytochromes essential for electron transport. Results of genetic crosses between wild-type and *poky* strains suggest that *poky* is an extranuclear trait inherited through the cytoplasm. If one mating type is *poky* and the other is wild type, all progeny colonies are *poky*. The reciprocal cross, where *poky* is transmitted by the other mating type, produces normal wild-type colonies.

Another extensive study of mitochondrial mutations has been performed with the yeast *Saccharomyces cerevisiae*. The first such mutation, described by Boris Ephrussi and his coworkers in 1956, was named *petite* because of the small size of the yeast colonies (**Figure 4–18**). Many independent *petite* mutations have since been discovered and studied, and all have a common characteristic—a deficiency in cellular respiration involving abnormal electron transport. The majority of them demonstrate cytoplasmic transmission, indicating mutations in the DNA of the mitochondria. This organism is a facultative anaerobe and can grow by fermenting glucose through glycolysis; thus, it may survive the loss of mitochondrial function by generating energy anaerobically.

The complex genetics of *petite* mutations has revealed that a small proportion are the result of nuclear DNA changes. They exhibit Mendelian inheritance and illustrate that mitochondria function depends on both nuclear and organellar gene products.

Mitochondrial Mutations: Human Genetic Disorders

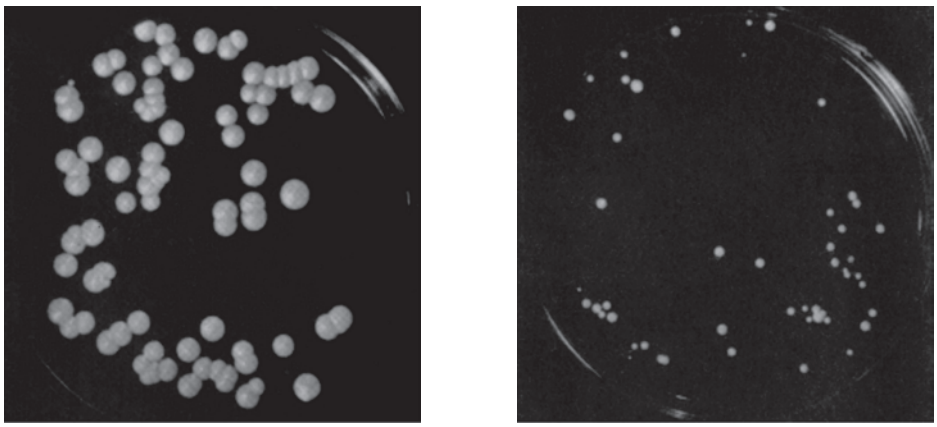
Our knowledge of the genetics of mitochondria has now greatly expanded. The DNA found in human mitochondria has been completely sequenced and contains 16,569 base pairs. Mitochondrial gene products have been identified and include the following:

13 proteins, required for aerobic cellular respiration

22 transfer RNAs (tRNAs), required for translation

2 ribosomal RNAs (rRNAs), required for translation

Because a cell's energy supply is largely dependent on aerobic cellular respiration, disruption of any



Normal colonies Petite colonies

FIGURE 4–18 Photos comparing normal versus petite colonies of the yeast *Saccharomyces cerevisiae*.

mitochondrial gene by mutation may potentially have a severe impact on that organism, such as we saw in our previous discussion of the *petite* mutation in yeast. In fact, mtDNA is particularly vulnerable to mutations for two possible reasons. First, the ability to repair mtDNA damage does not appear to be equivalent to that of nuclear DNA. Second, the concentration of highly mutagenic free radicals generated by cell respiration that accumulate in such a confined space very likely raises the mutation rate in mtDNA.

Fortunately, a zygote receives a large number of organelles through the egg, so if only one organelle or a few of them contain a mutation (an illustration of *heteroplasmy*), the impact is greatly diluted by the many mitochondria that lack the mutation and function normally. If a deleterious mutation arises or is present in the initial population of organelles, adults will have cells with a variable mixture of both normal and abnormal organelles. From a genetic standpoint, this condition of heteroplasmy makes analysis quite difficult.

Many disorders in humans are known to be due to mutations in mitochondrial genes. For example, **myoclonic epilepsy and ragged-red fiber disease (MERRF)** demonstrates a pattern of inheritance consistent with maternal transmission. Only the offspring of affected mothers inherit this disorder, while the offspring of affected fathers are normal. Individuals with this rare disorder express ataxia (lack of muscular coordination), deafness, dementia, and epileptic seizures. The disease is named for the presence of "ragged-red" skeletal-muscle fibers that exhibit blotchy red patches resulting from the proliferation of aberrant mitochondria. Brain function, which has a high energy demand, is also affected in this disorder, leading to the neurological symptoms described above.

The mutation that causes MERRF has now been identified and is in a mitochondrial gene whose altered product interferes with the capacity for translation of proteins within the organelle. This, in turn, leads to the various manifestations of the disorder. The cells of MERRF individuals exhibit heteroplasmy, containing a mixture of normal and abnormal mitochondria. Different patients display different proportions of the two, and even different cells from the same patient exhibit various levels of abnormal mitochondria. Were it not for heteroplasmy, the mutation would very likely be lethal, testifying to the essential nature of mitochondrial function and its reliance on the genes encoded by DNA within the organelle.

Mitochondria, Human Health, and Aging

The study of hereditary mitochondrial-based disorders provides insights into the critical importance of this organelle during normal development. In fact, mitochondrial dysfunction seems to be implicated in a large number of major human disease conditions, including anemia, blindness, Type 2 (late-onset) diabetes, autism, atherosclerosis, infertility, neurodegenerative diseases such as Parkinson, Alzheimer, and Huntington disease, schizophrenia and bipolar disorders, and a variety of cancers. It is becoming evident, for example, that mutations in mtDNA are present in such human malignancies as skin, colorectal, liver, breast, pancreatic, lung, prostate, and bladder cancers.

Over 400 mtDNA mutations associated with more than 150 distinct mtDNA-based genetic syndromes have been identified. Genetic tests for detecting mutations in the mtDNA genome that may serve as early-stage disease markers have been developed. However, it is still unclear whether mtDNA mutations are causative effects contributing to development of malignant tumors or whether they are the consequences of tumor formation. Nonetheless, there is an interesting link between mtDNA mutations and cancer, including data suggesting that many chemical carcinogens have significant mutation effects on mtDNA.

The study of hereditary mitochondrial-based disorders has also suggested a link between the progressive decline of mitochondrial function and the aging process. It has been hypothesized that the accumulation of sporadic mutations in mtDNA leads to an increased prevalence of defective

mitochondria (and the concomitant decrease in the supply of ATP) in cells over a lifetime. This condition in turn plays a significant role in aging.

These, and other studies, continue to speak to the importance of normal mitochondrial function. As cells undergo genetic damage, which appears to be a natural phenomenon, their function declines, which may be an underlying factor in aging as well as in the progression of age-related disorders.

Maternal Effect

In **maternal effect,** also referred to as maternal influence, an offspring's phenotype for a particular trait is under the control of the mother's *nuclear gene products* present in the egg. This is in contrast to biparental inheritance, where both parents transmit information on genes in the nucleus that determines the offspring's phenotype. In cases of maternal effect, the nuclear genes of the female gamete are transcribed, and the genetic products (either proteins or yet untranslated mRNAs) accumulate in the egg ooplasm. After fertilization, these products are distributed among newly formed cells and influence the patterns or traits established during early development. The following example will illustrate such an influence of the maternal genome on particular traits.

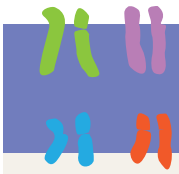
Embryonic Development in *Drosophila*

A recently documented example that illustrates maternal effect involves various genes that control embryonic development in *Drosophila melanogaster*. The genetic control of embryonic development in *Drosophila*, discussed in greater detail in Chapter 20, is a fascinating story. The protein products of the maternal-effect genes function to activate other genes, which may in turn activate still other genes. This cascade of gene activity leads to a normal embryo whose subsequent development yields a normal adult fly. The extensive work by Edward B. Lewis, Christiane Nüsslein-Volhard, and Eric Wieschaus (who shared the 1995 Nobel Prize for Physiology or Medicine for their findings) has clarified how these and other genes function. Genes that illustrate maternal effect have products that are synthesized by the developing egg and stored in the oocyte prior to fertilization. Following fertilization, these products specify molecular gradients that determine spatial organization as development proceeds.

For example, the gene *bicoid* (*bcd*) plays an important role in specifying the development of the anterior portion of the fly. Embryos derived from mothers who are homozygous for this mutation (bcd^-/bcd^-) fail to develop anterior areas that normally give rise to the head and thorax of the adult fly. Embryos whose mothers contain at least one wild-type allele (bcd^+) develop normally, even if the genotype of the embryo is homozygous for the mutation. Consistent with the concept of maternal effect, the *genotype of the female parent*, not the *genotype of the embryo*, determines the phenotype of the offspring. Nusslein-Volhard and Wieschaus, using large-scale mutant screens, discovered many other maternal-effect genes critical to early development in *Drosophila*.

ESSENTIAL POINT

When patterns of inheritance vary from that expected due to biparental transmission of nuclear genes, phenotypes are often found to be under the control of DNA present in mitochondria or chloroplasts, or are influenced during development by the expression of the maternal genotype in the egg. ■



GENETICS, TECHNOLOGY, AND SOCIETY

Improving the Genetic Fate of Purebred Dogs

For dog lovers, nothing is quite so heartbreaking as watching a dog slowly go blind, struggling to adapt to a life of perpetual darkness. That's what happens in progressive retinal atrophy (PRA), a group of inherited disorders first described in Gordon setters in 1909. Since then, PRA has been detected in more than 100 other breeds of dogs, including Irish setters, border collies, Norwegian elkhounds, toy poodles, miniature schnauzers, cocker spaniels, and Siberian huskies.

The products of many genes are required for the development and maintenance of healthy retinas, and a defect in any one of these genes may cause retinal dysfunction. Decades of research have led to the identification of five such genes (*PDE6A, PDE6B, PRCD, rhodopsin,* and *PRGR*), and more may be discovered. Different mutant alleles are present in different breeds, and each allele is associated with a different form of PRA that varies slightly in its clinical symptoms and rate of progression. Mutations of *PDE6A, PDE6B,* and *PRCD* genes are inherited in a recessive pattern, mutations of the *rhodopsin* gene (such as those found in Mastiffs) are dominant, and

(continued)

PRGR mutations (in Siberian huskies and Samoyeds) are X-linked.

PRA is almost ten times more common in certain purebred dogs than in mixed breeds. The development of distinct breeds of dogs has involved intensive selection for desirable attributes, such as a particular size, shape, color, or behavior. Many desired characteristics are determined by recessive alleles. The fastest way to increase the homozygosity of these alleles is to mate close relatives, which are likely to carry the same alleles. For example, dogs may be mated to a cousin or a grandparent. Some breeders, in an attempt to profit from impressive pedigrees, also produce hundreds of offspring from individual dogs that have won major prizes at dog shows. This "popular sire effect," as it has been termed, further increases the homozygosity of alleles in purebred dogs.

Unfortunately, the generations of inbreeding that have established favorable characteristics in purebreeds have also increased the homozygosity of certain harmful recessive alleles, resulting in a high incidence of inherited diseases. More than 300 genetic diseases have been characterized in purebred dogs, and many breeds have a predisposition to more than 20 of them. According to researchers at Cornell University, purebred dogs suffer the highest incidence of inherited disease of any animal: 25 percent of the 20 million purebred dogs in America are affected with one genetic ailment or another.

Fortunately, advances in canine genetics are beginning to provide new tools to increase the health of purebred dogs. As of 2007, genetic tests are available to detect 30 different retinal diseases in dogs. Tests for PRA are now being used to identify heterozygous carriers of *PRCD* mutations. These carriers show no symptoms of PRA but, if mated with other carriers, pass the trait on to about 25 percent of their offspring. Eliminating PRA carriers from breeding programs has almost eradicated this condition from Portuguese water dogs and has greatly reduced its prevalence in other breeds.

Scientists will be able to identify more genes underlying canine inherited diseases thanks to the completion of the Dog Genome Project in 2005. In addition, new therapies that correct gene-based defects will emerge.

The Dog Genome Project may have benefits for humans beyond the reduction of disease in their canine companions. Eighty-five percent of the genes in the dog genome have equivalents in humans, and over 300 diseases affecting dogs also affect humans, including heart disease, epilepsy, allergies, and some cancers. The identification of a disease-causing gene in dogs can be a shortcut to the isolation of the corresponding gene in humans. By contributing to the cure of human diseases, dogs may prove to be "man's best friend" in an entirely new way.

Your Turn

Take time, individually or in groups, to answer the following questions. Investigate the references and links, to help you understand some of the issues surrounding the genetics of purebred dogs.

1. What are some of the limitations of genetic tests, especially as they apply to purebred dog genetic diseases?

This topic is discussed on the OptiGen website (http://www.optigen.com). *OptiGen is a company that offers gene tests for all known forms of PRA in dogs and is developing tests for other inherited disorders. From their TESTS list, select prcd-PRA, and visit the link "Benefits and Limits to All Genetic Testing."*

2. Which human disease is similar to PRA in the Siberian husky?

To learn more about these genes and diseases, visit the "Inherited Diseases in Dogs" database (http://www.vet.cam.ac.uk/idid/) *and search the database for the Siberian husky and Progressive Retinal Atrophy. Once there, follow the OMIM reference link to learn about the human version of PRA in the Siberian husky.*

3. Chinese Shar-Pei dogs have been selected for their unique wrinkled skin. Unfortunately, these purebred dogs are also predisposed to hereditary fevers. Recently, a gene called *HAS2* was identified in these dogs. It encodes the enzyme hyaluronic acid synthase 2, and a mutation associated with the gene is responsible for both skin appearance and fever. Describe how the *HAS2* gene can have these pleiotropic effects and why this finding is important to human health.

Read about the HAS2 gene discovery in Olsson, M. et al. March 11, 2011. A novel unstable duplication upstream of *HAS2* predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs. *PLoS Genetics* 7(3):e1001332.

CASE STUDY | Sudden blindness

A man aged 21 suddenly started having blurred vision in both his eyes. Over the course of six weeks, his sight got worse until most of his central vision was gone. He was diagnosed with Leber hereditary optic neuropathy (LHON), which is caused by a point mutation in the mitochondrial DNA (mtDNA) in the subunits of complex I of the oxidative phosphorylation chain. It leads to the sudden death of cells in the optic nerve. This disease can be inherited, and the symptoms might not show until the disease is triggered by oxidative stress. This case raises several questions:

1. Did the young man inherit the mitochondrial mutation from his father or his mother?
2. Will he pass it on to his children? Explain why or why not.
3. Which other members of his family should be tested for the mutation?

INSIGHTS AND SOLUTIONS

Genetic problems take on added complexity if they involve two independent characters and multiple alleles, incomplete dominance, or epistasis. The most difficult types of problems are those that pioneering geneticists faced during laboratory or field studies. They had to determine the mode of inheritance by working backward from the observations of offspring to parents of unknown genotype.

1. Consider the problem of comb-shape inheritance in chickens, where walnut, pea, rose, and single are the observed distinct phenotypes (see the photographs below). How is comb shape inherited, and what are the genotypes of the P_1 generation of each cross? Use the following data:

Cross 1 :	single	×	single	⟶	all single
Cross 2 :	walnut	×	walnut	⟶	all walnut
Cross 3 :	rose	×	pea	⟶	all walnut
Cross 4 :	$F_1 \times F_1$ of cross 3				
	walnut	×	walnut	⟶	93 walnut
					28 rose
					32 pea
					10 single

Walnut

Pea

Rose

Single

Solution: At first glance, this problem appears quite difficult. However, applying a systematic approach and breaking the analysis into steps usually simplifies it. Our approach involves two steps. First, analyze the data carefully for any useful information. Then, once you identify something that is clearly helpful, follow an empirical approach—that is, formulate a hypothesis and, in a sense, test it against the given data. Look for a pattern of inheritance that is consistent with all cases.

This problem gives two immediately useful facts. First, in cross 1, P_1 singles breed true. Second, while P_1 walnuts breed true in cross 2, a walnut phenotype is also produced in cross 3 between rose and pea. When these F_1 walnuts are crossed in cross 4, all four comb shapes are produced in a ratio that approximates 9:3:3:1. This observation immediately suggests a cross involving two gene pairs, because the resulting data display the same ratio as in Mendel's dihybrid crosses. Since only one trait is involved (comb shape), epistasis may be occurring. This could serve as your working hypothesis, and you must now propose how the two gene pairs "interact" to produce each phenotype.

If you call the allele pairs *A*, *a* and *B*, *b*, you can predict that because walnut represents 9/16 in cross 4, *A*−*B*− will produce walnut. You might also hypothesize that in cross 2, the genotypes are *AABB* × *AABB*, where walnut bred true. (Recall that *A*− and *B*− mean *AA* or *Aa* and *BB* or *Bb*, respectively.)

The phenotype representing 1/16 of the offspring of cross 4 is single; therefore, you could predict that this phenotype is the result of the *aabb* genotype. This is consistent with cross 1.

Now you have only to determine the genotypes for rose and pea. The most logical prediction is that at least one dominant *A* or *B* allele combined with the double recessive condition of the other allele pair accounts for these phenotypes. For example,

A−*bb*	⟶	rose
aaB−	⟶	pea

If *AAbb* (rose) is crossed with *aaBB* (pea) in cross 3, all offspring will be *AaBb* (walnut). This is consistent with the data, and you must now look at only cross 4. We predict these walnut genotypes to be *AaBb* (as above), and from the cross *AaBb* (walnut) × *AaBb* (walnut) we expect

9/16	*A*−*B*−	(walnut)
3/16	*A*−*bb*	(rose)
3/16	*aaB*−	(pea)
1/16	*aabb*	(single)

Our prediction is consistent with the information given. The initial hypothesis of the epistatic interaction of two gene pairs proves consistent throughout, and the problem is solved.

This problem demonstrates the need for a basic theoretical knowledge of transmission genetics. Then, you can search for appropriate clues that will enable you to proceed in a stepwise fashion toward a solution. Mastering problem solving requires practice but can give you a great deal of satisfaction. Apply this general approach to the following problems.

2. In radishes, flower color may be red, purple, or white. The edible portion of the radish may be long or oval. When only flower color is studied, no dominance is evident, and red × white crosses yield all purple. If these F_1 purples are interbred, the F_2 generation consists of 1/4 red: 1/2 purple: 1/4 white. Regarding radish shape, long is dominant to oval in a normal Mendelian fashion.

   (a) Determine the F_1 and F_2 phenotypes from a cross between a true-breeding red, long radish and one that is white and oval. Be sure to define all gene symbols initially.

Solution: This is a modified dihybrid cross in which the gene pair controlling color exhibits incomplete dominance. Shape is controlled conventionally. First, establish gene symbols:

RR = red $O-$ = long
Rr = purple oo = oval
rr = white

P_1: *RROO* × *rroo*
(red long) (white oval)

F_1: all *RrOo* (purple long)

$F_1 \times F_1$: *RrOo* × *RrOo*

F_2: 1/4 *RR*	3/4 *O*–	3/16 *RRO*–	red long
	1/4 *oo*	1/16 *RRoo*	red oval
2/4 *Rr*	3/4 *O*–	6/16 *RrO*–	purple long
	1/4 *oo*	2/16 *Rroo*	purple oval
1/4 *rr*	3/4 *O*–	3/16 *rrO*–	white long
	1/4 *oo*	1/16 *rroo*	white oval

Note that to generate the F_2 results, we have used the forked-line method. First, we consider the outcome of crossing F_1 parents for the color genes (*Rr* × *Rr*). Then the outcome of shape is considered (*Oo* × *Oo*).

3. In humans, red–green color blindness is inherited as an X-linked recessive trait. A woman with normal vision whose father is color blind marries a male who has normal vision. Predict the color vision of their male and female offspring.

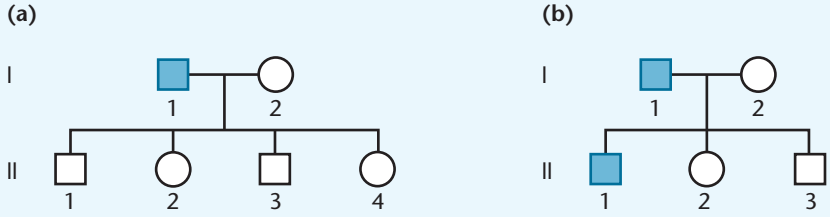
Solution: The female is heterozygous since she inherited an X chromosome with the mutant allele from her father. Her husband is normal. Therefore, the parental genotypes are

Cc × C↑ (↑ represents the Y chromosome)

All female offspring are normal (*CC* or *Cc*). One-half of the male children will be color blind (*c* ↑), and the other half will have normal vision (*C*↑).

4. Consider the two very limited unrelated pedigrees shown below. Of the four combinations of X-linked recessive, X-linked dominant, autosomal recessive, and autosomal dominant, which modes of inheritance can be absolutely ruled out in each case?

Solution: For both pedigrees, X-linked recessive and autosomal recessive remain possible, provided that the maternal parent is heterozygous in pedigree (b). At first glance autosomal dominance seems unlikely in pedigree (a), since at least half of the offspring should express a dominant trait expressed by one of their parents. However, while it is true that if the affected parent carries an autosomal dominant gene heterozygously, each offspring has a 50 percent chance of inheriting and expressing the mutant gene, the sample size of four offspring is too small to rule out this possibility. In pedigree (b), autosomal dominance is clearly possible. In both cases, one can rule out X-linked dominance because the female offspring would inherit and express the dominant allele, and they do not express the trait in either pedigree.



Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we focused on many extensions and modifications of Mendelian principles and ratios. In the process, we encountered many opportunities to consider how this information was acquired. Answer the following fundamental questions:

   (a) How were early geneticists able to ascertain inheritance patterns that did not fit typical Mendelian ratios?
   (b) How did geneticists determine that inheritance of some phenotypic characteristics involves the interactions of two or more gene pairs? How were they able to determine how many gene pairs were involved?
   (c) How do we know that specific genes are located on the sex-determining chromosomes rather than on autosomes?
   (d) For genes whose expression seems to be tied to the sex of individuals, how do we know whether a gene is X-linked in contrast to exhibiting sex-limited or sex-influenced inheritance?
   (e) How was extranuclear inheritance discovered?

CONCEPT QUESTION

2. Review the Chapter Concepts list on page 69. These all relate to exceptions to the inheritance patterns encountered by Mendel. Write a short essay that explains why multiple and lethal alleles often result in a modification of the classic Mendelian monohybrid and dihybrid ratios. ■

3. In mice, there is a set of multiple alleles of a gene for coat color. Four of those alleles are as follows:

   C = full color (wild)
   cch = chinchilla
   cd = dilution
   c = albino

   Given that the gene locus is not sex-linked and that each allele is dominant to those lower in the list, diagram the crosses indicated below and give the phenotypic ratios expected from each.
   (a) wild (heterozygous for dilution) × chinchilla (heterozygous for albino)

(b) chinchilla (heterozygous for albino) × albino

4. The trait of medium-sized leaves in iris is determined by the genetic condition *PP′*. Plants with large leaves are *PP*, whereas plants with small leaves are *P′P′*. A cross is made between two plants each with medium-sized leaves. If they produce 80 seedlings, what would be the expected phenotypes, and in what numbers would they be expected? What is the term for this allelic relationship?
5. The creeper gene in chickens causes short and stunted legs (creeper condition) in the heterozygous state (*Cc*) and lethality in the homozygous state (*CC*). The genotype *cc* produces normal chickens. What ratio is obtained when creeper chickens are interbred? Is the *C* allele behaving dominantly or recessively in causing lethality?
6. Three gene pairs located on separate autosomes determine flower color and shape as well as plant height. The first pair exhibits incomplete dominance, where color can be red, pink (the heterozygote), or white. The second pair leads to the dominant personate or recessive peloric flower shape, while the third gene pair produces either the dominant tall trait or the recessive dwarf trait. Homozygous plants that are red, personate, and tall are crossed with those that are white, peloric, and dwarf. Determine the F_1 genotype(s) and phenotype(s). If the F_1 plants are interbred, what proportion of the offspring will exhibit the same phenotype as the F_1 plants?
7. As in the plants of Problem 6, color may be red, white, or pink; and flower shape may be personate or peloric. Determine the P_1 and F_1 genotypes for the following crosses:

(a) red, peloric × white, personate → F_1: all pink, personate

(b) red, personate × white, peloric → F_1: all pink, personate

(c) pink, personate × red, peloric → F_1:
- 1/4 red, personate
- 1/4 red, peloric
- 1/4 pink, personate
- 1/4 pink, peloric

(d) pink, personate × white, peloric → F_1:
- 1/4 white, personate
- 1/4 white, peloric
- 1/4 pink, personate
- 1/4 pink, peloric

(e) What phenotypic ratios woud result from crossing the F_1 of (a) to the F_1 of (b)?

8. Two different genes, located on two different chromosomes, are responsible for color production in the aleurone layer of corn kernels. For color production (either purple or red), the dominant alleles of these two genes (*C* and *R*) must come together. Furthermore, a third gene, located on a third chromosome, interacts with the *C* and *R* alleles to determine whether the aleurone will be red or purple. While the dominant allele (*P*) ensures purple color, the homozygous recessive condition (*pp*) makes the aleurone red. Determine the F_1 phenotypic ratio of the following crosses: (a) *CCrrPP* × *ccRRpp*; (b) *CcRRpp* × *CCRrpp*; (c) *CcRrPp* × *CcRrpp*.
9. Given the inheritance pattern of aleurone color in corn kernels described in Problem 8, predict the genotype and phenotype of the parents that produced the following F_1 offspring: (a) 27/64 purple: 9/64 red: 28/64 colorless; (b) 9/32 purple: 0/32 red: 3/32 colorless.
10. A husband and wife have normal vision, although both of their fathers are red–green color-blind, inherited as an X-linked recessive condition. What is the probability that their first child will be (a) a normal son, (b) a normal daughter, (c) a color-blind son, (d) a color-blind daughter?
11. Morbid obesity in humans is an autosomal recessive disorder. Rett syndrome is a neurological disorder that is marked by the X-linked dominant allele. If two non-obese parents with Rett syndrome produce an obese, non-Rett son, what is the probability that their next child will be a female who is obese and also has Rett syndrome?
12. In goats, development of the beard is due to a recessive gene. The following cross involving true-breeding goats was made and carried to the F_2 generation:

P_1: bearded female × beardless male
↓
F_1: all bearded males and beardless females

$F_1 \times F_1 \rightarrow$
- 1/8 beardless males
- 3/8 bearded males
- 3/8 beardless females
- 1/8 bearded females

Offer an explanation for the inheritance and expression of this trait, diagramming the cross. Propose one or more crosses to test your hypothesis.
13. Duchenne muscular dystrophy (DMD), marked by muscular degeneration, results from an X- linked recessive gene. Thus, a female who is heterozygous for this gene and does not have the disease can be a carrier. What kind of offspring can you expect from a DMD-affected male and a carrier female? Can there be a carrier male?
14. Hemophilia is an X-linked recessive mutation in humans that causes delayed blood clotting. What kinds of F_1 and F_2 offspring would be expected from matings between (a) a hemophilic female and a normal male, and (b) a hemophilic male and a normal female? Compare these results to those that would be obtained if the hemophilic gene was autosomal.
15. Premature graying of hair in humans results from an autosomal recessive mutation. Thus, a homozygous recessive individual will experience premature graying. What phenotypic F_1 male and female ratios will result from a mating between a hemophilic, prematurely grayed female and a male who is normal for both phenotypes, but had a prematurely grayed mother?
16. While *vermilion* is X-linked in *Drosophila* and causes eye color to be bright red, *brown* is an autosomal recessive mutation that causes the eye to be brown. Flies carrying both mutations lose all pigmentation and are white-eyed. Predict the F_1 and F_2 results of the following crosses:

(a) vermilion females × brown males
(b) brown females × vermilion males
(c) white females × wild males

17. In pigs, coat color may be sandy, red, or white. A geneticist spent several years mating true-breeding pigs of all different color combinations, even obtaining true-breeding lines from different parts of the country. For crosses 1 and 4 in the following table, she encountered a major problem: her computer crashed and she lost the F_2 data. She nevertheless persevered and, using the limited data shown here, was able to predict the mode of inheritance and the number of genes involved, as well as to assign genotypes to each coat color. Attempt to duplicate her analysis, based on the available data generated from the crosses shown.

Cross	P_1	F_1	F_2
1	sandy × sandy	All red	Data lost
2	red × sandy	All red	3/4 red: 1/4 sandy
3	sandy × white	All sandy	3/4 sandy: 1/4 white
4	white × red	All red	Data lost

When you have formulated a hypothesis to explain the mode of inheritance and assigned genotypes to the respective coat colors, predict the outcomes of the F_2 generations where the data were lost.

18. A geneticist from an alien planet that prohibits genetic research brought with him two true-breeding lines of frogs. One frog line croaks by *uttering* "rib-it rib-it" and has purple eyes. The other frog line croaks by *muttering* "knee-deep knee-deep" and has green eyes. He mated the two frog lines, producing F_1 frogs that were all utterers with blue eyes. A large F_2 generation then yielded the following ratios:

27/64 blue, utterer
12/64 green, utterer
9/64 blue, mutterer
9/64 purple, utterer
4/64 green, mutterer
3/64 purple, mutterer

(a) How many total gene pairs are involved in the inheritance of both eye color and croaking?
(b) Of these, how many control eye color, and how many control croaking?
(c) Assign gene symbols for all phenotypes, and indicate the genotypes of the P_1, F_1, and F_2 frogs.
(d) After many years, the frog geneticist isolated true-breeding lines of all six F_2 phenotypes. Indicate the F_1 and F_2 phenotypic ratios of a cross between a blue, mutterer and a purple, utterer.

19. In another cross, the frog geneticist from Problem 18 mated two purple, utterers with the results shown here. What were the genotypes of the parents?

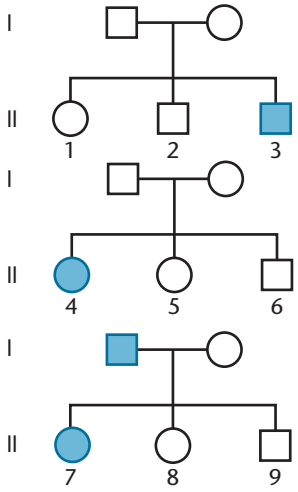
9/16 purple, utterers
3/16 purple, mutterers
3/16 green, utterers
1/16 green, mutterers

20. In cattle, coats may be solid white, solid black, or black-and-white spotted. When true-breeding solid whites are mated with true-breeding solid blacks, the F_1 generation consists of all solid white individuals. After many $F_1 \times F_1$ matings, the following ratio was observed in the F_2 generation:

12/16 solid white
3/16 black-and-white spotted
1/16 solid black

Explain the mode of inheritance governing coat color by determining how many gene pairs are involved and which genotypes yield which phenotypes. Is it possible to isolate a true-breeding strain of black-and-white spotted cattle? If so, what genotype would they have? If not, explain why not.

21. Consider the following three pedigrees, all involving the same human trait:



(a) Which sets of conditions, if any, can be excluded?

dominant and X-linked
dominant and autosomal
recessive and X-linked
recessive and autosomal

(b) For any set of conditions that you excluded, indicate the *single individual* in generation II (1–9) that was instrumental in your decision to exclude that condition. If none were excluded, answer "none apply."
(c) Given your conclusions in parts (a) and (b), indicate the *genotype* of individuals II-1, II-6, and II-9. If more than one possibility applies, list all possibilities. Use the symbols *A* and *a* for the genotypes.

22. Labrador retrievers may be black, brown, or golden in color (see the chapter opening photograph on p. 69). Although each color may breed true, many different outcomes occur if numerous litters are examined from a variety of matings, where the parents are not necessarily true-breeding. The following results show some of the possibilities. Propose a mode of inheritance that is consistent with these data, and indicate the corresponding genotypes of the parents in each mating. Indicate as well the genotypes of dogs that breed true for each color.

(a) black × brown ⟶ all black
(b) black × brown ⟶ 1/2 black
1/2 brown
(c) black × brown ⟶ 3/4 black
1/4 golden
(d) black × golden ⟶ all black
(e) black × golden ⟶ 4/8 golden
3/8 black
1/8 brown
(f) black × golden ⟶ 2/4 golden
1/4 black
1/4 brown
(g) brown × brown ⟶ 3/4 brown
1/4 golden
(h) black × black ⟶ 9/16 black
4/16 golden
3/16 brown

23. Three autosomal recessive mutations in yeast, all producing the same phenotype (*m1*, *m2*, and *m3*), are subjected to complementation analysis. Of the results shown below, which, if any, are alleles of one another? Predict the results of the cross that is not shown—that is, *m2* × *m3*.

Cross 1: *m1* × *m2* ⟶ F_1: all wild-type progeny
Cross 2: *m1* × *m3* ⟶ F_1: all mutant progeny

24. Horses can be cremello (a light cream color), chestnut (a reddish brown color), or palomino (a golden color with white in the horse's tail and mane).

Chestnut

Palomino

Cremello

Of these phenotypes, only palominos never breed true. The following results have been observed:

cremello × palomino	⟶	1/2 cremello 1/2 palomino
chestnut × palomino	⟶	1/2 chestnut 1/2 palomino
palomino × palomino	⟶	1/4 chestnut 1/2 palomino 1/4 cremello

(a) From these results, determine the mode of inheritance by assigning gene symbols and indicating which genotypes yield which phenotypes.
(b) Predict the F_1 and F_2 results of many initial matings between cremello and chestnut horses.

25. Pigment in the mouse is produced only when the *C* allele is present. Individuals of the *cc* genotype have no color. If color is present, it may be determined by the *A* and *a* alleles. *AA* or *Aa* results in agouti color, whereas *aa* results in black coats.

(a) What F_1 and F_2 genotypic and phenotypic ratios are obtained from a cross between *AACC* and *aacc* mice?
(b) In the three crosses shown here between agouti females whose genotypes were unknown and males of the *aacc* genotype, what are the genotypes of the female parents for each of the following phenotypic ratios?

(1)	(2)	(3)
8 agouti 8 colorless	9 agouti 10 black	4 agouti 5 black 10 colorless

26. Five human matings numbered 1–5 are shown in the following table. Included are both maternal and paternal phenotypes for ABO and MN blood-group antigen status.

Parental Phenotypes					Offspring	
(1)	A, M	×	A, N		(a)	A, N
(2)	B, M	×	B, M		(b)	O, N
(3)	O, N	×	B, N		(c)	O, MN
(4)	AB, M	×	O, N		(d)	B, M
(5)	AB, MN	×	AB, MN		(e)	B, MN

Each mating resulted in one of the five offspring shown to the right (a–e). Match each offspring with one correct set of parents, using each parental set only once. Is there more than one set of correct answers?

27. Two mothers give birth to sons at the same time at a busy urban hospital. The son of mother 1 is afflicted with hemophilia, a disease caused by an X-linked recessive allele. Neither parent has the disease. Mother 2 has a normal son, despite the fact that the father has hemophilia. Several years later, couple 1 sues the hospital, claiming that these two newborns were swapped in the nursery following their birth. As a genetic counselor, you are called to testify. What information can you provide the jury concerning the allegation?

28. In Dexter and Kerry cattle, animals may be polled (hornless) or horned. The Dexter animals have short legs, whereas the Kerry animals have long legs. When many offspring were obtained from matings between polled Kerrys and horned Dexters, half were found to be polled Dexters and half polled Kerrys. When these two types of F_1 cattle were mated to one another, the following F_2 data were obtained:

3/8 polled Dexters	3/8 polled Kerrys
1/8 horned Dexters	1/8 horned Kerrys

A geneticist was puzzled by these data and interviewed farmers who had bred these cattle for decades. She learned that Kerrys were true-breeding. Dexters, on the other hand, were not true-breeding and never produced as many offspring as Kerrys. Provide a genetic explanation for these observations.

29. What genetic criteria distinguish a case of extranuclear inheritance from (a) a case of Mendelian autosomal inheritance; (b) a case of X-linked inheritance?

30. The specification of the anterior–posterior axis in *Drosophila* embryos is initially controlled by various gene products that are synthesized and stored in the mature egg following oogenesis. Mutations in these genes result in abnormalities of the axis during embryogenesis, illustrating maternal effect. How do such mutations vary from those involved in organelle heredity that illustrate extranuclear inheritance? Devise a set of parallel crosses and expected outcomes involving mutant genes that contrast maternal effect and organelle heredity.

31. The maternal-effect mutation *bicoid* (*bcd*) is recessive. In the absence of the bicoid protein product, embryogenesis is not completed. Consider a cross between a female heterozygous for the bicoid mutation (bcd^+/bcd^-) and a homozygous male (bcd^-/bcd^-)

(a) How is it possible for a male homozygous for the mutation to exist?
(b) Predict the outcome (normal vs. failed embryogenesis) in the F_1 and F_2 generations of the cross described.

32. Students taking a genetics exam were expected to answer the following question by converting data to a "meaningful ratio" and then solving the problem. The instructor assumed that the final ratio would reflect two gene pairs, and most correct answers did. Here is the exam question:
"Flowers may be white, orange, or brown. When plants with white flowers are crossed with plants with brown flowers, all the F_1 flowers are white. For F_2 flowers, the following data were obtained:

48 white
12 orange
4 brown

Convert the F_2 data to a meaningful ratio that allows you to explain the inheritance of color. Determine the number of genes involved and the genotypes that yield each phenotype."

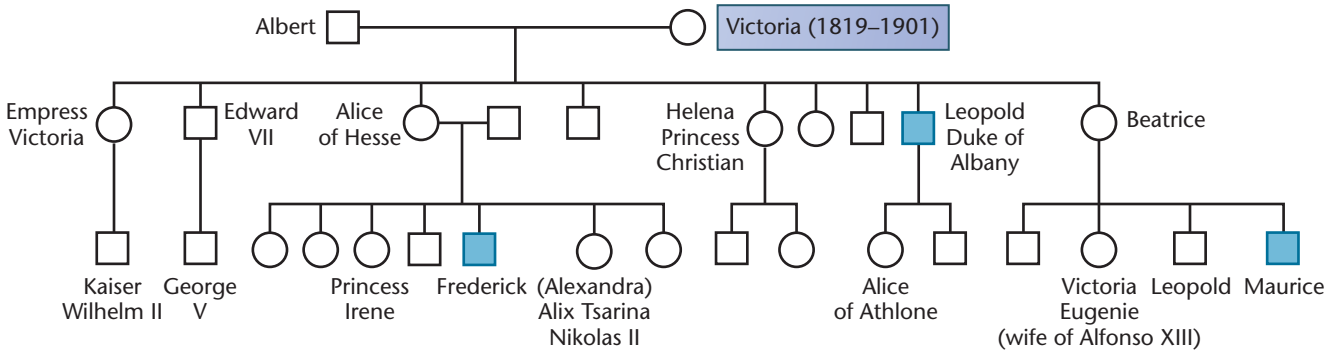
(a) Solve the problem for two gene pairs. What is the final F_2 ratio?
(b) A number of students failed to reduce the ratio for two gene pairs as described above and solved the problem using three gene pairs. When examined carefully, their solution was deemed a valid response by the instructor. Solve the problem using three gene pairs.

33. In four o'clock plants, many flower colors are observed. In a cross involving two true-breeding strains, one crimson and the other white, all of the F_1 generation were rose color. In the F_2, four new phenotypes appeared along with the P_1 and F_1 parental colors. The following ratio was obtained:

1/16 crimson
2/16 orange
1/16 yellow
2/16 magenta
4/16 rose
2/16 pale yellow
4/16 white

Propose an explanation for the inheritance of these flower colors.

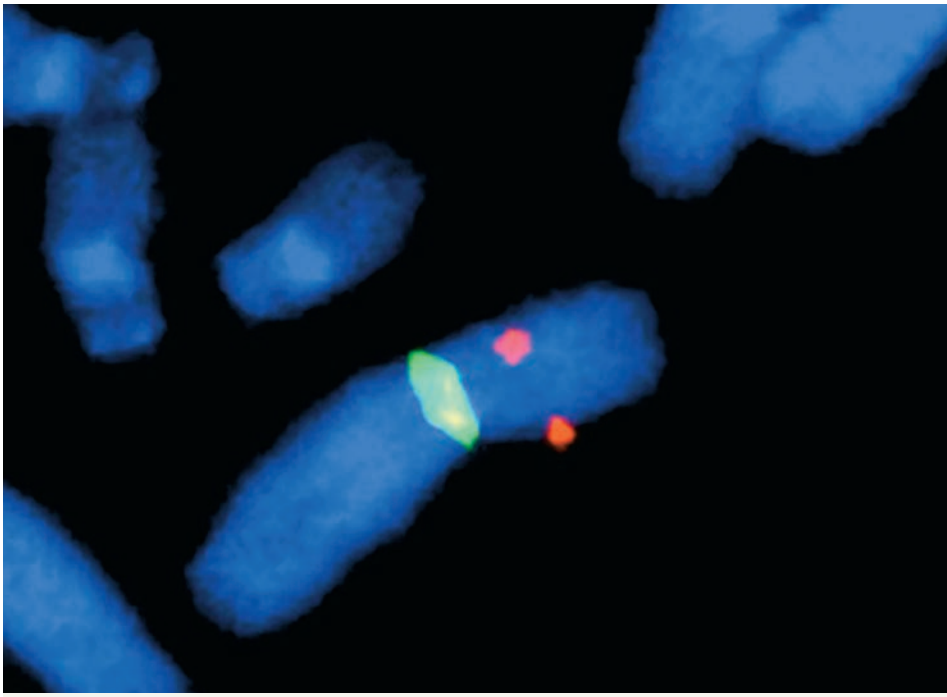
34. Below is a partial pedigree of hemophilia in the British Royal Family descended from Queen Victoria, who is believed to be the original "carrier" in this pedigree. Analyze the pedigree and indicate which females are also certain to be carriers. What is the probability that Princess Irene is a carrier?



5 Sex Determination and Sex Chromosomes

CHAPTER CONCEPTS

- A variety of mechanisms have evolved that result in sexual differentiation, leading to sexual dimorphism and greatly enhancing the production of genetic variation within species.
- Often, specific genes, usually on a single chromosome, cause maleness or femaleness during development.
- In humans, the presence of extra X or Y chromosomes beyond the diploid number may be tolerated but often leads to syndromes demonstrating distinctive phenotypes.
- While segregation of sex-determining chromosomes should theoretically lead to a one-to-one sex ratio of males to females, in humans the actual ratio favors males at conception.
- In mammals, females inherit two X chromosomes compared to one in males, but the extra genetic information in females is compensated for by random inactivation of one of the X chromosomes early in development.
- In some reptilian species, temperature during incubation of eggs determines the sex of offspring.



A human X chromosome highlighted using fluorescence *in situ* hybridization (FISH), a method in which specific probes bind to specific sequences of DNA. The probe producing green fluorescence binds to DNA at the centromere of X chromosomes. The probe producing red fluorescence binds to the DNA sequence of the X-linked Duchenne muscular dystrophy (DMD) gene.

In the biological world, a wide range of reproductive modes and life cycles are observed. Some organisms are entirely asexual, displaying no evidence of sexual reproduction. Other organisms alternate between short periods of sexual reproduction and prolonged periods of asexual reproduction. In most diploid eukaryotes, however, sexual reproduction is the only natural mechanism for producing new members of the species. The perpetuation of all sexually reproducing organisms depends ultimately on an efficient union of gametes during fertilization. In turn, successful fertilization depends on some form of **sexual differentiation** in the reproductive organisms. Even though it is not overtly evident, this differentiation occurs in organisms as low on the evolutionary scale as bacteria and single-celled eukaryotic algae. In more complex forms of life, the differentiation of the sexes is more evident as phenotypic dimorphism of males and females. The ancient symbol for iron and for Mars, depicting a shield and spear (♂), and the ancient symbol for copper and for Venus, depicting a mirror (♀), have also come to symbolize maleness and femaleness, respectively.

Dissimilar, or **heteromorphic, chromosomes,** such as the XY pair in mammals, characterize one sex or the other in a wide range of species, resulting in their label as **sex chromosomes.** Nevertheless, it is genes,

rather than chromosomes, that ultimately serve as the underlying basis of **sex determination.** As we will see, some of these genes are present on sex chromosomes, but others are autosomal. Extensive investigation has revealed a wide variation in sex-chromosome systems—even in closely related organisms—suggesting that mechanisms controlling sex determination have undergone rapid evolution many times in the history of life.

In this chapter, we delve more deeply into what is known about the genetic basis for the determination of sexual differences, with a particular emphasis on two organisms: our own species, representative of mammals; and *Drosophila*, on which pioneering sex-determining studies were performed.

5.1 X and Y Chromosomes Were First Linked to Sex Determination Early in the Twentieth Century

How sex is determined has long intrigued geneticists. In 1891, Hermann Henking identified a nuclear structure in the sperm of certain insects, which he labeled the X-body. Several years later, Clarence McClung showed that some of the sperm in grasshoppers contain an unusual genetic structure, called a *heterochromosome,* but the remainder of the sperm lack this structure. He mistakenly associated the presence of the heterochromosome with the production of male progeny. In 1906, Edmund B. Wilson clarified Henking and McClung's findings when he demonstrated that female somatic cells in the butterfly *Protenor* contain 14 chromosomes, including two X chromosomes. During oogenesis, an even reduction occurs, producing gametes with seven chromosomes, including one X chromosome. Male somatic cells, on the other hand, contain only 13 chromosomes, including one X chromosome. During spermatogenesis, gametes are produced containing either six chromosomes, without an X, or seven chromosomes, one of which is an X. Fertilization by X-bearing sperm results in female offspring, and fertilization by X-deficient sperm results in male offspring [**Figure 5–1(a)**].

The presence or absence of the X chromosome in male gametes provides an efficient mechanism for sex determination in this species and also produces a 1:1 sex ratio in the resulting offspring. As we saw earlier, *C. elegans* exhibits this system of sex determination.

Wilson also experimented with the milkweed bug *Lygaeus turcicus*, in which both sexes have 14 chromosomes. Twelve of these are autosomes (A). In addition, the females have two X chromosomes, while the males have only a single X and a smaller heterochromosome labeled the **Y chromosome**. Females in this species produce only

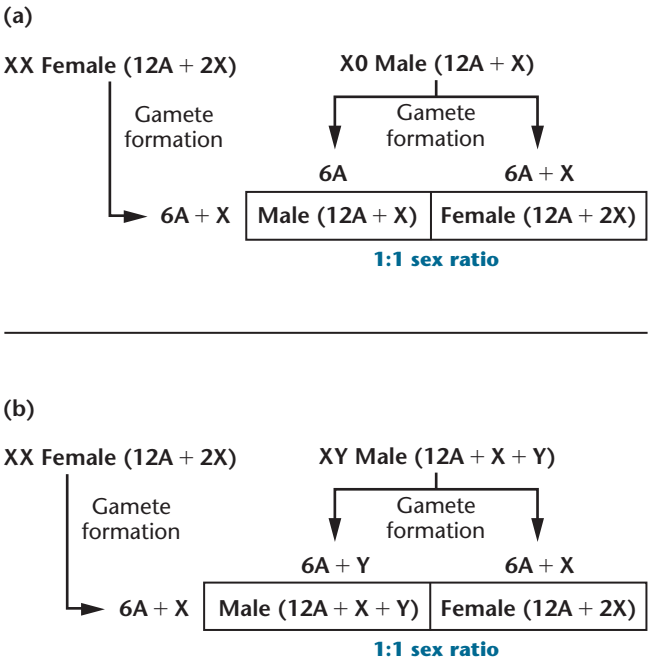


FIGURE 5–1 (a) Sex determination where the heterogametic sex (the male in this example) is X0 and produces gametes with or without the X chromosome; (b) sex determination, where the heterogametic sex (again, the male in this example) is XY and produces gametes with either an X or a Y chromosome. In both cases, the chromosome composition of the offspring determines its sex.

gametes of the (6A + X) constitution, but males produce two types of gametes in equal proportions, (6A + X) and (6A + Y). Therefore, following random fertilization, equal numbers of male and female progeny will be produced with distinct chromosome complements [**Figure 5–1(b)**].

In *Protenor* and *Lygaeus* insects, males produce unlike gametes. As a result, they are described as the **heterogametic sex**, and in effect, their gametes ultimately determine the sex of the progeny in those species. In such cases, the female, which has like sex chromosomes, is the **homogametic sex,** producing uniform gametes with regard to chromosome numbers and types.

The male is not always the heterogametic sex. In some organisms, the female produces unlike gametes, exhibiting either the *Protenor* XX/XO or *Lygaeus* XX/XY mode of sex determination. Examples include certain moths and butterflies, some fish, reptiles, amphibians, at least one species of plants (*Fragaria orientalis*), and most birds. To immediately distinguish situations in which the female is the heterogametic sex, some geneticists use the notation **ZZ/ZW,** where ZZ is the homogametic male and ZW is the heterogametic female, instead of the XX/XY notation. For example, chickens are so denoted.

ESSENTIAL POINT

Specific sex chromosomes contain genetic information that controls sex determination and sexual differentiation. ■

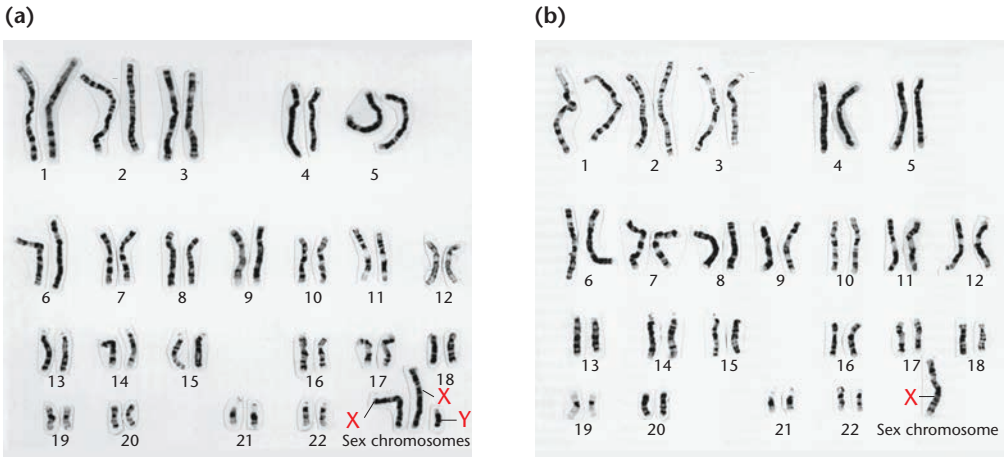


FIGURE 5–2 The karyotypes of individuals with (a) Klinefelter syndrome (47,XXY) and (b) Turner syndrome (45,X).

5.2 The Y Chromosome Determines Maleness in Humans

The first attempt to understand sex determination in our own species occurred almost 100 years ago and involved the visual examination of chromosomes in dividing cells. Efforts were made to accurately determine the diploid chromosome number of humans, but because of the relatively large number of chromosomes, this proved to be quite difficult. Then, in 1956, Joe Hin Tjio and Albert Levan discovered an effective way to prepare chromosomes for accurate viewing. This technique led to a strikingly clear demonstration of metaphase stages showing that 46 was indeed the human diploid number. Later that same year, C. E. Ford and John L. Hamerton, also working with testicular tissue, confirmed this finding. The familiar karyotypes of a human male (Figure 2–4) illustrate the difference in size between the human X and Y chromosomes.

Of the normal 23 pairs of human chromosomes, one pair was shown to vary in configuration in males and females. These two chromosomes were designated the X and Y sex chromosomes. The human female has two X chromosomes, and the human male has one X and one Y chromosome.

We might believe that this observation is sufficient to conclude that the Y chromosome determines maleness. However, several other interpretations are possible. The Y could play no role in sex determination; the presence of two X chromosomes could cause femaleness; or maleness could result from the lack of a second X chromosome. The evidence that clarified which explanation was correct came from study of the effects of human sex-chromosome variations, described in the following section. As such investigations revealed, the Y chromosome does indeed determine maleness in humans.

Klinefelter and Turner Syndromes

Around 1940, scientists identified two human abnormalities characterized by aberrant sexual development, **Klinefelter syndrome (47,XXY)** and **Turner syndrome (45,X).*** Individuals with Klinefelter syndrome are generally tall and have long arms and legs and large hands and feet. They usually have genitalia and internal ducts that are male, but their testes are rudimentary and fail to produce sperm. At the same time, feminine sexual development is not entirely suppressed. Slight enlargement of the breasts (gynecomastia) is common, and the hips are often rounded. This ambiguous sexual development, referred to as intersexuality, can lead to abnormal social development. Intelligence is often below the normal range as well.

In Turner syndrome, the affected individual has female external genitalia and internal ducts, but the ovaries are rudimentary. Other characteristic abnormalities include short stature (usually under 5 feet), skin flaps on the back of the neck, and underdeveloped breasts. A broad, shieldlike chest is sometimes noted. Intelligence is usually normal.

In 1959, the karyotypes of individuals with these syndromes were determined to be abnormal with respect to the sex chromosomes. Individuals with Klinefelter syndrome have more than one X chromosome. Most often they have an XXY complement in addition to 44 autosomes [**Figure 5–2(a)**], which is why people with this karyotype are designated 47,XXY. Individuals with Turner syndrome most often have only 45 chromosomes, including just a single X chromosome; thus, they are designated 45,X [**Figure 5–2(b)**]. Note the convention used in designating these chromosome compositions.

* Although the possessive form of the names of eponymous syndromes is sometimes used (e.g., Klinefelter's syndrome), the current preference is to use the nonpossessive form.

The number states the total number of chromosomes present, and the information after the comma indicates the deviation from the normal diploid content. Both conditions result from **nondisjunction,** the failure of the sex chromosomes to segregate properly during meiosis (nondisjunction is described in Chapter 6 and illustrated in Figure 6–1).

These Klinefelter and Turner karyotypes and their corresponding sexual phenotypes led scientists to conclude that the Y chromosome determines maleness in humans. In its absence, the person's sex is female, even if only a single X chromosome is present. The presence of the Y chromosome in the individual with Klinefelter syndrome is sufficient to determine maleness, even though male development is not complete. Similarly, in the absence of a Y chromosome, as in the case of individuals with Turner syndrome, no masculinization occurs. Note that we cannot conclude anything regarding sex determination under circumstances where a Y chromosome is present without an X because Y-containing human embryos lacking an X chromosome (designated 45,Y) do not survive.

Klinefelter syndrome occurs in about 1 of every 660 male births. The karyotypes **48,XXXY, 48,XXYY, 49,XXXXY**, and **49,XXXYY** are similar phenotypically to 47,XXY, but manifestations are often more severe in individuals with a greater number of X chromosomes.

Turner syndrome can also result from karyotypes other than 45,X, including individuals called **mosaics,** whose somatic cells display two different genetic cell lines, each exhibiting a different karyotype. Such cell lines result from a mitotic error during early development, the most common chromosome combinations being **45,X/46,XY** and **45,X/46,XX.** Thus, an embryo that began life with a normal karyotype can give rise to an individual whose cells show a mixture of karyotypes and who exhibits varying aspects of this syndrome.

Turner syndrome is observed in about 1 in 2000 female births, a frequency much lower than that for Klinefelter syndrome. One explanation for this difference is the observation that a substantial majority of 45,X fetuses die *in utero* and are aborted spontaneously. Thus, a similar frequency of the two syndromes may occur at conception.

47,XXX Syndrome

The abnormal presence of three X chromosomes along with a normal set of autosomes (**47,XXX**) results in female differentiation. The highly variable syndrome that accompanies this genotype, often called **triplo-X,** occurs in about 1 of 1000 female births. Frequently, 47,XXX women are perfectly normal and may remain unaware of their abnormality in chromosome number unless a karyotype is done. In other cases, underdeveloped secondary sex characteristics, sterility, delayed development of language and motor skills, and mental retardation may occur. In rare instances, **48,XXXX** (tetra-X) and **49,XXXXX** (penta-X) karyotypes have been reported. The syndromes associated with these karyotypes are similar to but more pronounced than the 47,XXX syndrome. Thus, in many cases, the presence of additional X chromosomes appears to disrupt the delicate balance of genetic information essential to normal female development.

47,XYY Condition

Another human condition involving the sex chromosomes is **47,XYY.** Studies of this condition, in which the only deviation from diploidy is the presence of an additional Y chromosome in an otherwise normal male karyotype, were initiated in 1965 by Patricia Jacobs. She discovered that 9 of 315 males in a Scottish maximum security prison had the 47,XYY karyotype. These males were significantly above average in height and had been incarcerated as a result of dangerous, violent, or criminal propensities. Of the nine males studied, seven were of subnormal intelligence, and all suffered personality disorders. Several other studies produced similar findings.

The possible correlation between this chromosome composition and criminal behavior piqued considerable interest, and extensive investigation of the phenotype and frequency of the 47,XYY condition in both criminal and noncriminal populations ensued. Above-average height (usually over 6 feet) and subnormal intelligence were substantiated, and the frequency of males displaying this karyotype was indeed revealed to be higher in penal and mental institutions compared with unincarcerated populations (one study showed 29 XYY males when 28,366 were examined [0.10%]). A particularly relevant question involves the characteristics displayed by the XYY males who are not incarcerated. The only nearly constant association is that such individuals are over 6 feet tall.

A study to further address this issue was initiated in 1974 to identify 47,XYY individuals at birth and to follow their behavioral patterns during preadult and adult development. While the study was considered unethical and soon abandoned, it has became clear that there are many XYY males present in the population who do not exhibit antisocial behavior and who lead normal lives. Therefore, we must conclude that there is a high, but not constant, correlation between the extra Y chromosome and the predisposition of these males to exhibit behavioral problems.

Sexual Differentiation in Humans

Once researchers had established that, in humans, it is the Y chromosome that houses genetic information necessary for maleness, they attempted to pinpoint a specific gene or genes capable of providing the "signal" responsible for sex determination. Before we delve into this topic, it is useful to

consider how sexual differentiation occurs in order to better comprehend how humans develop into sexually dimorphic males and females. During early development, every human embryo undergoes a period when it is potentially hermaphroditic. By the fifth week of gestation, gonadal primordia (the tissues that will form the gonad) arise as a pair of **gonadal (genital) ridges** associated with each embryonic kidney. The embryo is potentially hermaphroditic because at this stage its gonadal phenotype is sexually indifferent—male or female reproductive structures cannot be distinguished, and the gonadal ridge tissue can develop to form male or female gonads. As development progresses, primordial germ cells migrate to these ridges, where an outer cortex and inner medulla form (*cortex* and *medulla* are the outer and inner tissues of an organ, respectively). The cortex is capable of developing into an ovary, while the medulla may develop into a testis. In addition, two sets of undifferentiated ducts called the Wolffian and Müllerian ducts exist in each embryo. Wolffian ducts differentiate into other organs of the male reproductive tract, while Müllerian ducts differentiate into structures of the female reproductive tract.

Because gonadal ridges can form either ovaries or testes, they are commonly referred to as **bipotential gonads.** What switch triggers gonadal ridge development into testes or ovaries? The presence or absence of a Y chromosome is the key. If cells of the ridge have an XY constitution, development of the medulla into a testis is initiated around the seventh week. However, in the absence of the Y chromosome, no male development occurs, the cortex of the ridge subsequently forms ovarian tissue, and the Müllerian duct forms oviducts (Fallopian tubes), uterus, cervix, and portions of the vagina. Depending on which pathway is initiated, parallel development of the appropriate male or female duct system then occurs, and the other duct system degenerates. If testes differentiation is initiated, the embryonic testicular tissue secretes hormones that are essential for continued male sexual differentiation. As we will discuss in the next section, the presence of a Y chromosome and the development of the testes also inhibit formation of female reproductive organs.

In females, as the twelfth week of fetal development approaches, the oogonia within the ovaries begin meiosis, and primary oocytes can be detected. By the twenty-fifth week of gestation, all oocytes become arrested in meiosis and remain dormant until puberty is reached some 10 to 15 years later. In males, on the other hand, primary spermatocytes are not produced until puberty is reached (see Figure 2–11).

The Y Chromosome and Male Development

The human Y chromosome, unlike the X, was long thought to be mostly blank genetically. It is now known that this is not true, even though the Y chromosome contains far fewer genes than does the X. Data from the Human Genome Project

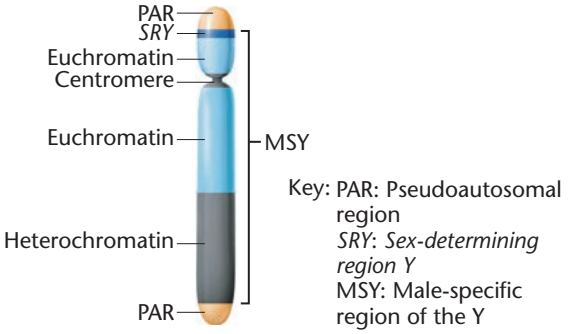


FIGURE 5–3 The regions of the human Y chromosome.

indicate that the Y chromosome has at least 75 genes, compared to 900–1400 genes on the X. Current analysis of these genes and regions with potential genetic function reveals that some have homologous counterparts on the X chromosome and others do not. For example, present on both ends of the Y chromosome are so-called **pseudoautosomal regions (PARs)** that share homology with regions on the X chromosome and synapse and recombine with it during meiosis. The presence of such a pairing region is critical to segregation of the X and Y chromosomes during male gametogenesis. The remainder of the chromosome, about 95 percent of it, does not synapse or recombine with the X chromosome. As a result, it was originally referred to as the *nonrecombining region of the Y (NRY)*. More recently, researchers have designated this region as the **male-specific region of the Y (MSY).** Some portions of the MSY share homology with genes on the X chromosome, and others do not.

The human Y chromosome is diagrammed in **Figure 5–3**. The MSY is divided about equally between *euchromatic* regions, containing functional genes, and *heterochromatic* regions, lacking genes. Within euchromatin, adjacent to the PAR of the short arm of the Y chromosome, is a critical gene that controls male sexual development, called the ***sex-determining region Y (SRY).*** In humans, the absence of a Y chromosome almost always leads to female development; thus, this gene is absent from the X chromosome. At six to eight weeks of development, the *SRY* gene becomes active in XY embryos. *SRY* encodes a protein that causes the undifferentiated gonadal tissue of the embryo to form testes. This protein is called the **testis-determining factor (TDF).** *SRY* (or a closely related version) is present in all mammals thus far examined, indicative of its essential function throughout this diverse group of animals.*

Our ability to identify the presence or absence of DNA sequences in rare individuals whose sex-chromosome

* It is interesting to note that in chickens, a similar gene has recently been identified. Called *DMRTI*, it is located on the Z chromosome. This gene is the subject of Problem 27 in the Problems section at the end of the chapter.

composition does not correspond to their sexual phenotype has provided evidence that *SRY* is the gene responsible for male sex determination. For example, there are human males who have two X and no Y chromosomes. Often, attached to one of their X chromosomes is the region of the Y that contains *SRY*. There are also females who have one X and one Y chromosome. Their Y is almost always missing the *SRY* gene. These observations argue strongly in favor of the role of *SRY* in providing the primary signal for male development.

Further support of this conclusion involves an experiment using **transgenic mice.** These animals are produced from fertilized eggs injected with foreign DNA that is subsequently incorporated into the genetic composition of the developing embryo. In normal mice, a chromosome region designated *Sry* has been identified that is comparable to *SRY* in humans. When mouse DNA containing *Sry* is injected into normal XX mouse eggs, most of the offspring develop into males.

The question of how the product of this gene triggers development of embryonic gonadal tissue into testes rather than ovaries is the key question under investigation. TDF is now believed to function as a *transcription factor*, a DNA-binding protein that interacts directly with the regulatory sequences of other genes to stimulate their expression. Thus, TDF behaves as a master switch that controls other genes downstream in the process of sexual differentiation. Interestingly, many identified thus far reside on autosomes, including the human *SOX9* gene located on chromosome 17 and the subject of the following Now Solve This entry.

NOW SOLVE THIS

5–1 Campomelic dysplasia (CMD1) is a congenital human syndrome featuring malformation of bone and cartilage. It is caused by an autosomal dominant mutation of a gene located on chromosome 17. Consider the following observations in sequence, and in each case, draw whatever appropriate conclusions are warranted.

(a) Of those with the syndrome who are karyotypically 46,XY, approximately 75 percent are sex reversed, exhibiting a wide range of female characteristics.
(b) The nonmutant form of the gene, called *SOX9*, is expressed in the developing gonad of the XY male, but not the XX female.
(c) The *SOX9* gene shares 71 percent amino acid coding sequence homology with the Y-linked *SRY* gene.
(d) CMD1 patients who exhibit a 46,XX karyotype develop as females, with no gonadal abnormalities.

■ **HINT:** *This problem asks you to apply the information presented in this chapter to a real-life example. The key to its solution is knowing that some genes are activated and produce their normal product as a result of expression of products of other genes found on different chromosomes.*

A more recent area of investigation has involved the Y chromosome and paternal aging. For many years, it has been known that maternal age is correlated with an elevated rate of offspring with chromosomal defects, including Down syndrome (see Chapter 6). Advanced paternal age has now been associated with an increased risk in offspring of congenital disorders with a genetic basis, including certain cancers, schizophrenia, autism, and other conditions, collectively known as *paternal age effects (PAE)*. Studies in which the genomes of sperm have been sequenced have demonstrated the presence of specific PAE mutations, including numerous ones on the Y chromosome. Evidence suggests that PAE mutations are positively selected for and result in an enrichment of mutant sperm over time.

ESSENTIAL POINT

The presence or absence of a Y chromosome that contains an intact *SRY* gene is responsible for causing maleness in humans. ■

5.3 The Ratio of Males to Females in Humans Is Not 1.0

The presence of heteromorphic sex chromosomes in one sex of a species but not the other provides a potential mechanism for producing equal proportions of male and female offspring. This potential depends on the segregation of the X and Y (or Z and W) chromosomes during meiosis, such that half of the gametes of the heterogametic sex receive one of the chromosomes and half receive the other one. As we learned in the previous section, small pseudoautosomal regions of pairing homology do exist at both ends of the human X and Y chromosomes, suggesting that the X and Y chromosomes do synapse and then segregate into different gametes. Provided that both types of gametes are equally successful in fertilization and that the two sexes are equally viable during development, a 1:1 ratio of male and female offspring should result.

The actual proportion of male to female offspring, referred to as the **sex ratio,** has been assessed in two ways. The **primary sex ratio** (**PSR**) reflects the proportion of males to females conceived in a population. The **secondary sex ratio** reflects the proportion of each sex that is born. The secondary sex ratio is much easier to determine but has the disadvantage of not accounting for any disproportionate embryonic or fetal mortality.

When the secondary sex ratio in the human population was determined in 1969 by using worldwide census data, it did not equal 1.0. For example, in the Caucasian population in the United States, the secondary ratio was a little less than 1.06, indicating that about 106 males were born for each 100 females. (In 1995, this ratio dropped to slightly less than 1.05.) In the African-American population in the United States, the ratio was 1.025. In other countries, the excess of

male births is even greater than is reflected in these values. For example, in Korea, the secondary sex ratio was 1.15.

Despite these ratios, it is possible that the PSR is 1.0 and is altered between conception and birth. For the secondary ratio to exceed 1.0, then, prenatal female mortality would have to be greater than prenatal male mortality. However, when this hypothesis was first examined, it was deemed to be false. In a Carnegie Institute study, reported in 1948, the sex of approximately 6000 embryos and fetuses recovered from miscarriages and abortions was determined, and fetal mortality was actually higher in males. On the basis of the data derived from that study, the PSR in U.S. Caucasians was estimated to be 1.079, suggesting that more males than females are conceived in the human population.

To explain why, researchers examined the assumptions on which the theoretical ratio is based:

1. Because of segregation, males produce equal numbers of X- and Y-bearing sperm.
2. Each type of sperm has equivalent viability and motility in the female reproductive tract.
3. The egg surface is equally receptive to both X- and Y-bearing sperm.

No direct experimental evidence contradicts any of these assumptions.

A PSR favoring male conceptions remained dogma for many decades until, in 2015, a study using an extensive data set was published that concludes that the PSR is 1.0—suggesting that equal numbers of males and females are indeed conceived. Among other parameters, the examination of the sex of 3-day-old and 6-day-old embryos conceived using assisted reproductive technology provided the most direct assessment. Following conception, however, mortality was then shown to fluctuate between the sexes, until at birth, more males than females are born. Thus, female mortality during embryonic and fetal development exceeds that of males. Clearly, this is a difficult topic to investigate but one of continued interest. For now, the most recent findings are convincing and contradict the earlier studies.

ESSENTIAL POINT

In humans, the sex ratio at conception and birth remains an active area of research. The most current study shows that equal numbers of males and females are conceived, but that more males than females are born. ■

5.4 Dosage Compensation Prevents Excessive Expression of X-Linked Genes in Humans and Other Mammals

The presence of two X chromosomes in normal human females and only one X in normal human males is unique compared with the equal numbers of autosomes present in the cells of both sexes. On theoretical grounds alone, it is possible to speculate that this disparity should create a "genetic dosage" difference between males and females, with attendant problems, for all X-linked genes. There is the potential for females to produce twice as much of each product of all X-linked genes. The additional X chromosomes in both males and females exhibiting the various syndromes discussed earlier in this chapter are thought to compound this dosage problem. Embryonic development depends on proper timing and precisely regulated levels of gene expression. Otherwise, disease phenotypes or embryonic lethality can occur. In this section, we will describe research findings regarding X-linked gene expression that demonstrate a genetic mechanism of **dosage compensation** that balances the dose of X chromosome gene expression in females and males.

Barr Bodies

Murray L. Barr and Ewart G. Bertram's experiments with cats, as well as Keith Moore and Barr's subsequent study in humans, demonstrate a genetic mechanism in mammals that compensates for X chromosome dosage disparities. Barr and Bertram observed a darkly staining body in the interphase nerve cells of female cats that was absent in similar cells of males. In humans, this body can be easily demonstrated in female cells derived from the buccal mucosa (cheek cells) or in fibroblasts (undifferentiated connective tissue cells), but not in similar male cells (**Figure 5–4**). This highly condensed structure, about 1 μm in diameter, lies against the nuclear envelope of interphase cells, and it stains positively for a number of different DNA-binding dyes.

This chromosome structure, called a **sex chromatin body,** or simply a **Barr body,** is an inactivated X chromosome. Susumo Ohno was the first to suggest that the Barr body arises from one of the two X chromosomes. This hypothesis is attractive because it provides a possible mechanism for dosage compensation. If one of the two

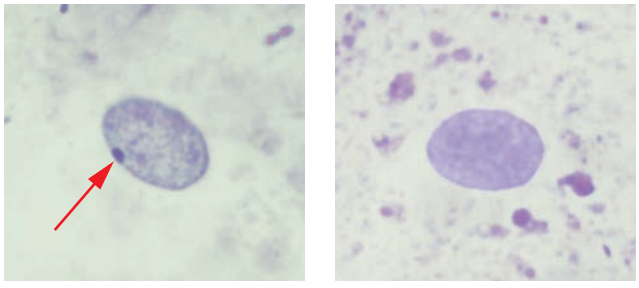


FIGURE 5–4 Photomicrographs comparing cheek epithelial cell nuclei from a male that fails to reveal Barr bodies (right) with a nucleus from a female that demonstrates a Barr body (indicated by the arrow in the left image). This structure, also called a sex chromatin body, represents an inactivated X chromosome.

X chromosomes is inactive in the cells of females, the dosage of genetic information that can be expressed in males and females will be equivalent. Convincing, though indirect, evidence for this hypothesis comes from study of the sex-chromosome syndromes described earlier in this chapter. Regardless of how many X chromosomes a somatic cell possesses, all but one of them appear to be inactivated and can be seen as Barr bodies. For example, no Barr body is seen in the somatic cells of Turner 45,X females; one is seen in Klinefelter 47,XXY males; two in 47,XXX females; three in 48,XXXX females; and so on (**Figure 5–5**). Therefore, the number of Barr bodies follows an $N - 1$ rule, where N is the total number of X chromosomes present.

Although this apparent inactivation of all but one X chromosome increases our understanding of dosage compensation, it further complicates our perception of other matters. For example, because one of the two X chromosomes is inactivated in normal human females, why then is the Turner 45,X individual not entirely normal? Why aren't females with the triplo-X and tetra-X karyotypes (47,XXX and 48,XXXX) completely unaffected by the additional X chromosome? Furthermore, in Klinefelter syndrome (47,XXY), X chromosome inactivation effectively renders the person 46,XY. Why aren't these males unaffected by the extra X chromosome in their nuclei?

One possible explanation is that chromosome inactivation does not normally occur in the very early stages of development of those cells destined to form gonadal tissues. Another possible explanation is that not all genes on each X chromosome forming a Barr body are inactivated. Recent studies have indeed demonstrated that as many as 15 percent of the human X chromosomal genes actually escape inactivation. Clearly, then, not every gene on the X requires inactivation. In either case, excessive expression of certain X-linked genes might still occur at critical times during development despite apparent inactivation of superfluous X chromosomes.

The Lyon Hypothesis

In mammalian females, one X chromosome is of maternal origin, and the other is of paternal origin. Which one is inactivated? Is the inactivation random? Is the same chromosome inactive in all somatic cells? In the early 1960s, Mary Lyon, Liane Russell, and Ernest Beutler independently proposed a hypothesis that answers these questions. They postulated that the inactivation of X chromosomes occurs randomly in somatic cells at a point early in embryonic development, most likely sometime during the blastocyst stage of development. Once inactivation has occurred, all descendant cells have the same X chromosome inactivated as their initial progenitor cell.

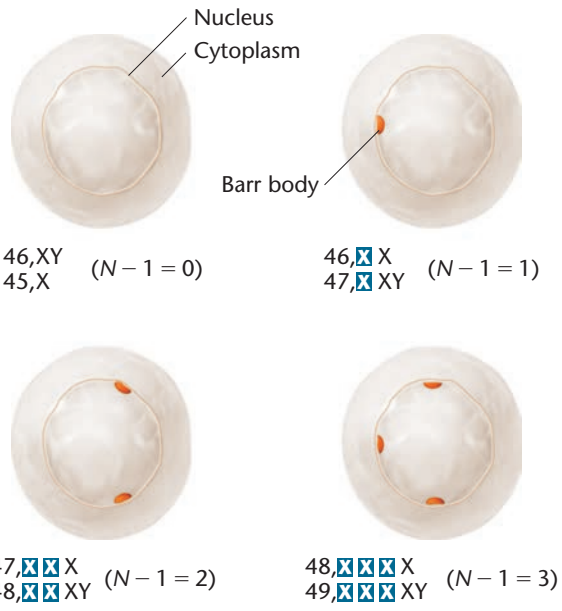


FIGURE 5–5 Occurrence of Barr bodies in various human karyotypes, where all X chromosomes except one ($N - 1$) are inactivated.

This explanation, which has come to be called the **Lyon hypothesis,** was initially based on observations of female mice heterozygous for X-linked coat-color genes. The pigmentation of these heterozygous females was mottled, with large patches expressing the color allele on one X and other patches expressing the allele on the other X. This is the phenotypic pattern that would be expected if different X chromosomes were inactive in adjacent patches of cells. Similar mosaic patterns occur in the black and yellow-orange patches of female tortoiseshell and calico cats (**Figure 5–6**). Such X-linked coat-color patterns do not occur in male cats because all their cells contain the single maternal X chromosome and are therefore hemizygous for only one X-linked coat-color allele.

The most direct evidence in support of the Lyon hypothesis comes from studies of gene expression in clones of human fibroblast cells. Individual cells are isolated following biopsy and cultured *in vitro*. A culture of cells derived from a single cell is called a **clone.** The synthesis of the enzyme glucose-6-phosphate dehydrogenase (G6PD) is controlled by an X-linked gene. Numerous mutant alleles of this gene have been detected, and their gene products can be differentiated from the wild-type enzyme by their migration pattern in an electrophoretic field.

Fibroblasts have been taken from females heterozygous for different allelic forms of *G6PD* and studied. The Lyon hypothesis predicts that if inactivation of an X chromosome occurs randomly early in development, and thereafter all progeny cells have the same X chromosome inactivated as their progenitor, such a female should show two

(a) (b)

FIGURE 5–6 (a) The random distribution of orange and black patches in a calico cat illustrates the Lyon hypothesis. The white patches are due to another gene, distinguishing calico cats from tortoiseshell cats (b), which lack the white patches.

types of clones, each containing only one electrophoretic form of *G6PD,* in approximately equal proportions. This prediction has been confirmed experimentally, and studies involving modern techniques in molecular biology have clearly established that X chromosome inactivation occurs.

One ramification of X-inactivation is that mammalian females are mosaics for all heterozygous X-linked alleles—some areas of the body express only the maternally derived alleles, and others express only the paternally derived alleles. An especially interesting example involves **red–green color blindness,** an X-linked recessive disorder. In humans, hemizygous males are fully color-blind in all retinal cells. However, heterozygous females display mosaic retinas, with patches of defective color perception and surrounding areas with normal color perception. In this example, random inactivation of one or the other X chromosome early in the development of heterozygous females has led to these phenotypes.

The Mechanism of Inactivation

The least understood aspect of the Lyon hypothesis is the mechanism of X chromosome inactivation. Somehow, either DNA, the attached histone proteins, or both DNA and histone proteins, are chemically modified, silencing most genes that are part of that chromosome. Once silenced, a memory is created that keeps the same homolog inactivated following chromosome replications and cell divisions. Such a process, whereby expression of genes on one homolog, but not the other, is affected, is referred to as **imprinting.** This term also applies to a number of other examples in which genetic information is modified and gene expression is repressed. Collectively, such events are part of the growing field of **epigenetics** (see Special Topics Chapter ST1).

NOW SOLVE THIS

5–2 Carbon Copy (CC), the first cat produced from a clone, was created from an ovarian cell taken from her genetic donor, Rainbow, a calico cat. The diploid nucleus from the cell was extracted and then injected into an enucleated egg. The resulting zygote was then allowed to develop in a petri dish, and the cloned embryo was implanted in the uterus of a surrogate mother cat, who gave birth to CC. CC's surrogate mother was a tabby (see the photo below). Geneticists were very interested in the outcome of cloning a calico cat because they were not certain if the cloned cat would have patches of orange and black, just orange, or just black. Taking into account the Lyon hypothesis, explain the basis of the uncertainty. Would you expect CC to appear identical to Rainbow? Explain why or why not.

Carbon Copy with her surrogate mother.

HINT: *This problem involves an understanding of the Lyon hypothesis. The key to its solution is to realize that the donor nucleus was from a differentiated ovarian cell of an adult female cat, which itself had inactivated one of its X chromosomes.*

Ongoing investigations are beginning to clarify the mechanism of inactivation. A region of the mammalian X chromosome is the major control unit. This region, located on the proximal end of the p arm in humans, is called the **X inactivation center (*Xic*),** and its genetic expression *occurs only on the X chromosome that is inactivated*. The *Xic* is about 1 Mb (10^6 base pairs) in length and is known to contain several putative regulatory units and four genes. One of these, ***X-inactive specific transcript (XIST),*** is now known to be a critical gene for X-inactivation.

Several interesting observations have been made regarding the RNA that is transcribed from the *XIST* gene, many coming from experiments that used the equivalent gene in the mouse (*Xist*). First, the RNA product is quite large and does not encode a protein, and thus is not translated. The RNA products of *Xist* spread over and coat the X chromosome *bearing the gene that produced them*. Two other noncoding genes at the *Xic* locus, *Tsix* (an antisense partner of *Xist*) and *Xite,* are also believed to play important roles in X-inactivation.

A second observation is that transcription of *Xist* initially occurs at low levels on all X chromosomes. As the inactivation process begins, however, transcription continues, and is enhanced, only on the X chromosome that becomes inactivated. In 1996, a research group led by Neil Brockdorff and Graeme Penny provided convincing evidence that transcription of *Xist* is the critical event in chromosome inactivation. These researchers introduced a targeted deletion (7 kb) into this gene, disrupting its sequence. As a result, the chromosome bearing the deletion lost its ability to become inactivated.

ESSENTIAL POINT

In mammals, female somatic cells randomly inactivate one of two X chromosomes during early embryonic development, a process important for balancing the expression of X chromosome-linked genes in males and females. ■

5.5 The Ratio of X Chromosomes to Sets of Autosomes Can Determine Sex

We now discuss two interesting cases where the Y chromosome does not play a role in sex determination. First, in the fruit fly, *Drosophila melanogaster*, even though most males contain a Y chromosome, the Y plays no role. Second, in the roundworm, *Caenorhabditis elegans*, the organism lacks a Y chromosome altogether. In both cases, we shall see that the critical factor is the ratio of X chromosomes to the number of sets of autosomes.

D. melanogaster

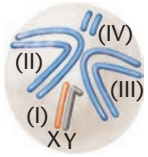
Because males and females in *Drosophila melanogaster* (and other *Drosophila* species) have the same general sex-chromosome composition as humans (males are XY and females are XX), we might assume that the Y chromosome also causes maleness in these flies. However, the elegant work of Calvin Bridges in 1921 showed this not to be true. His studies of flies with quite varied chromosome compositions led him to the conclusion that the Y chromosome is not involved in sex determination in this organism. Instead, Bridges proposed that the X chromosomes and autosomes together play a critical role in sex determination.

Bridges' work can be divided into two phases: (1) A study of offspring resulting from nondisjunction of the X chromosomes during meiosis in females and (2) subsequent work with progeny of females containing three copies of each chromosome, called triploid ($3n$) females. As we have seen previously in this chapter (and as you will see in Figure 6—1), nondisjunction is the failure of paired chromosomes to segregate or separate during the anaphase stage of the first or second meiotic divisions. The result is the production of two types of abnormal gametes, one of which contains an extra chromosome ($n + 1$) and the other of which lacks a chromosome ($n - 1$). Fertilization of such gametes with a haploid gamete produces ($2n + 1$) or ($2n - 1$) zygotes. As in humans, if nondisjunction involves the X chromosome, in addition to the normal complement of autosomes, both an XXY and an X0 sex-chromosome composition may result. (The "0" signifies that neither a second X nor a Y chromosome is present, as occurs in X0 genotypes of individuals with Turner syndrome.)

Contrary to what was later discovered in humans, Bridges found that the XXY flies were normal females and the X0 flies were sterile males. The presence of the Y chromosome in the XXY flies did not cause maleness, and its absence in the X0 flies did not produce femaleness. From these data, Bridges concluded that the Y chromosome in *Drosophila* lacks male-determining factors, but since the X0 males were sterile, it does contain genetic information essential to male fertility.

Bridges was able to clarify the mode of sex determination in *Drosophila* by studying the progeny of triploid females ($3n$), which have three copies each of the haploid complement of chromosomes. *Drosophila* has a haploid number of 4, thereby possessing three pairs of autosomes in addition to its pair of sex chromosomes. Triploid females apparently originate from rare diploid eggs fertilized by normal haploid sperm. Triploid females have heavy-set bodies, coarse bristles, and coarse eyes, and they may be fertile. Because of the odd number of each chromosome (3), during meiosis, a variety of different chromosome complements are distributed into gametes that give rise to offspring with a variety of abnormal chromosome constitutions. Correlations between the sexual morphology and chromosome

Normal diploid male



2 sets of autosomes
+
X Y

Chromosome formulation	Ratio of X chromosomes to autosome sets	Sexual morphology
3X:2A	1.5	Metafemale
3X:3A	1.0	Female
2X:2A	1.0	Female
3X:4A	0.75	Intersex
2X:3A	0.67	Intersex
X:2A	0.50	Male
XY:2A	0.50	Male
XY:3A	0.33	Metamale

FIGURE 5–7 The ratios of X chromosomes to sets of autosomes and the resultant sexual morphology seen in *Drosophila melanogaster.*

composition, along with Bridges' interpretation, are shown in **Figure 5–7.**

Bridges realized that the critical factor in determining sex is the ratio of X chromosomes to the number of haploid sets of autosomes (A) present. Normal (2X:2A) and triploid (3X:3A) females each have a ratio equal to 1.0, and both are fertile. As the ratio exceeds unity (3X:2A, or 1.5, for example), what was once called a *superfemale* is produced. Because such females are most often inviable, they are now more appropriately called **metafemales.**

Normal (XY:2A) and sterile (X0:2A) males each have a ratio of 1:2, or 0.5. When the ratio decreases to 1:3, or 0.33, as in the case of an XY:3A male, infertile **metamales** result. Other flies recovered by Bridges in these studies had an (X:A) ratio intermediate between 0.5 and 1.0. These flies were generally larger, and they exhibited a variety of morphological abnormalities and rudimentary bisexual gonads and genitalia. They were invariably sterile and expressed both male and female morphology, thus being designated as **intersexes.**

Bridges' results indicate that in *Drosophila*, factors that cause a fly to develop into a male are not located on the sex chromosomes but are instead found on the autosomes. Some female-determining factors, however, are located on the X chromosomes. Thus, with respect to primary sex determination, male gametes containing one of each autosome plus a Y chromosome result in male offspring not because of the presence of the Y but because they fail to contribute an X chromosome. This mode of sex determination is explained by the **genic balance theory.** Bridges proposed that a threshold for maleness is reached when the X:A ratio is 1:2 (X:2A), but that the presence of an additional X (XX:2A) alters the balance and results in female differentiation.

Numerous genes involved in sex determination in *Drosophila* have been identified. The recessive autosomal gene *transformer* (*tra*), discovered over 50 years ago by Alfred H. Sturtevant, clearly demonstrated that a single autosomal gene could have a profound impact on sex determination. Females homozygous for *tra* are transformed into sterile males, but homozygous males are unaffected. More recently, another gene, *Sex-lethal* (*Sxl*), has been shown to play a critical role, serving as a "master switch" in sex determination. Activation of the X-linked *Sxl* gene, which relies on a ratio of X chromosomes to sets of autosomes that equals 1.0, is essential to female development. In the absence of activation—as when, for example, the X:A ratio is 0.5—male development occurs.

Although it is not yet exactly clear how this ratio influences the *Sxl* locus, we do have some insights into the question. The *Sxl* locus is part of a hierarchy of gene expression and exerts control over other genes, including *tra* (discussed in the previous paragraph) and *dsx (doublesex)*. The wild-type allele of *tra* is activated by the product of *Sxl* only in females and in turn influences the expression of *dsx*. Depending on how the initial RNA transcript of *dsx* is processed (spliced, as explained below), the resultant dsx protein activates either male- or female-specific genes required for sexual differentiation. Each step in this regulatory cascade requires a form of processing called **RNA splicing,** in which portions of the RNA are removed and the remaining fragments are "spliced" back together prior to translation into a protein. In the case of the *Sxl* gene, the RNA transcript may be spliced in different ways, a phenomenon called **alternative splicing.** A different RNA transcript is produced in females than in males. In potential females, the transcript is active and initiates a cascade of regulatory gene expression, ultimately leading to female differentiation. In potential males, the transcript is inactive, leading to a different pattern of gene activity, whereby male differentiation occurs. We will return to this topic in Chapter 15, where alternative splicing is again addressed as one of the mechanisms involved in the regulation of genetic expression in eukaryotes.

Caenorhabditis elegans

The nematode worm *C. elegans* [**Figure 5–8(a)**] has become a popular organism in genetic studies, particularly for

(a)

(b)

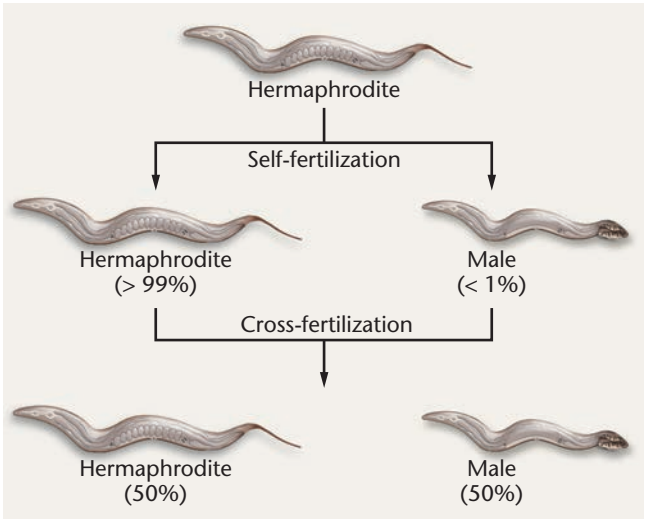


FIGURE 5–8 (a) Photomicrograph of a hermaphroditic nematode, *C. elegans*; (b) the outcomes of self-fertilization in a hermaphrodite, and a mating of a hermaphrodite and a male worm.

investigating the genetic control of development. Its usefulness is based on the fact that adults consist of approximately 1000 cells, the precise lineage of which can be traced back to specific embryonic origins. There are two sexual phenotypes in these worms: males, which have only testes, and hermaphrodites, which contain both testes and ovaries. During larval development of hermaphrodites, testes form that produce sperm, which is stored. Ovaries are also produced, but oogenesis does not occur until the adult stage is reached several days later. The eggs that are produced are fertilized by the stored sperm in a process of self-fertilization.

The outcome of this process is quite interesting [**Figure 5–8(b)**]. The vast majority of organisms that result are hermaphrodites, like the parental worm; less than 1 percent of the offspring are males. As adults, males can mate with hermaphrodites, producing about half male and half hermaphrodite offspring.

The genetic signal that determines maleness in contrast to hermaphroditic development is provided by genes located on both the X chromosome and autosomes. *C. elegans* lacks a Y chromosome altogether—hermaphrodites have two X chromosomes, while males have only one X chromosome. It is believed that, as in *Drosophila*, it is the ratio of X chromosomes to the number of sets of autosomes that ultimately determines the sex of these worms. A ratio of 1.0 (two X chromosomes and two copies of each autosome) results in hermaphrodites, and a ratio of 0.5 results in males. The absence of a heteromorphic Y chromosome is not uncommon in organisms.

5.6 Temperature Variation Controls Sex Determination in Reptiles

We conclude this chapter by discussing several cases involving reptiles, in which the environment—specifically temperature—has a profound influence on sex determination. In contrast to **chromosomal, or genotypic, sex determination (CSD** or **GSD),** in which sex is determined genetically (as is true of all examples thus far presented in the chapter), the cases that we will now discuss are categorized as **temperature-dependent sex determination (TSD).** As we shall see, the investigations leading to this information may well have come closer to revealing the true nature of the underlying basis of sex determination than any findings previously discussed.

In many species of reptiles, sex is predetermined at conception by sex-chromosome composition, as is the case in many organisms already considered in this chapter. For example, in many snakes, including vipers, a ZZ/ZW mode is in effect, in which the female is the heterogamous sex (ZW). However, in boas and pythons, it is impossible to distinguish one sex chromosome from the other in either sex. In many lizards, both the XX/XY and ZZ/ZW systems are found, depending on the species.

In still other reptilian species, however, TSD is the norm, including all crocodiles, most turtles, and some lizards, where sex determination is achieved according to the incubation temperature of eggs during a critical period of embryonic development. Three distinct patterns of TSD emerge (cases I–III in **Figure 5–9**). In case I, low temperatures yield 100 percent females, and high temperatures yield 100 percent males. Just the opposite occurs in case II. In case III, low *and* high temperatures yield 100 percent females, while intermediate temperatures yield various proportions of males. The third pattern is seen in various species of crocodiles, turtles, and lizards, although other members of these groups are known to exhibit the other patterns.

Two observations are noteworthy. First, in all three patterns, certain temperatures result in both male and female offspring; second, this pivotal temperature T_P range is fairly narrow, usually spanning less than 5°C, and sometimes only 1°C. The central question raised by these

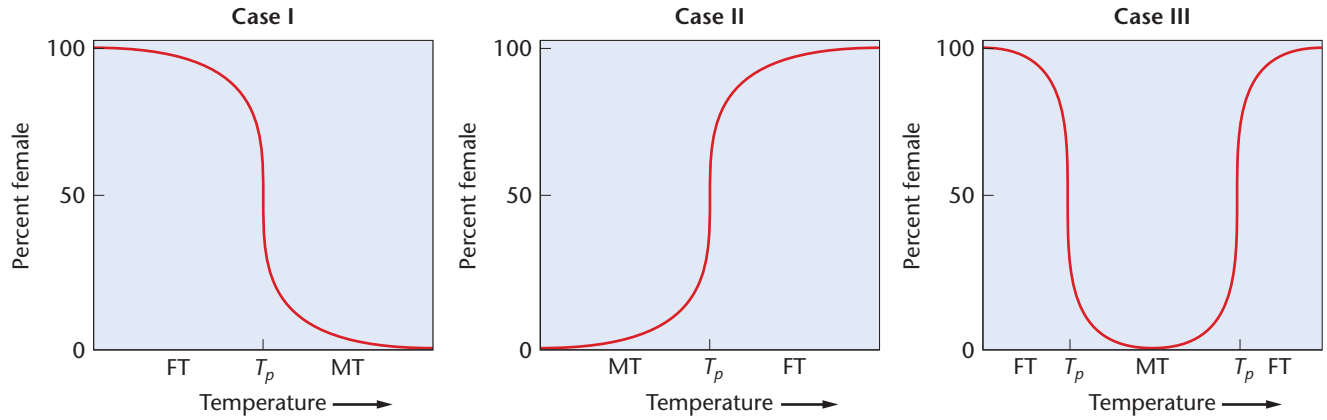


FIGURE 5–9 Three different patterns of temperature-dependent sex determination (TSD) in reptiles, as described in the text. The relative pivotal temperature T_p is crucial to sex determination during a critical point during embryonic development. FT = Female-determining temperature; MT = male-determining temperature

observations is: What are the metabolic or physiological parameters affected by temperature that lead to the differentiation of one sex or the other?

The answer is thought to involve steroids (mainly estrogens) and the enzymes involved in their synthesis. Studies clearly demonstrate that the effects of temperature on estrogens, androgens, and inhibitors of the enzymes controlling their synthesis are involved in the sexual differentiation of ovaries and testes. One enzyme in particular, **aromatase,** converts androgens (male hormones such as testosterone) to estrogens (female hormones such as estradiol). The activity of this enzyme is correlated with the pathway of reactions that occurs during gonadal differentiation activity and is high in developing ovaries and low in developing testes. Researchers in this field, including Claude Pieau and colleagues, have proposed that a thermosensitive factor mediates the transcription of the reptilian aromatase gene, leading to temperature-dependent sex determination. Several other genes are likely to be involved in this mediation.

The involvement of sex steroids in gonadal differentiation has also been documented in birds, fishes, and amphibians. Thus, sex-determining mechanisms involving estrogens seem to be characteristic of nonmammalian vertebrates. The regulation of such systems, while temperature-dependent in many reptiles, appears to be controlled by sex chromosomes (XX/XY or ZZ/ZW) in many of these other organisms. A final intriguing thought on this matter is that the product of *SRY*, a key component in mammalian sex determination, has been shown to bind *in vitro* to a regulatory portion of the aromatase gene, suggesting a mechanism whereby it could act as a repressor of ovarian development.

ESSENTIAL POINT

Many reptiles show temperature-dependent effects on sex determination. Although specific sex chromosomes determine genotypic sex in many reptiles, temperature effects on genes involved in sexual determination affect whether an embryo develops a male or female phenotype. ■

CASE STUDY | Not reaching puberty

Three adolescent girls were referred for genetic testing because their menstrual cycles had not started and their secondary female sex characteristics had not developed. While they all had the same symptoms, they differed greatly in height and appearance. One was short, one was very tall, and one was of the average height. One had skin flaps on the back of her neck. The tall girl had mild mental retardation, while the other two had normal intelligence. When their karyotypes were examined, they were found to be 45,X, 46,XY, and 48,XXXX.

1. What is the karyotype of the tall girl?
2. Which of these girls has Turner syndrome? Describe her appearance and karyotype.
3. Under what circumstances can a girl have a karyotype of 46,XY? What effects will this have on her appearance and intelligence?
4. What, if any, treatments can be given to each of these girls to enable them to develop normally? Will they be able to have children?

INSIGHTS AND SOLUTIONS

1. In *Drosophila*, the X chromosomes may become attached to one another ($\hat{XX}$) such that they always segregate together. Some flies thus contain a set of attached X chromosomes plus a Y chromosome.

   (a) What sex would such a fly be? Explain why this is so.

   (b) Given the answer to part (a), predict the sex of the offspring that would occur in a cross between this fly and a normal one of the opposite sex.

   (c) If the offspring described in part (b) are allowed to interbreed, what will be the outcome?

   Solution:
   (a) The fly will be a female. The ratio of X chromosomes to sets of autosomes—which determines sex in *Drosophila*—will be 1.0, leading to normal female development. The Y chromosome has no influence on sex determination in *Drosophila*.

   (b) All progeny flies will have two sets of autosomes along with one of the following sex-chromosome compositions:

   (1) $\hat{XXX}$ → a metafemale with 3 X's (called a trisomic)

   (2) $\hat{XXY}$ → a female like her mother

   (3) XY → a normal male

   (4) YY → no development occurs

   (c) A stock will be created that maintains attached-X females generation after generation.

2. The Xg cell-surface antigen is coded for by a gene located on the X chromosome. No equivalent gene exists on the Y chromosome. Two codominant alleles of this gene have been identified: *Xg1* and *Xg2*. A woman of genotype *Xg2/Xg2* bears children with a man of genotype *Xg1/Y*, and they produce a son with Klinefelter syndrome of genotype *Xg1/Xg2Y*. Using proper genetic terminology, briefly explain how this individual was generated. In which parent and in which meiotic division did the mistake occur?

   Solution: Because the son with Klinefelter syndrome is *Xg1/Xg2Y*, he must have received both the *Xg1* allele and the Y chromosome from his father. Therefore, nondisjunction must have occurred during meiosis I in the father.

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we have focused on sex differentiation, sex chromosomes, and genetic mechanisms involved in sex determination. At the same time, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, you should answer the following fundamental questions:
   (a) How do we know that in humans the X chromosomes play no role in sex determination, while the Y chromosome causes maleness and its absence causes femaleness?
   (b) How did we originally (in the late 1940s) analyze the sex ratio at conception in humans, and how has our approach to studying this issue changed in 2015?
   (c) How do we know that X chromosomal inactivation of either the paternal or maternal homolog is a random event during early development in mammalian females?
   (d) How do we know that *Drosophila* utilizes a different sex-determination mechanism than mammals, even though it has the same sex-chromosome compositions in males and females?

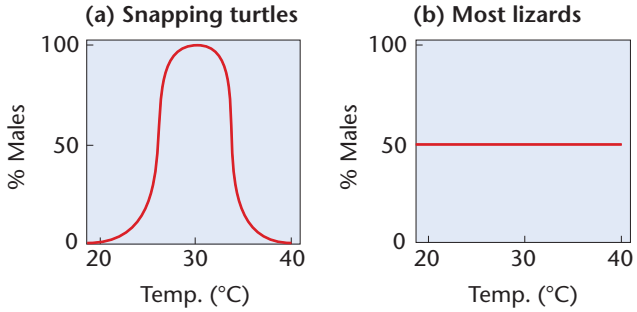
CONCEPT QUESTION

2. Review the Chapter Concepts list on p. 100. These all center on sex determination or the expression of genes encoded on sex chromosomes. Write a short essay that discusses sex chromosomes as they contrast with autosomes. ■

3. As related to sex determination, what is meant by
   (a) homomorphic and heteromorphic chromosomes; and
   (b) isogamous and heterogamous organisms?
4. Contrast the life cycle of a plant such as *Zea mays* with an animal such as *C. elegans*.
5. Distinguish between (a) the concepts of sexual differentiation and sex determination and (b) *Protenor* and *Lygaeus* modes of sex determination.
6. Describe the major differences between XO individuals in *Drosophila* and those in humans.
7. How do mammals, including humans, solve the "dosage problem" caused by the presence of an X and Y chromosome in one sex and two X chromosomes in the other sex?
8. What specific observations (evidence) support the conclusions about sex determination in *Drosophila* and humans?
9. Describe how nondisjunction in human female gametes can give rise to Klinefelter and Turner syndrome offspring following fertilization by a normal male gamete.
10. An insect species is discovered in which the heterogametic sex is unknown. An X-linked recessive mutation for *reduced wing* (*rw*) is discovered. Contrast the F_1 and F_2 generations from a cross between a female with reduced wings and a male with normal-sized wings when
    (a) the female is the heterogametic sex; and
    (b) the male is the heterogametic sex.
11. Given your answers to Problem 10, is it possible to distinguish between the *Protenor* and *Lygaeus* mode of sex determination based on the outcome of these crosses?
12. A group of scientists developing an XX zygote *in vitro* are curious to see the impact of certain chemicals on the development of the said organism. They incubate the zygote with the help of testosterone and some transcription factors, which are usually produced by the activity of the Y chromosome. They discover that the zygote develops into a sterile female with masculinized reproductive organs. Explain why this happens.
13. An attached-X female fly, $\hat{XXY}$ (see the "Insights and Solutions" box), expresses the recessive X-linked *white*-eye phenotype. It is crossed to a male fly that expresses the X-linked recessive

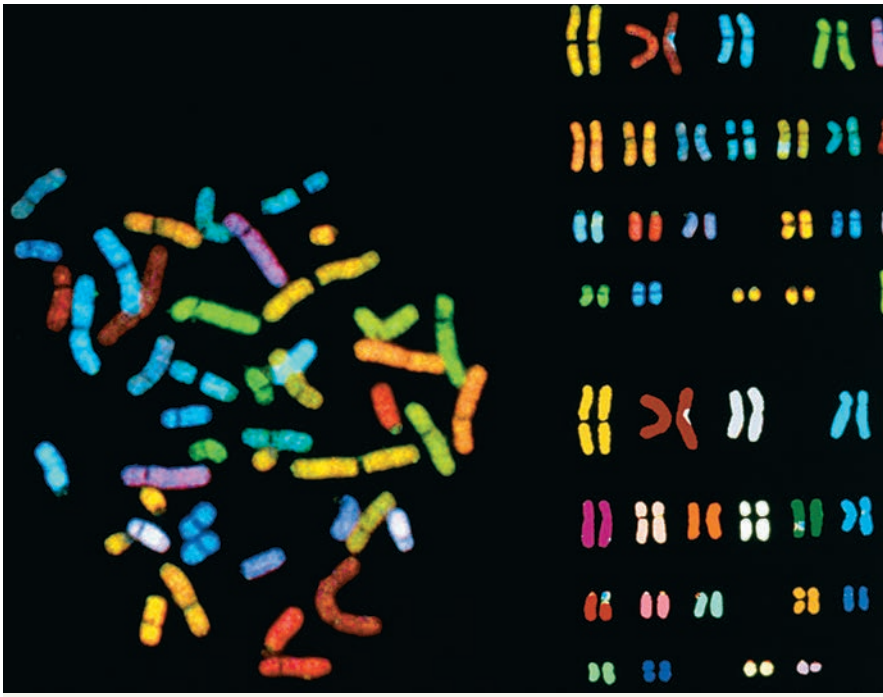
miniature wing phenotype. Determine the outcome of this cross in terms of sex, eye color, and wing size of the offspring.

14. Assume that on rare occasions the attached X chromosomes in female gametes become unattached. Based on the parental phenotypes in Problem 13, what outcomes in the F_1 generation would indicate that this has occurred during female meiosis?
15. It is believed that any male-determining genes contained on the Y chromosome in humans are not located in the limited region that synapses with the X chromosome during meiosis. What might be the outcome if such genes were located in this region?
16. What is a Barr body, and where is it found in a cell?
17. Indicate the expected number of Barr bodies in interphase cells of individuals with (a) triple X syndrome (XXX), (b) XYY syndrome, (c) Klinefelter syndrome, (d) Turner syndrome, and (e) karyotype 48, XXXX.
18. Define the Lyon hypothesis.
19. Can the Lyon hypothesis be tested in a human female who is homozygous for one allele of the X-linked *G6PD* gene? Why, or why not?
20. A cross is made between a female calico cat and a male cat having the gene for black fur on his X chromosome. What fraction of the offspring would one expect to be calico?
21. Under what circumstances can a male cat exhibit a tortoiseshell coat pattern?
22. What does the apparent need for dosage compensation mechanisms suggest about the expression of genetic information in normal diploid individuals?
23. A color-blind, chromatin-positive male child (one Barr body) has a maternal grandfather who was color blind. The boy's mother and father are phenotypically normal. Construct and support (using appropriately labeled diagrams) a rationale whereby the chromosomal and genetic attributes of the chromatin-positive male are fully explained.
24. In *Drosophila*, an individual female fly was observed to be of the XXY chromosome complement (normal autosomal complement) and to have white eyes as contrasted with the normal red eye color of wild type. The female's father had red eyes, and the mother had white eyes. Knowing that white eyes are X-linked and recessive, present an explanation for the genetic and chromosomal constitution of the XXY, white-eyed individual. It is important that you state in which parent and at what stage the chromosomal event occurred that caused the genetic and cytogenetic abnormality.
25. In mice, the X-linked dominant mutation *Testicular feminization* (*Tfm*) eliminates the normal response to the testicular hormone testosterone during sexual differentiation. An XY mouse bearing the *Tfm* allele on the X chromosome develops testes, but no further male differentiation occurs—the external genitalia of such an animal are female. From this information, what might you conclude about the role of the *Tfm* gene product and the X and Y chromosomes in sex determination and sexual differentiation in mammals? Can you devise an experiment, assuming you can "genetically engineer" the chromosomes of mice, to test and confirm your explanation?
26. Shown here are graphs that plot the percentage of fertilized eggs containing males against the atmospheric temperature during early development in (a) snapping turtles and (b) most lizards. Interpret these data as they relate to the effect of temperature on sex determination.



27. In chickens, a key gene involved in sex determination has recently been identified. Called *DMRT1*, it is located on the Z chromosome and is absent on the W chromosome. Like *SRY* in humans, it is male determining. Unlike *SRY* in humans, however, female chickens (ZW) have a single copy while males (ZZ) have two copies of the gene. Nevertheless, it is transcribed only in the developing testis. Working in the laboratory of Andrew Sinclair (a co-discoverer of the human *SRY* gene), Craig Smith and colleagues were able to "knock down" expression of *DMRT1* in ZZ embryos using RNA interference techniques (see Chapter 15). In such cases, the developing gonads look more like ovaries than testes [*Nature* 461: 267 (2009)]. What conclusions can you draw about the role that the *DMRT1* gene plays in chickens in contrast to the role the *SRY* gene plays in humans?

6 Chromosome Mutations: Variation in Number and Arrangement



Spectral karyotyping of human chromosomes utilizing differentially labeled "painting" probes.

CHAPTER CONCEPTS

- The failure of chromosomes to properly separate during meiosis results in variation in the chromosome content of gametes and subsequently in offspring arising from such gametes.
- Plants often tolerate an abnormal genetic content, but, as a result, they often manifest unique phenotypes. Such genetic variation has been an important factor in the evolution of plants.
- In animals, genetic information is in a delicate equilibrium whereby the gain or loss of a chromosome, or part of a chromosome, in an otherwise diploid organism often leads to lethality or to an abnormal phenotype.
- The rearrangement of genetic information within the genome of a diploid organism may be tolerated by that organism but may affect the viability of gametes and the phenotypes of organisms arising from those gametes.
- Chromosomes in humans contain fragile sites—regions susceptible to breakage, which lead to abnormal phenotypes.

In previous chapters, we have emphasized how mutations and the resulting alleles affect an organism's phenotype and how traits are passed from parents to offspring according to Mendelian principles. In this chapter, we look at phenotypic variation that results from more substantial changes than alterations of individual genes—modifications at the level of the chromosome.

Although most members of diploid species normally contain precisely two haploid chromosome sets, many known cases vary from this pattern. Modifications include a change in the total number of chromosomes, the deletion or duplication of genes or segments of a chromosome, and rearrangements of the genetic material either within or among chromosomes. Taken together, such changes are called **chromosome mutations** or **chromosome aberrations** in order to distinguish them from gene mutations. Because the chromosome is the unit of genetic transmission, according to Mendelian laws, chromosome aberrations are passed to offspring in a predictable manner, resulting in many unique genetic outcomes.

Because the genetic component of an organism is delicately balanced, even minor alterations of either content or location of genetic information within the genome can result in some form of phenotypic variation. More substantial changes may be lethal, particularly in animals. Throughout this chapter, we consider many types of chromosomal aberrations, the phenotypic consequences for the organism that harbors an aberration, and the impact of the aberration on the offspring of an affected individual. We will also discuss the role of chromosome aberrations in the evolutionary process.

6.1 Variation in Chromosome Number: Terminology and Origin

Variation in chromosome number ranges from the addition or loss of one or more chromosomes to the addition of one or more haploid sets of chromosomes. Before we embark on our discussion, it is useful to clarify the terminology that describes such changes. In the general condition known as **aneuploidy,** an organism gains or loses one or more chromosomes but not a complete set. The loss of a single chromosome from an otherwise diploid genome is called *monosomy.* The gain of one chromosome results in *trisomy.* These changes are contrasted with the condition of **euploidy,** where complete haploid sets of chromosomes are present. If more than two sets are present, the term **polyploidy** applies. Organisms with three sets are specifically *triploid,* those with four sets are *tetraploid,* and so on. Table 6.1 provides an organizational framework for you to follow as we discuss each of these categories of aneuploid and euploid variation and the subsets within them.

As we consider cases that include the gain or loss of chromosomes, it is useful to examine how such aberrations originate. For instance, how do the syndromes arise where the number of sex-determining chromosomes in humans is altered, as described in Chapter 5? As you may recall, the gain (47,XXY) or loss (45,X) of an X chromosome from an otherwise diploid genome affects the phenotype, resulting in **Klinefelter syndrome** or **Turner syndrome,** respectively (see Figure 5–4). Human females may contain extra X chromosomes (e.g., 47,XXX, 48,XXXX), and some males contain an extra Y chromosome (47,XYY).

TABLE 6.1 Terminology for Variation in Chromosome Numbers

Term	Explanation
Aneuploidy	$2n \pm x$ chromosomes
Monosomy	$2n - 1$
Disomy	$2n$
Trisomy	$2n + 1$
Tetrasomy, pentasomy, etc.	$2n + 2$, $2n + 3$, etc.
Euploidy	Multiples of n
Diploidy	$2n$
Polyploidy	$3n$, $4n$, $5n$, . . .
Triploidy	$3n$
Tetraploidy, pentaploidy, etc.	$4n$, $5n$, etc.
Autopolyploidy	Multiples of the same genome
Allopolyploidy (amphidiploidy)	Multiples of closely related genomes

Such chromosomal variation originates as a random error during the production of gametes, a phenomenon referred to as **nondisjunction,** whereby paired homologs fail to disjoin during segregation. This process disrupts the normal distribution of chromosomes into gametes. The results of nondisjunction during meiosis I and meiosis II for a single chromosome of a diploid organism are shown in Figure 6–1.

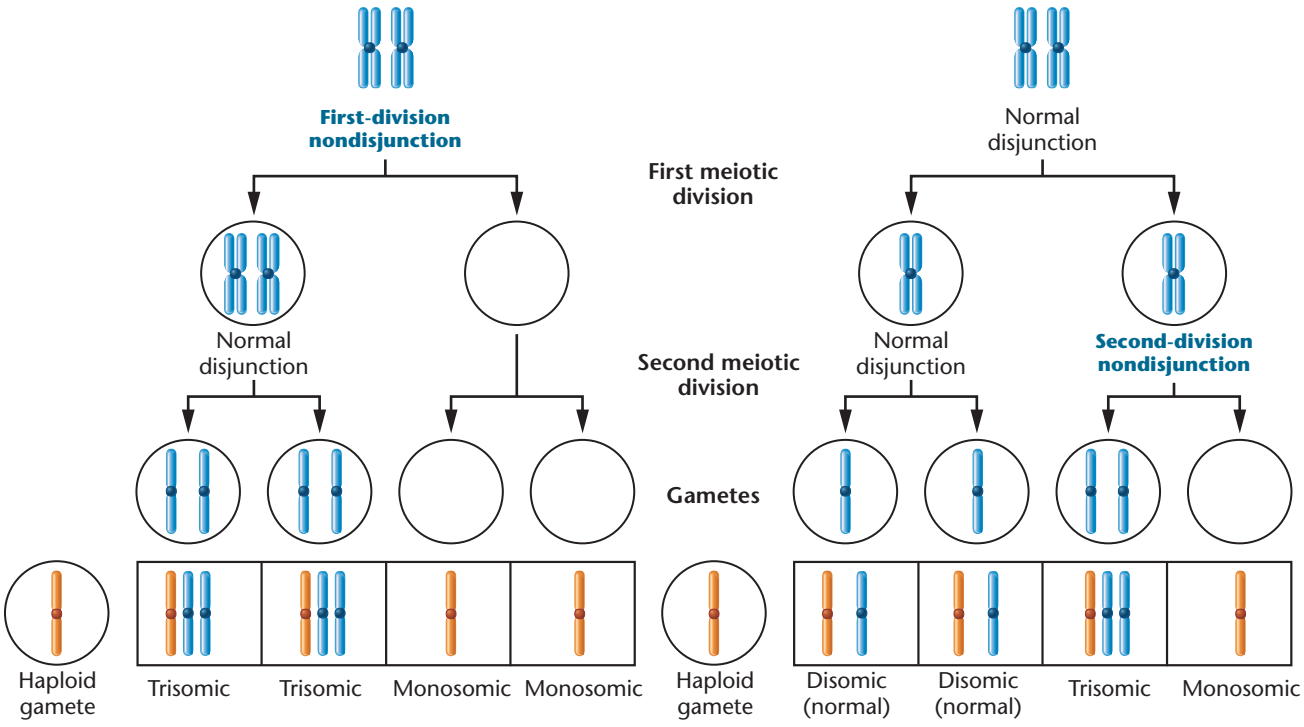


FIGURE 6–1 Nondisjunction during the first and second meiotic divisions. In both cases, some of the gametes that are formed either contain two members of a specific chromosome or lack that chromosome. After fertilization by a gamete with normal haploid content, monosomic, disomic (normal), or trisomic zygotes are produced.

As you can see, abnormal gametes can form that contain either two members of the affected chromosome or none at all. Fertilizing these with a normal haploid gamete produces a zygote with either three members (trisomy) or only one member (monosomy) of this chromosome. Nondisjunction leads to a variety of aneuploid conditions in humans and other organisms.

NOW SOLVE THIS

6–1 A human female with Turner syndrome (45,X) also expresses the X-linked trait hemophilia, as did her father. Which of her parents underwent nondisjunction during meiosis, giving rise to the gamete responsible for the syndrome?

■ **HINT:** *This problem involves an understanding of how nondisjunction leads to aneuploidy. The key to its solution is first to review Turner syndrome, discussed above and in more detail in Chapter 5, then to factor in that she expresses hemophilia, and finally, to consider which parent contributed a gamete with an X chromosome that underwent normal meiosis.*

ESSENTIAL POINT

Alterations of the precise diploid content of chromosomes are referred to as chromosomal aberrations or chromosomal mutations. ■

6.2 Monosomy and Trisomy Result in a Variety of Phenotypic Effects

We turn now to a consideration of variations in the number of autosomes and the genetic consequence of such changes. The most common examples of aneuploidy, where an organism has a chromosome number other than an exact multiple of the haploid set, are cases in which a single chromosome is either added to, or lost from, a normal diploid set.

Monosomy

The loss of one chromosome produces a $2n - 1$ complement called **monosomy.** Although monosomy for the X chromosome occurs in humans, as we have seen in 45,X Turner syndrome, monosomy for any of the autosomes is not usually tolerated in humans or other animals. In *Drosophila,* flies that are monosomic for the very small chromosome IV (containing less than 5 percent of the organism's genes) develop more slowly, exhibit reduced body size, and have impaired viability. Monosomy for the larger chromosomes II and III is apparently lethal because such flies have never been recovered.

The failure of monosomic individuals to survive is at first quite puzzling, since at least a single copy of every gene is present in the remaining homolog. However, one explanation is that if just one of those genes is represented by a lethal allele, monosomy unmasks the recessive lethal that is tolerated in heterozygotes carrying the corresponding wild-type allele, leading to the death of the organism. In other cases, a single copy of a recessive gene due to monosomy may be insufficient to provide life-sustaining function for the organism, a phenomenon called **haploinsufficiency.**

Aneuploidy is better tolerated in the plant kingdom. Monosomy for autosomal chromosomes has been observed in maize, tobacco, the evening primrose (*Oenothera*), and the jimson weed (*Datura*), among many other plants. Nevertheless, such monosomic plants are usually less viable than their diploid derivatives. Haploid pollen grains, which undergo extensive development before participating in fertilization, are particularly sensitive to the lack of one chromosome and are seldom viable.

Trisomy

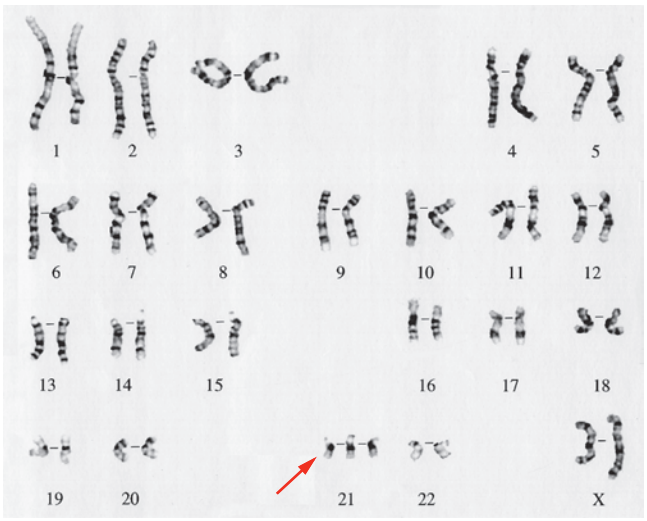
In general, the effects of **trisomy** ($2n + 1$) parallel those of monosomy. However, the addition of an extra chromosome produces somewhat more viable individuals in both animal and plant species than does the loss of a chromosome. In animals, this is often true, provided that the chromosome involved is relatively small. However, the addition of a large autosome to the diploid complement in both *Drosophila* and humans has severe effects and is usually lethal during development.

In plants, trisomic individuals are viable, but their phenotype may be altered. A classical example involves the jimson weed, *Datura,* whose diploid number is 24. Twelve primary trisomic conditions are possible, and examples of each one have been recovered. Each trisomy alters the phenotype of the plant's capsule sufficiently to produce a unique phenotype. These capsule phenotypes were first thought to be caused by mutations in one or more genes.

Still another example is seen in the rice plant (*Oryza sativa*), which has a haploid number of 12. Trisomic strains for each chromosome have been isolated and studied—the plants of 11 strains can be distinguished from one another and from wild-type plants. Trisomics for the longer chromosomes are the most distinctive, and the plants grow more slowly. This is in keeping with the belief that larger chromosomes cause greater genetic imbalance than smaller ones. Leaf structure, foliage, stems, grain morphology, and plant height also vary among the various trisomies.

Down Syndrome: Trisomy 21

The only human autosomal trisomy in which a significant number of individuals survive longer than a year past birth was discovered in 1866 by Langdon Down. The condition is



© Design Pics/Alamy

FIGURE 6–2 The karyotype and a photograph of a child with Down syndrome (hugging her unaffected sister on the right). In the karyotype, three members of the G-group chromosome 21 are present, creating the 47,21+ condition.

now known to result from trisomy of chromosome 21, one of the G group* (Figure 6–2), and is called **Down syndrome** or simply **trisomy 21** (designated 47,21+). This trisomy is found in approximately 1 infant in every 800 live births. While this might seem to be a rare, improbable event, there are approximately 4000–5000 such births annually in the United States, and there are currently over 250,000 individuals with Down syndrome.

Typical of other conditions classified as syndromes, many phenotypic characteristics *may* be present in trisomy 21, but any single affected individual usually exhibits only a subset of these. In the case of Down syndrome, there are 12 to 14 such characteristics, with each individual, on average, expressing 6 to 8 of them. Nevertheless, the outward appearance of these individuals is very similar, and they bear a striking resemblance to one another. This is, for the most part, due to a prominent epicanthic fold in each eye** and the typically flat face and round head. People with Down syndrome are also characteristically short and may have a protruding, furrowed tongue (which causes the mouth to remain partially open) and short, broad hands with characteristic palm and fingerprint patterns. Physical, psychomotor, and mental development are retarded, and poor muscle tone is characteristic. While life expectancy is shortened to an average of about 50 years, individuals are known to survive into their 60s.

Children afflicted with Down syndrome are prone to respiratory disease and heart malformations, and they show an incidence of leukemia approximately 20 times higher than that of the normal population. However, careful medical scrutiny and treatment throughout their lives can extend their survival significantly. A striking observation is that death in older Down syndrome adults is frequently due to Alzheimer disease. The onset of this disease occurs at a much earlier age than in the normal population.

Because Down syndrome is common in our population, a comprehensive understanding of the underlying genetic basis has long been a research goal. Investigations have given rise to the idea that a critical region of chromosome 21 contains the genes that are dosage sensitive in this trisomy and responsible for the many phenotypes associated with the syndrome. This hypothetical portion of the chromosome has been called the **Down syndrome critical region (DSCR).** A mouse model was created in 2004 that is trisomic for the DSCR, although some mice do not exhibit the characteristics of the syndrome. Nevertheless, this remains an important investigative approach.

Current studies of the DSCR region in both humans and mice have led to several interesting findings. We now believe that the three copies of the genes present in this region are necessary, but themselves not sufficient for the cognitive deficiencies characteristic of the syndrome. Another finding involves the important observation that Down syndrome individuals have a decreased risk of developing a number of cancers involving solid tumors, including lung cancer and melanoma. This health benefit has been correlated with the presence of an extra copy of the

*On the basis of size and centromere placement, human autosomal chromosomes are divided into seven groups: A (1–3), B (4–5), C (6–12), D (13–15), E (16–18), F (19–20), and G (21–22).

**The epicanthic fold, or epicanthus, is a skin fold of the upper eyelid, extending from the nose to the inner side of the eyebrow. It covers and appears to lower the inner corner of the eye, giving the eye a slanted, or almond-shaped, appearance. The epicanthus is a prominent normal component of the eyes in many Asian groups.

DSCR1 gene, which encodes a protein that suppresses *vascular endothelial growth factor* (*VEGF*). This suppression, in turn, blocks the process of angiogenesis. As a result, the overexpression of this gene inhibits tumors from forming proper vascularization, diminishing their growth. A 14-year study published in 2002 involving 17,800 Down syndrome individuals revealed an approximate 10 percent reduction in cancer mortality in contrast to a control population. No doubt, further information will be forthcoming from the study of the DSCR region.

The Origin of the Extra 21st Chromosome in Down Syndrome

Most frequently, this trisomic condition occurs through nondisjunction of chromosome 21 during meiosis. Failure of paired homologs to disjoin during either anaphase I or II may lead to gametes with the $n + 1$ chromosome composition. About 75 percent of these errors leading to Down syndrome are attributed to nondisjunction during the first meiotic division. Subsequent fertilization with a normal gamete creates the trisomic condition.

Chromosome analysis has shown that, while the additional chromosome may be derived from either the mother or father, the ovum is the source in about 95 percent of 47,21+ trisomy cases. Before the development of techniques using polymorphic markers to distinguish paternal from maternal homologs, this conclusion was supported by the more indirect evidence derived from studies of the age of mothers giving birth to infants afflicted with Down syndrome. **Figure 6–3** shows the relationship between the incidence of Down syndrome births and maternal age, illustrating the dramatic increase as the age of the mother increases.

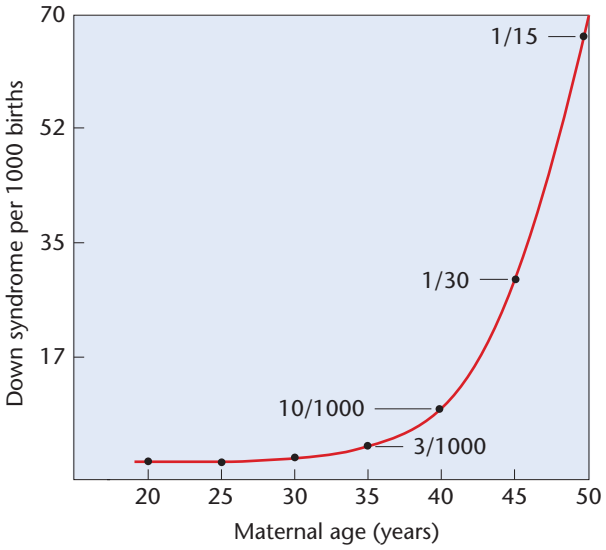


FIGURE 6–3 Incidence of Down syndrome births related to maternal age.

While the frequency is about 1 in 1000 at maternal age 30, a tenfold increase to a frequency of 1 in 100 is noted at age 40. The frequency increases still further to about 1 in 30 at age 45. A very alarming statistic is that as the age of childbearing women exceeds 45, the probability of a Down syndrome birth continues to increase substantially. In spite of this high probability, substantially more than half of Down syndrome births occur to women younger than 35 years, because the overwhelming proportion of pregnancies in the general population involve women under that age.

Although the nondisjunctional event that produces Down syndrome seems more likely to occur during oogenesis in women over the age of 35, we do not know with certainty why this is so. However, one observation may be relevant. Meiosis is initiated in all the eggs of a human female when she is still a fetus, until the point where the homologs synapse and recombination has begun. Then oocyte development is arrested in meiosis I. Thus, all primary oocytes have been formed by birth. When ovulation begins at puberty, meiosis is reinitiated in one egg during each ovulatory cycle and continues into meiosis II. The process is once again arrested after ovulation and is not completed unless fertilization occurs.

The end result of this progression is that each ovum that is released has been arrested in meiosis I for about a month longer than the one released during the preceding cycle. As a consequence, women 30 or 40 years old produce ova that are significantly older and that have been arrested longer than those they ovulated 10 or 20 years previously. In spite of the logic underlying this hypothesis explaining the cause of the increased incidence of Down syndrome as women age, it remains difficult to prove directly.

These statistics obviously pose a serious problem for the woman who becomes pregnant late in her reproductive years. Genetic counseling early in such pregnancies is highly recommended. Counseling informs prospective parents about the probability that their child will be affected and educates them about Down syndrome. Although some individuals with Down syndrome must be institutionalized, others benefit greatly from special education programs and may be cared for at home. Down syndrome children in general are noted for their affectionate, loving nature. A genetic counselor may also recommend a prenatal diagnostic technique in which fetal cells are isolated and cultured.

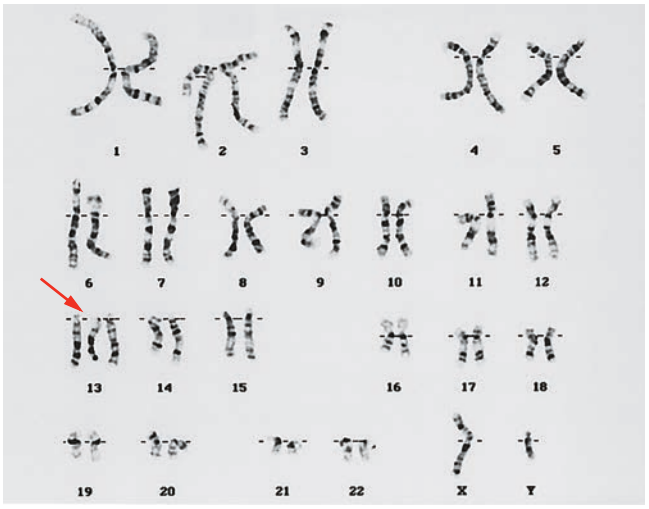
In **amniocentesis** and **chorionic villus sampling (CVS),** the two most familiar approaches, fetal cells are obtained from the amniotic fluid or the chorion of the placenta, respectively. In a newer approach, fetal cells and DNA are derived directly from the maternal circulation, a technique referred to as **noninvasive prenatal genetic diagnosis (NIPGD).** Requiring only a 10-mL maternal blood sample, this procedure will become increasingly

more common because it poses no risk to the fetus. After fetal cells are obtained and cultured, the karyotype can be determined by cytogenetic analysis. If the fetus is diagnosed as being affected, a therapeutic abortion is one option currently available to parents. Obviously, this is a difficult decision involving a number of religious and ethical issues.

Since Down syndrome is caused by a random error—nondisjunction of chromosome 21 during maternal or paternal meiosis—the occurrence of the disorder is *not* expected to be inherited. Nevertheless, Down syndrome occasionally runs in families. These instances, referred to as familial Down syndrome, involve a translocation of chromosome 21, another type of chromosomal aberration, which we will discuss later in the chapter.

Human Aneuploidy

Besides Down syndrome, only two human trisomies, and no autosomal monosomies, survive to term: **Patau** and **Edwards syndromes** (47,13+ and 47,18+, respectively). Even so, these individuals manifest severe malformations and early lethality. **Figure 6–4** illustrates the abnormal karyotype and the many defects characterizing Patau infants.



Mental retardation
Growth failure
Low-set, deformed ears
Deafness
Atrial septal defect
Ventricular septal defect
Abnormal polymorphonuclear granulocytes
Microcephaly
Cleft lip and palate
Polydactyly
Deformed finger nails
Kidney cysts
Double ureter
Umbilical hernia
Developmental uterine abnormalities
Cryptorchidism

FIGURE 6–4 The karyotype and phenotypic description of an infant with Patau syndrome, where three members of the D-group chromosome 13 are present, creating the 47,13+ condition.

The above observation leads us to ask whether many other aneuploid conditions arise but that the affected fetuses do not survive to term. That this is the case has been confirmed by karyotypic analysis of spontaneously aborted fetuses. These studies reveal two striking statistics: (1) Approximately 20 percent of all conceptions terminate in spontaneous abortion (some estimates are considerably higher); and (2) about 30 percent of all spontaneously aborted fetuses demonstrate some form of chromosomal imbalance. This suggests that at least 6 percent (0.20×0.30) of conceptions contain an abnormal chromosome complement. A large percentage of fetuses demonstrating chromosomal abnormalities are aneuploids.

An extensive review of this subject by David H. Carr has revealed that a significant percentage of aborted fetuses are trisomic for one of the chromosome groups. Trisomies for every human chromosome have been recovered. Interestingly, the monosomy with the highest incidence among abortuses is the 45,X condition, which produces an infant with Turner syndrome if the fetus survives to term. Autosomal monosomies are seldom found, however, even though nondisjunction should produce $n - 1$ gametes with a frequency equal to $n + 1$ gametes. This finding suggests that gametes lacking a single chromosome are functionally impaired to a serious degree or that the embryo dies so early in its development that recovery occurs infrequently. We discussed the potential causes of monosomic lethality earlier in this chapter. Carr's study also found various forms of polyploidy and other miscellaneous chromosomal anomalies.

These observations support the hypothesis that normal embryonic development requires a precise diploid complement of chromosomes to maintain the delicate equilibrium in the expression of genetic information. The prenatal mortality of most aneuploids provides a barrier against the introduction of these genetic anomalies into the human population.

ESSENTIAL POINT

Studies of monosomic and trisomic disorders are increasing our understanding of the delicate genetic balance that is essential for normal development. ■

6.3 Polyploidy, in Which More Than Two Haploid Sets of Chromosomes Are Present, Is Prevalent in Plants

The term *polyploidy* describes instances in which more than two multiples of the haploid chromosome set are found. The naming of polyploids is based on the number

of sets of chromosomes found: A triploid has 3*n* chromosomes; a tetraploid has 4*n;* a pentaploid, 5*n;* and so forth (Table 6.1). Several general statements can be made about polyploidy. This condition is relatively infrequent in many animal species but is well known in lizards, amphibians, and fish, and is much more common in plant species. Usually, odd numbers of chromosome sets are not reliably maintained from generation to generation because a polyploid organism with an uneven number of homologs often does not produce genetically balanced gametes. For this reason, triploids, pentaploids, and so on, are not usually found in plant species that depend solely on sexual reproduction for propagation.

Polyploidy originates in two ways: (1) The addition of one or more extra sets of chromosomes, identical to the normal haploid complement of the same species, resulting in **autopolyploidy;** or (2) the combination of chromosome sets from different species occurring as a consequence of hybridization, resulting in **allopolyploidy** (from the Greek word *allo,* meaning "other" or "different"). The distinction between auto- and allopolyploidy is based on the genetic origin of the extra chromosome sets, as shown in **Figure 6–5**.

In our discussion of polyploidy, we use the following symbols to clarify the origin of additional chromosome sets. For example, if *A* represents the haploid set of chromosomes of any organism, then

$$A = a_1 + a_2 + a_3 + a_4 + \cdots + a_n$$

where a_1, a_2, and so on, are individual chromosomes and *n* is the haploid number. A normal diploid organism is represented simply as *AA*.

Autopolyploidy

In autopolyploidy, each additional set of chromosomes is identical to the parent species. Therefore, triploids are represented as *AAA,* tetraploids are *AAAA,* and so forth.

Autotriploids arise in several ways. A failure of all chromosomes to segregate during meiotic divisions can produce a diploid gamete. If such a gamete is fertilized by a haploid gamete, a zygote with three sets of chromosomes is produced. Or, rarely, two sperm may fertilize an ovum, resulting in a triploid zygote. Triploids are also produced under experimental conditions by crossing diploids with tetraploids. Diploid organisms produce gametes with *n* chromosomes, while tetraploids produce 2*n* gametes. Upon fertilization, the desired triploid is produced.

Because they have an even number of chromosomes, **autotetraploids** (4*n*) are theoretically more likely to be found in nature than are autotriploids. Unlike triploids, which often produce genetically unbalanced gametes with odd numbers of chromosomes, tetraploids are more likely to produce balanced gametes when involved in sexual reproduction.

How polyploidy arises naturally is of great interest to geneticists. In theory, if chromosomes have replicated, but the parent cell never divides and instead reenters interphase, the chromosome number will be doubled. That this very likely occurs is supported by the observation that tetraploid cells can be produced experimentally from diploid cells. This is accomplished by applying cold or heat shock to meiotic cells or by applying colchicine to somatic cells undergoing mitosis. Colchicine, an alkaloid derived from the autumn crocus, interferes with spindle formation, and thus replicated chromosomes cannot separate at anaphase and do not migrate to the poles. When colchicine is removed, the cell can reenter interphase. When the paired sister chromatids separate and uncoil, the nucleus contains twice the diploid number of chromosomes and is therefore 4*n.* This process is shown in **Figure 6–6** .

In general, autopolyploids are larger than their diploid relatives. This increase seems to be due to larger cell size rather than greater cell number. Although autopolyploids do not contain new or unique information compared with their diploid relatives, the flower and fruit of plants are often increased in size, making such varieties of greater horticultural or commercial value. Economically important triploid plants include several potato species of the genus *Solanum,* Winesap apples, commercial bananas, seedless watermelons,

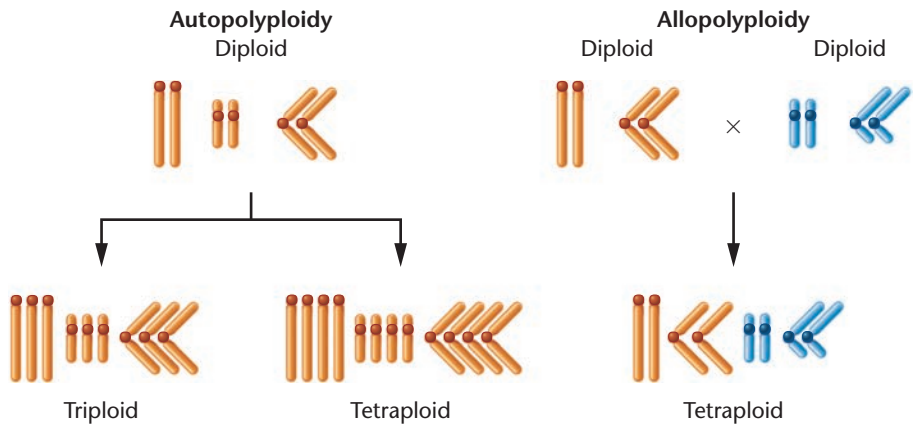


FIGURE 6–5 Contrasting chromosome origins of an autopolyploid versus an allopolyploid karyotype.

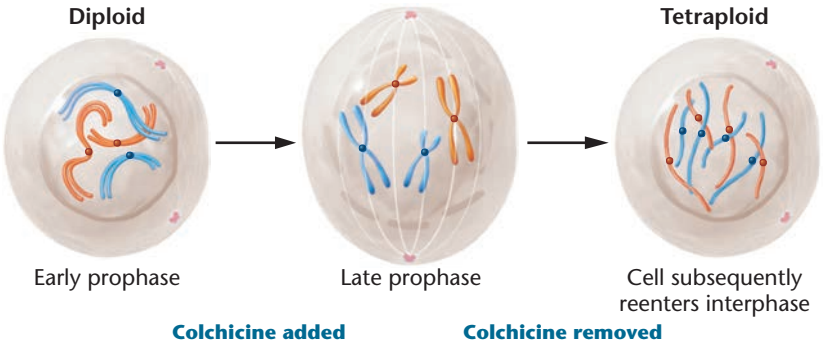


FIGURE 6–6 The potential involvement of colchicine in doubling the chromosome number. Two pairs of homologous chromosomes are shown. While each chromosome had replicated its DNA earlier during interphase, the chromosomes do not appear as double structures until late prophase. When anaphase fails to occur normally, the chromosome number doubles if the cell reenters interphase.

and the cultivated tiger lily *Lilium tigrinum*. These plants are propagated asexually. Diploid bananas contain hard seeds, but the commercial, triploid, "seedless" variety has edible seeds. Tetraploid alfalfa, coffee, peanuts, and McIntosh apples are also of economic value because they are either larger or grow more vigorously than do their diploid or triploid counterparts. Many of the most popular varieties of hosta plant are tetraploid. In each case, leaves are thicker and larger, the foliage is more vivid, and the plant grows more vigorously. The commercial strawberry is an octoploid.

How cells with increased ploidy values express different phenotypes from their diploid counterparts has been investigated. Gerald Fink and his colleagues created strains of the yeast *Saccharomyces cerevisiae* with one, two, three, or four copies of the genome and then examined the expression levels of all genes during the cell cycle. Using the stringent standards of at least a tenfold increase or decrease of gene expression, Fink and coworkers identified numerous cases where, as ploidy increased, gene expression either increased or decreased at least tenfold. Among these cases are two genes that encode **G1 cyclins,** which are repressed when ploidy increases. G1 cyclins facilitate the cell's movement through G1 of the cell cycle, which is thus delayed when expression of these genes is repressed. The polyploid cell stays in the G1 phase longer and, on average, grows to a larger size before it moves beyond the G1 stage of the cell cycle, providing a clue as to how other polyploids demonstrate increased cell size.

Allopolyploidy

Polyploidy can also result from hybridizing two closely related species. If a haploid ovum from a species with chromosome sets *AA* is fertilized by sperm from a species with sets *BB*, the resulting hybrid is *AB,* where $A = a_1, a_2, a_3, \ldots a_n$ and $B = b_1, b_2, b_3, \ldots b_n$. The hybrid organism may be sterile because of its inability to produce viable gametes. Most often, this occurs when some or all of the *a* and *b* chromosomes are not homologous and therefore cannot synapse in meiosis. As a result, unbalanced genetic conditions result. If, however, the new *AB* genetic combination undergoes a natural or an induced chromosomal doubling, two copies of all *a* chromosomes and two copies of all *b* chromosomes will be present, and they will pair during meiosis. As a result, a fertile *AABB* tetraploid is produced. These events are shown in Figure 6–7 . Since this polyploid contains the equivalent of four haploid genomes derived from separate species, such an organism is called an **allotetraploid.** When both original species are known, an equivalent term, **amphidiploid,** is preferred in describing the allotetraploid.

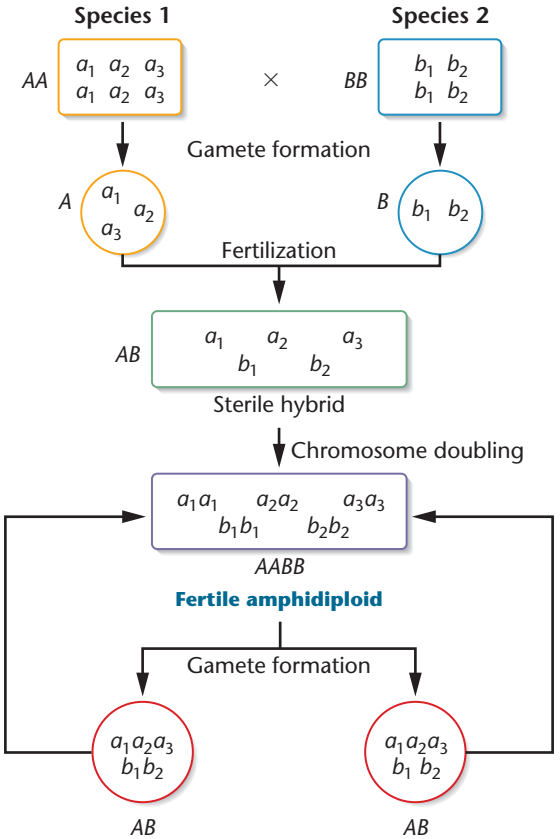


FIGURE 6–7 The origin and propagation of an amphidiploid. Species 1 contains genome *A* consisting of three distinct chromosomes, a_1, a_2, and a_3. Species 2 contains genome *B* consisting of two distinct chromosomes, b_1 and b_2. Following fertilization between members of the two species and chromosome doubling, a fertile amphidiploid containing two complete diploid genomes (*AABB*) is formed.

FIGURE 6–8 The pods of the amphidiploid form of *Gossypium,* the cultivated American cotton plant.

Amphidiploid plants are often found in nature. Their reproductive success is based on their potential for forming balanced gametes. Since two homologs of each specific chromosome are present, meiosis occurs normally (Figure 6–7) and fertilization successfully propagates the plant sexually. This discussion assumes the simplest situation, where none of the chromosomes in set *A* are homologous to those in set *B.* In amphidiploids formed from closely related species, some homology between *a* and *b* chromosomes is likely. Allopolyploids are rare in most animals because mating behavior is most often species-specific, and thus the initial step in hybridization is unlikely to occur.

A classical example of amphidiploidy in plants is the cultivated species of American cotton, *Gossypium* (**Figure 6–8**). This species has 26 pairs of chromosomes: 13 are large and 13 are much smaller. When it was discovered that Old World cotton had only 13 pairs of large chromosomes, allopolyploidy was suspected. After an examination of wild American cotton revealed 13 pairs of small chromosomes, this speculation was strengthened. J. O. Beasley reconstructed the origin of cultivated cotton experimentally by crossing the Old World strain with the wild American strain and then treating the hybrid with colchicine to double the chromosome number. The result of these treatments was a fertile amphidiploid variety of cotton. It contained 26 pairs of chromosomes as well as characteristics similar to the cultivated variety.

Amphidiploids often exhibit traits of both parental species. An interesting example involves the grasses wheat and rye. Wheat (genus *Triticum*) has a basic haploid genome of seven chromosomes. In addition to normal diploids ($2n = 14$), cultivated autopolyploids exist, including tetraploid ($4n = 28$) and hexaploid ($6n = 42$) species. Rye (genus *Secale*) also has a genome consisting of seven chromosomes. The only cultivated species is the diploid plant ($2n = 14$).

Using the technique outlined in Figure 6–7, geneticists have produced various hybrids. When tetraploid wheat is crossed with diploid rye and the F_1 plants are treated with colchicine, a hexaploid variety ($6n = 42$) is obtained; the hybrid, designated *Triticale,* represents a new genus. Other *Triticale* varieties have been created. These hybrid plants demonstrate characteristics of both wheat and rye. For example, they combine the high-protein content of wheat with rye's high content of the amino acid lysine, which is low in wheat and thus is a limiting nutritional factor. Wheat is considered to be a high-yielding grain, whereas rye is noted for its versatility of growth in unfavorable environments. *Triticale* species that combine both traits have the potential of significantly increasing grain production. This and similar programs designed to improve crops through hybridization have long been under way in several developing countries.

NOW SOLVE THIS

6–2 When two plants belonging to the same genus but different species are crossed, the F_1 hybrid is viable and has more ornate flowers. Unfortunately, this hybrid is sterile and can only be propagated by vegetative cuttings. Explain the sterility of the hybrid and what would have to occur for the sterility of this hybrid to be reversed.

■ **HINT:** *This problem involves an understanding of allopolyploid plants. The key to its solution is to focus on the origin and composition of the chromosomes in the F_1 and how they might be manipulated.*

ESSENTIAL POINT

When complete sets of chromosomes are added to the diploid genome, these sets can have an identical or a diverse genetic origin, creating either autopolyploidy or allopolyploidy, respectively. ■

6.4 Variation Occurs in the Composition and Arrangement of Chromosomes

The second general class of chromosome aberrations includes changes that delete, add, or rearrange substantial portions of one or more chromosomes. Included in this broad category are deletions and duplications of genes or part of a chromosome and rearrangements of genetic material in which a chromosome segment is inverted, exchanged with a segment of a nonhomologous chromosome, or merely transferred to another chromosome. Exchanges and transfers are called translocations, in

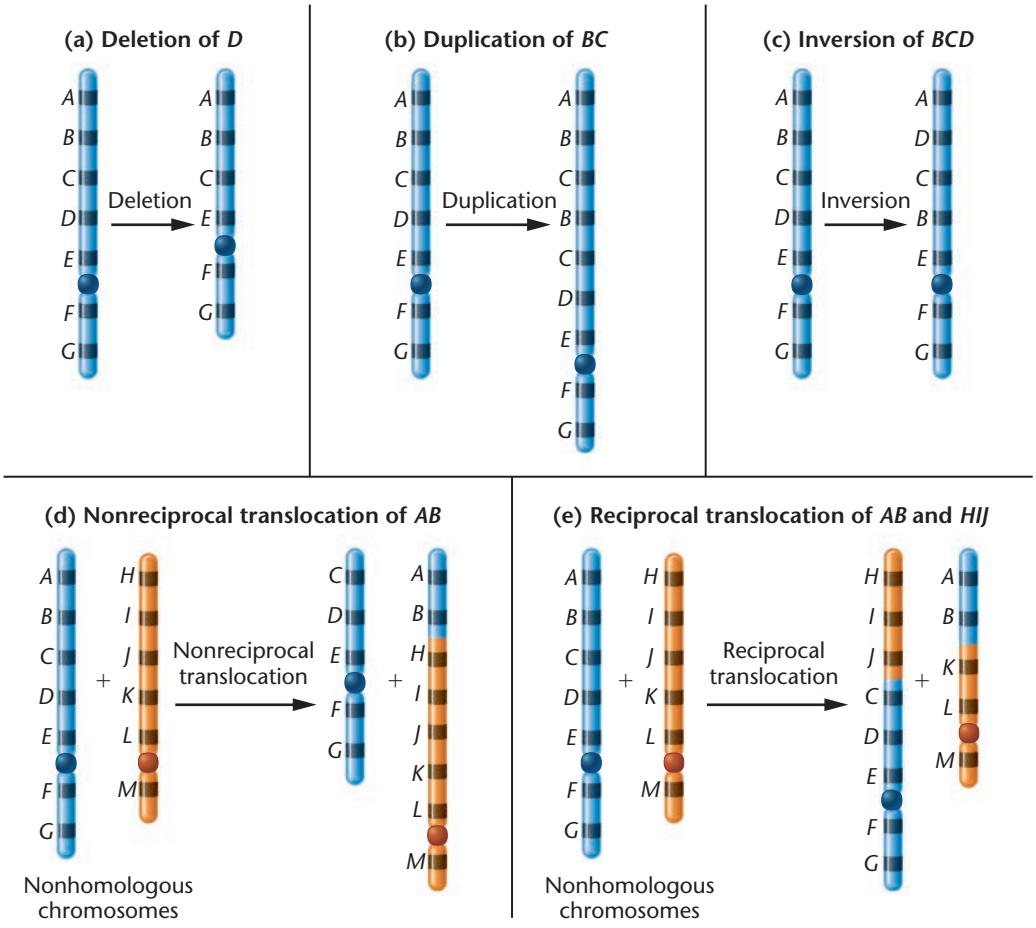


FIGURE 6–9 An overview of the five different types of gain, loss, or rearrangement of chromosome segments.

which the locations of genes are altered within the genome. These types of chromosome alterations are illustrated in **Figure 6–9**.

In most instances, these structural changes are due to one or more breaks along the axis of a chromosome, followed by either the loss or rearrangement of genetic material. Chromosomes can break spontaneously, but the rate of breakage may increase in cells exposed to chemicals or radiation. The ends produced at points of breakage are "sticky" and can rejoin other broken ends. If breakage and rejoining do not reestablish the original relationship and if the alteration occurs in germ plasm, the gametes will contain the structural rearrangement, which is heritable.

If the aberration is found in one homolog but not the other, the individual is said to be *heterozygous for the aberration.* In such cases, unusual but characteristic pairing configurations are formed during meiotic synapsis. These patterns are useful in identifying the type of change that has occurred. If no loss or gain of genetic material occurs, individuals bearing the aberration "heterozygously" are likely to be unaffected phenotypically. However, the unusual pairing arrangements often lead to gametes that are duplicated or deficient for some chromosomal regions. When this occurs, the offspring of "carriers" of certain aberrations have an increased probability of demonstrating phenotypic changes.

6.5 A Deletion Is a Missing Region of a Chromosome

When a chromosome breaks in one or more places and a portion of it is lost, the missing piece is called a **deletion** (or a **deficiency**). The deletion can occur either near one end or within the interior of the chromosome. These are **terminal** and **intercalary deletions,** respectively [**Figure 6–10(a) and (b)**]. The portion of the chromosome that retains the centromere region is usually maintained when the cell divides, whereas the segment without the centromere is eventually lost in progeny cells following mitosis or meiosis. For synapsis to occur between a chromosome with a large intercalary deletion and a normal homolog, the unpaired region of the normal homolog must "buckle out" into a **deletion,** or **compensation, loop** [**Figure 6–10(c)**].

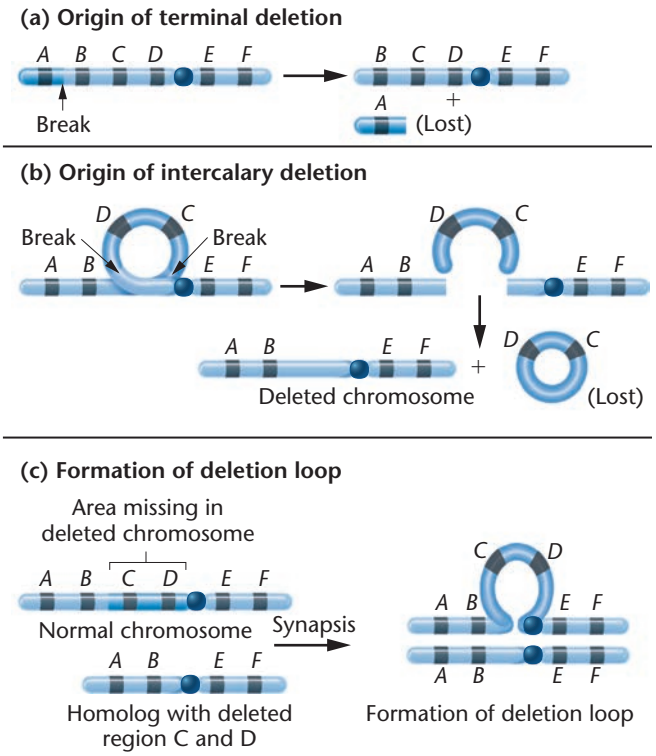


FIGURE 6–10 Origins of (a) a terminal and (b) an intercalary deletion. In (c), pairing occurs between a normal chromosome and one with an intercalary deletion by looping out the undeleted portion to form a deletion (or compensation) loop.

If only a small part of a chromosome is deleted, the organism might survive. However, a deletion of a portion of a chromosome need not be very great before the effects become severe. We see an example of this in the following discussion of the cri du chat syndrome in humans. If even more genetic information is lost as a result of a deletion, the aberration is often lethal, in which case the chromosome mutation never becomes available for study.

Cri du Chat Syndrome in Humans

In humans, the **cri du chat syndrome** results from the deletion of a small terminal portion of chromosome 5. It might be considered a case of *partial monosomy,* but since the region that is missing is so small, it is better referred to as a **segmental deletion.** This syndrome was first reported by Jérôme LeJeune in 1963, when he described the clinical symptoms, including an eerie cry similar to the meowing of a cat, after which the syndrome is named. This syndrome is associated with the loss of a small, variable part of the short arm of chromosome 5 (**Figure 6–11**). Thus, the genetic constitution may be designated as 46,5p–, meaning that the individual has all 46 chromosomes but that some or all of the p arm (the petite, or short, arm) of one member of the chromosome 5 pair is missing.

Infants with this syndrome may exhibit anatomic malformations, including gastrointestinal and cardiac complications, and they are often mentally retarded. Abnormal development of the glottis and larynx (leading to the characteristic cry) is typical of this syndrome.

Since 1963, hundreds of cases of cri du chat syndrome have been reported worldwide. An incidence of 1 in 25,000–50,000 live births has been estimated. Most often, the condition is not inherited but instead results from the sporadic loss of chromosomal material in gametes. The length of the short arm that is deleted varies somewhat; longer deletions appear to have a greater impact on the physical, psychomotor, and mental skill levels of those children who survive. Although the effects of the syndrome are severe, most individuals achieve motor and language skills and may be home-cared. In 2004, it was reported that the portion of the chromosome that is missing contains the *TERT* gene, which encodes telomerase reverse transcriptase, an enzyme essential for the maintenance of telomeres during DNA replication (see Chapter 10). Whether the absence of this gene on one homolog is related to the multiple phenotypes of cri du chat infants is still unknown.

FIGURE 6–11 A representative karyotype and a photograph of a child exhibiting cri du chat syndrome (46,5p–). In the karyotype, the arrow identifies the absence of a small piece of the short arm of one member of the chromosome 5 homologs.

6.6 A Duplication Is a Repeated Segment of a Chromosome

When any part of the genetic material—a single locus or a large piece of a chromosome—is present more than once in the genome, it is called a **duplication.** As in deletions, pairing in heterozygotes can produce a compensation loop. Duplications may arise as the result of unequal crossing over between synapsed chromosomes during meiosis (**Figure 6–12**) or through a replication error prior to meiosis. In the former case, both a duplication and a deletion are produced.

We consider three interesting aspects of duplications. First, they may result in gene redundancy. Second, as with deletions, duplications may produce phenotypic variation. Third, according to one convincing theory, duplications have also been an important source of genetic variability during evolution.

Gene Redundancy—Ribosomal RNA Genes

Although many gene products are not needed in every cell of an organism, other gene products are known to be essential components of all cells. For example, ribosomal RNA must be present in abundance to support protein synthesis. The more metabolically active a cell is, the higher the demand for this molecule. We might hypothesize that a single copy of the gene encoding rRNA is inadequate in many cells. Studies using the technique of molecular hybridization, which enables us to determine the percentage of the genome that codes for specific RNA sequences, show that our hypothesis is correct. Indeed, multiple copies of genes code for rRNA. Such DNA is called **rDNA,** and the general phenomenon is referred to as **gene redundancy.** For example, in the common intestinal bacterium *Escherichia coli* (*E. coli*), about 0.7 percent of the haploid genome consists of rDNA—the equivalent of seven copies of the gene. In *Drosophila melanogaster,* 0.3 percent of the haploid genome, equivalent to 130 gene copies, consists of rDNA. Although the presence of multiple copies of the same gene is not restricted to those coding for rRNA, we will focus on them in this section.

In some cells, particularly oocytes, even the normal amplification of rDNA is insufficient to provide adequate amounts of rRNA needed to construct ribosomes. For example, in the amphibian *Xenopus laevis,* 400 copies of rDNA are present per haploid genome. These genes are all found in a single area of the chromosome known as the **nucleolar organizer region (NOR).** In *Xenopus* oocytes, the NOR is selectively replicated to further increase rDNA copies, and each new set of genes is released from its template. Each set forms a small nucleolus, and as many as 1500 of these "micronucleoli" have been observed in a single oocyte. If we multiply the number of micronucleoli (1500) by the number of gene copies in each NOR (400), we see that amplification in *Xenopus* oocytes can result in over half a million gene copies! If each copy is transcribed only 20 times during the maturation of the oocyte, in theory, sufficient copies of rRNA are produced to result in well over 12 million ribosomes.

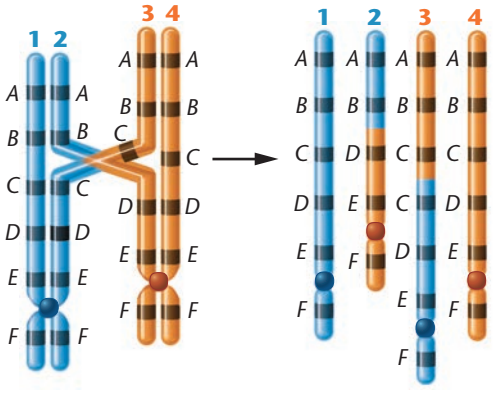


FIGURE 6–12 The origin of duplicated and deficient regions of chromosomes as a result of unequal crossing over. The tetrad on the left is mispaired during synapsis. A single crossover between chromatids 2 and 3 results in the deficient (chromosome 2) and duplicated (chromosome 3) chromosomal regions shown on the right. The two chromosomes uninvolved in the crossover event remain normal in gene sequence and content.

The *Bar* Mutation in *Drosophila*

Duplications can cause phenotypic variation that might at first appear to be caused by a simple gene mutation. The *Bar*-eye phenotype in *Drosophila* (**Figure 6–13**) is a classic example. Instead of the normal oval-eye shape, *Bar*-eyed flies have narrow, slitlike eyes. This phenotype is inherited in the same way as a dominant X-linked mutation.

In the early 1920s, Alfred H. Sturtevant and Thomas H. Morgan discovered and investigated this "mutation." Normal wild-type females (B^+/B^+) have about 800 facets in each eye. Heterozygous females (B/B^+) have about 350 facets, while homozygous females (B/B) average only about 70 facets. Females were occasionally recovered with even fewer facets and were designated as *double Bar* (B^D/B^+)

About 10 years later, Calvin Bridges and Herman J. Muller compared the polytene X chromosome banding pattern of the *Bar* fly with that of the wild-type fly. These chromosomes contain specific banding patterns that have been well categorized into regions. Their studies revealed that one copy of the region designated as 16A is present on both X chromosomes of wild-type flies but that this region was duplicated in *Bar* flies and triplicated in *double Bar* flies.

B^+/B^+ B/B^+ B/B

FIGURE 6–13 *Bar*-eye phenotypes in contrast to the wild-type eye in *Drosophila* (shown in the left panel).

These observations provided evidence that the *Bar* phenotype is not the result of a simple chemical change in the gene but is instead a duplication.

The Role of Gene Duplication in Evolution

During the study of evolution, it is intriguing to speculate on the possible mechanisms of genetic variation. The origin of unique gene products present in more recently evolved organisms but absent in ancestral forms is a topic of particular interest. In other words, how do "new" genes arise?

In 1970, Susumo Ohno published a provocative monograph, *Evolution by Gene Duplication,* in which he suggested that gene duplication is essential to the origin of new genes during evolution. Ohno's thesis is based on the supposition that the gene products of many genes, present as only a single copy in the genome, are indispensable to the survival of members of any species during evolution. Therefore, unique genes are not free to accumulate mutations sufficient to alter their primary function and give rise to new genes.

However, if an essential gene is duplicated in the germ line, major mutational changes in this extra copy will be tolerated in future generations because the original gene provides the genetic information for its essential function. The duplicated copy will be free to acquire many mutational changes over extended periods of time. Over short intervals, the new genetic information may be of no practical advantage. However, over long evolutionary periods, the duplicated gene may change sufficiently so that its product assumes a divergent role in the cell. The new function may impart an "adaptive" advantage to organisms, enhancing their fitness. Ohno has outlined a mechanism through which sustained genetic variability may have originated.

Ohno's thesis is supported by the discovery of genes that have a substantial amount of their organization and DNA sequence in common, but whose gene products are distinct. For example, trypsin and chymotrypsin fit this description, as do myoglobin and the various forms of hemoglobin. The DNA sequence is so similar (homologous) in each case that we may conclude that members of each pair of genes arose from a common ancestral gene through duplication. During evolution, the related genes diverged sufficiently that their products became unique.

Other support includes the presence of **gene families**—groups of contiguous genes whose products perform the same, or very similar functions. Again, members of a family show DNA sequence homology sufficient to conclude that they share a common origin and arose through the process of gene duplication. One of the most interesting supporting examples is the case of the *SRGAP2* gene in primates. This gene is known to be involved in the development of the brain. Humans have at least four similar copies of the gene, while all nonhuman primates have only a single copy. Several duplication events can be traced back to 3.4 million years ago, to 2.4 million years ago, and finally to 1 million years ago, resulting in distinct forms of *SRGAP2* labeled A–D. These evolutionary periods coincide with the emergence of the human lineage in primates. The function of these genes has now been related to the regulation and formation of dendritic spines in the brain, which is believed to contribute to the evolution of expanded brain function in humans, including the development of language and social cognition.

Other examples of gene families arising from duplication during evolution include the various types of human hemoglobin polypeptide chains, as well as the immunologically important T-cell receptors and antigens encoded by the major histocompatibility complex.

Duplications at the Molecular Level: Copy Number Variants (CNVs)

As we entered the era of genomics and became capable of sequencing entire genomes (see Chapter 17), we quickly realized that duplications of *portions of genes,* most often involving thousands of base pairs, occur on a regular basis. When individuals in the same species are compared, the number of copies of any given duplicated sequence is found to vary—sometimes there are larger and sometimes smaller numbers of copies. These variations, because they represent *quantitative differences in the number of large DNA sequences,* are termed **copy number variants (CNVs)** and are found in both coding and noncoding regions of the genome.

CNVs are of major interest in genetics because they are now believed to play crucial roles in the expression of many of our individual traits. In 2004, two research groups independently described the presence of CNVs in the genomes of healthy individuals with no known genetic disorders. CNVs were initially defined as regions of DNA at least 1 kb in length (1000 base pairs) that display at least 90 percent sequence identity. This initial study revealed 50 CNV loci, and in 2005, 300 additional sites were identified. In 2010 the number of CNV sites, when scaled down to sequences of DNA of at least 500 base pairs, was estimated at over 10,000 regions, representing a substantial proportion of the total genetic variability within humans. Future studies will no doubt uncover many more unique CNVs in the human genome.

Current studies have focused on finding associations with human diseases. CNVs appear to have both positive and negative associations with many diseases in which the genetic basis is not yet fully understood. For example, an association has been reported between CNVs and autism, the well-known neurodevelopmental disorder that impairs communication, behavior, and social interaction. Interestingly, a mutant CNV site has been found to appear *de novo* (anew) in 10 percent of so-called sporadic cases of autism, where unaffected parents lack the CNV mutation. This is in contrast to only 2 percent of affected individuals where the disease appears to be familial (that is, to run in the family).

Often, entire gene sequences are duplicated and impact individuals. For example, a higher than average copy number of the gene *CCL3L1* imparts an HIV-suppressive effect during viral infection, diminishing the progression to AIDS. Another finding has associated specific mutant CNV sites with certain subset populations of individuals with lung cancer—the greater number of copies of the *EGFR* (*Epidermal Growth Factor Receptor*) gene, the more responsive are patients with non–small-cell lung cancer to treatment. Finally, the greater the reduction in the copy number of the gene designated *DEFB,* the greater the risk of developing Crohn disease, a condition affecting the colon. Relevant to this chapter, these findings reveal that duplications and deletions are no longer restricted to textbook examples of these chromosomal mutations. We will return to this interesting topic later in the text (see Chapter 18), when genomics is discussed in detail.

ESSENTIAL POINT

Deletions or duplications of segments of a gene or a chromosome may be the source of mutant phenotypes such as cri du chat syndrome in humans and *Bar* eyes in *Drosophila,* while duplications can be particularly important as a source of redundant or new genes. ■

6.7 Inversions Rearrange the Linear Gene Sequence

The **inversion,** another class of structural variation, is a type of chromosomal aberration in which a segment of a chromosome is turned around 180 degrees within a chromosome. An inversion does not involve a loss of genetic information but simply rearranges the linear gene sequence. An inversion requires breaks at two points along the length of the chromosome and subsequent reinsertion of the inverted segment. **Figure 6–14** illustrates how an inversion might arise. By forming a chromosomal loop prior to breakage, the newly created "sticky ends" are brought close together and rejoined.

The inverted segment may be short or quite long and may or may not include the centromere. If the centromere is not part of the rearranged chromosome segment, it is a **paracentric inversion,** which is the type shown in Figure 6–14. If the centromere is part of the inverted segment, it is described as a **pericentric inversion.**

Consequences of Inversions during Gamete Formation

If only one member of a homologous pair of chromosomes has an inverted segment, normal *linear synapsis* during meiosis is not possible. Organisms with one inverted chromosome and one noninverted homolog are called **inversion heterozygotes.** Pairing between two such chromosomes in meiosis is accomplished only if they form an **inversion loop** (**Figure 6–15**).

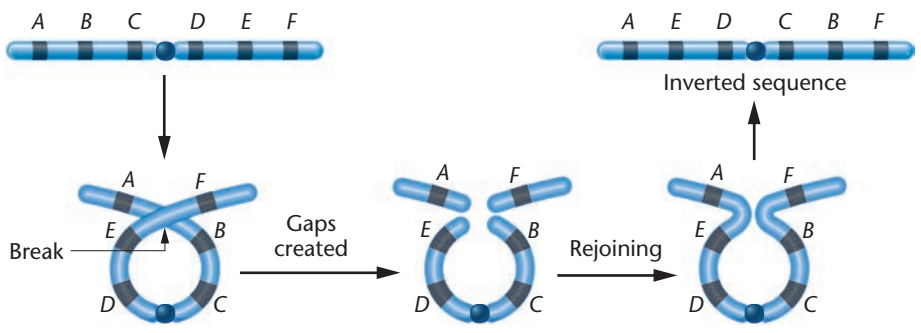


FIGURE 6–14 One possible origin of a paracentric inversion.

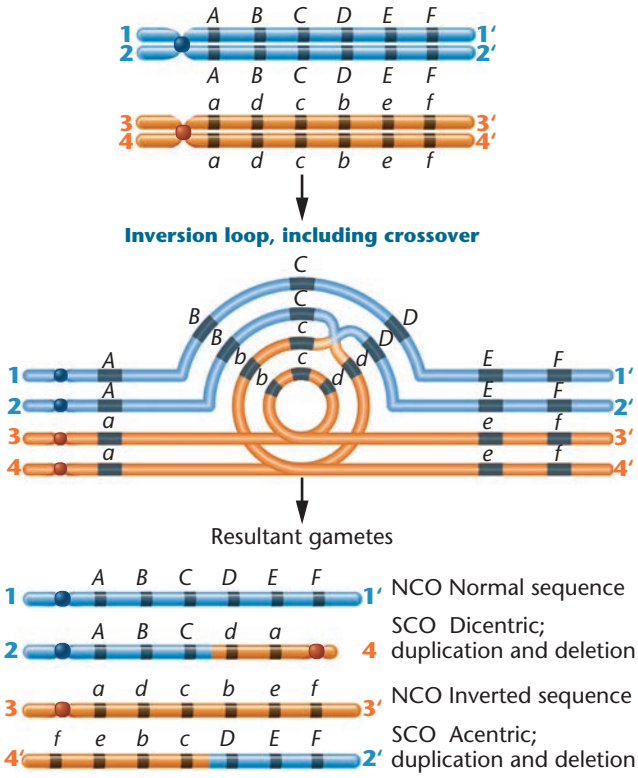


FIGURE 6–15 The effects of a single crossover (SCO) within an inversion loop in a paracentric inversion heterozygote, where two altered chromosomes are produced, one acentric and one dicentric. Both chromosomes also contain duplicated and deficient regions.

If crossing over does not occur within the inverted segment of the inversion loop, the homologs will segregate, which results in two normal and two inverted chromatids that are distributed into gametes. However, if crossing over does occur within the inversion loop, abnormal chromatids are produced. The effect of a single crossover (SCO) event within a paracentric inversion is diagrammed in Figure 6–15.

In any meiotic tetrad, a single crossover between nonsister chromatids produces two parental chromatids and two recombinant chromatids. When the crossover occurs within a paracentric inversion, however, one recombinant **dicentric chromatid** (two centromeres) and one recombinant **acentric chromatid** (lacking a centromere) are produced. Both contain duplications and deletions of chromosome segments as well. During anaphase, an acentric chromatid moves randomly to one pole or the other or may be lost, while a dicentric chromatid is pulled in two directions. This polarized movement produces *dicentric bridges* that are cytologically recognizable. A dicentric chromatid usually breaks at some point so that part of the chromatid goes into one gamete and part into another gamete during the reduction divisions. Therefore, gametes containing either recombinant chromatid are deficient in genetic material. In animals, when such a gamete participates in fertilization, the zygote most often develops abnormally, inviable embryos are produced, and lethality is the final result. In plants, gametes receiving such aberrant chromatids fail to develop normally, leading to aborted pollen or ovules. Thus, fertilization is not achieved.

Because offspring bearing crossover gametes are inviable and not recovered, it *appears* as if the inversion suppresses crossing over. Actually, in inversion heterozygotes, the inversion has the effect of *suppressing the recovery of crossover products* when chromosome exchange occurs within the inverted region. Moreover, up to one-half of the viable gametes have the inverted chromosome, and the inversion will be perpetuated within the species. The cycle will be repeated continuously during meiosis in future generations.

Evolutionary Advantages of Inversions

Because recovery of crossover products is suppressed in inversion heterozygotes, groups of specific alleles at adjacent loci within inversions may be preserved from generation to generation. If the alleles of the involved genes confer a survival advantage on the organisms maintaining them, the inversion is beneficial to the evolutionary survival of the species. For example, if a set of alleles *ABcDef* is more adaptive than sets *AbCdeF* or *abcdEF*, effective gametes will contain this favorable set of genes, undisrupted by crossing over.

In laboratory studies, the same principle is applied using **balancer chromosomes,** which contain inversions. When an organism is heterozygous for a balancer chromosome, desired sequences of alleles are preserved during experimental work.

NOW SOLVE THIS

6–3 What is the effect of a rare double crossover within a chromosome segment that is heterozygous for a paracentric inversion?

■ **HINT:** *This problem involves an understanding of how homologs synapse in the presence of a heterozygous paracentric inversion. The key to its solution is to draw out the tetrad and follow the chromatids undergoing a double crossover.*

6.8 Translocations Alter the Location of Chromosomal Segments in the Genome

Translocation, as the name implies, is the movement of a chromosomal segment to a new location in the genome. Reciprocal translocation, for example, involves the exchange of segments between two nonhomologous chromosomes. The least complex way for this event to occur is for two nonhomologous chromosome arms to come close to

each other so that an exchange is facilitated. **Figure 6–16(a)** shows a simple reciprocal translocation in which only two breaks are required. If the exchange includes internal chromosome segments, four breaks are required, two on each chromosome.

The genetic consequences of reciprocal translocations are, in several instances, similar to those of inversions. For example, genetic information is not lost or gained. Rather, there is only a rearrangement of genetic material. The presence of a translocation does not, therefore, directly alter the viability of individuals bearing it.

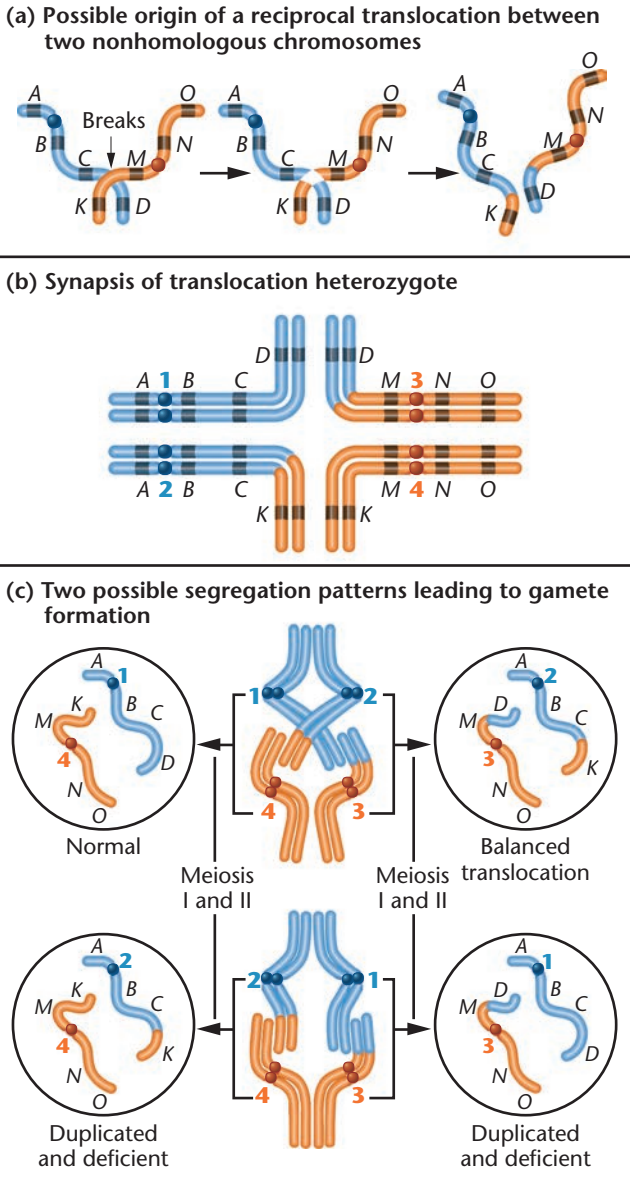


FIGURE 6–16 (a) Possible origin of a reciprocal translocation. (b) Synaptic configuration formed during meiosis in an individual that is heterozygous for the translocation. (c) Two possible segregation patterns, one of which leads to a normal and a balanced gamete (called alternate segregation) and one that leads to gametes containing duplications and deficiencies (called adjacent segregation).

Homologs that are heterozygous for a reciprocal translocation undergo unorthodox synapsis during meiosis. As shown in **Figure 6–16(b)**, pairing results in a crosslike configuration. As with inversions, genetically unbalanced gametes are also produced as a result of this unusual alignment during meiosis. In the case of translocations, however, aberrant gametes are not necessarily the result of crossing over. To see how unbalanced gametes are produced, focus on the homologous centromeres in Figure 6–16(b) and **Figure 6–16(c)**. According to the principle of independent assortment, the chromosome containing centromere 1 migrates randomly toward one pole of the spindle during the first meiotic anaphase; it travels along with *either* the chromosome having centromere 3 *or* the chromosome having centromere 4. The chromosome with centromere 2 moves to the other pole along with the chromosome containing *either* centromere 3 *or* centromere 4. This results in four potential meiotic products. The 1,4 combination contains chromosomes that are not involved in the translocation. The 2,3 combination, however, contains translocated chromosomes. These contain a complete complement of genetic information and are balanced. The other two potential products, the 1,3 and 2,4 combinations, contain chromosomes displaying duplicated and deleted segments. To simplify matters, crossover exchanges are ignored here.

When incorporated into gametes, the resultant meiotic products are genetically unbalanced. If they participate in fertilization, lethality often results. As few as 50 percent of the progeny of parents that are heterozygous for a reciprocal translocation survive. This condition, called **semisterility,** has an impact on the reproductive fitness of organisms, thus playing a role in evolution. Furthermore, in humans, such an unbalanced condition results in partial monosomy or trisomy, leading to a variety of birth defects.

Translocations in Humans: Familial Down Syndrome

Research conducted since 1959 has revealed numerous translocations in members of the human population. One common type of translocation involves breaks at the extreme ends of the short arms of two nonhomologous acrocentric chromosomes. These small segments are lost, and the larger segments fuse at their centromeric region. This type of translocation produces a new, large submetacentric or metacentric chromosome, often called a **Robertsonian translocation.**

One such translocation accounts for cases in which Down syndrome is familial (inherited). Earlier in this chapter, we pointed out that most instances of Down syndrome are due to trisomy 21. This chromosome composition results from nondisjunction during meiosis in one

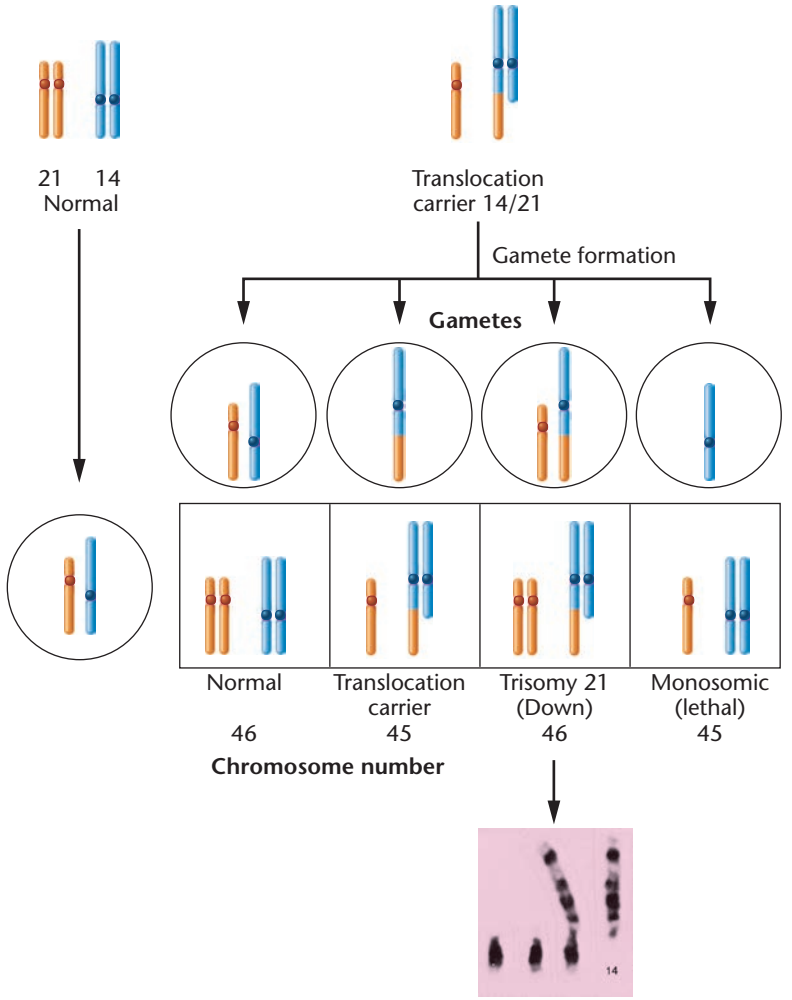


FIGURE 6–17 Chromosomal involvement and translocation in familial Down syndrome. The photograph shows the relevant chromosomes from a trisomy 21 offspring produced by a translocation carrier parent.

parent. Trisomy accounts for over 95 percent of all cases of Down syndrome. In such instances, the chance of the same parents producing a second affected child is extremely low. However, in the remaining families with a Down child, the syndrome occurs with a much higher frequency over several generations.

Cytogenetic studies of the parents and their offspring from these unusual cases explain the cause of **familial Down syndrome.** Analysis reveals that one of the parents contains a 14/21, D/G translocation (**Figure 6–17**). That is, one parent has the majority of the G-group chromosome 21 translocated to one end of the D-group chromosome 14. This individual is phenotypically normal, even though he or she has only 45 chromosomes. During meiosis, one-fourth of the individual's gametes have two copies of chromosome 21: a normal chromosome and a second copy translocated to chromosome 14. When such a gamete is fertilized by a standard haploid gamete, the resulting zygote has 46 chromosomes but three copies of chromosome 21. These individuals exhibit Down syndrome. Other potential surviving offspring contain either the standard diploid genome (without a translocation) or the balanced translocation like the parent. Both cases result in normal individuals. Although not illustrated in Figure 6–17, two other gametes may be formed, though rarely. Such gametes are unbalanced, and upon fertilization, lethality occurs.

Knowledge of translocations, as described above, has allowed geneticists to resolve the seeming paradox of an inherited trisomic phenotype in an individual with an apparent diploid number of chromosomes. It is also unique that the "carrier," who has 45 chromosomes and exhibits a normal phenotype, does not contain the *complete* diploid amount of genetic material. A small region is lost from both chromosomes 14 and 21 during the translocation event. This occurs because the ends of both chromosomes have broken off prior to their fusion. These specific regions are known to be two of many chromosomal locations housing multiple copies of the genes encoding rRNA, the major component of ribosomes. Despite the loss of up to 20 percent of these genes, the carrier is unaffected.

ESSENTIAL POINT

Inversions and translocations may initially cause little or no loss of genetic information or deleterious effects. However, heterozygous combinations of the involved chromosome segments may result in genetically abnormal gametes following meiosis, with lethality or inviability often ensuing. ■

6.9 Fragile Sites in Human Chromosomes Are Susceptible to Breakage

We conclude this chapter with a brief discussion of the results of an intriguing discovery made around 1970 during observations of metaphase chromosomes prepared following human cell culture. In cells derived from certain individuals, a specific area along one of the chromosomes failed to stain, giving the appearance of a gap. In other individuals whose chromosomes displayed such morphology, the gaps appeared at other positions within the set of chromosomes. Such areas eventually became known as **fragile sites,** since they appeared to be susceptible to chromosome

breakage when cultured in the absence of certain chemicals such as folic acid, which is normally present in the culture medium. Fragile sites were at first considered curiosities, until a strong association was subsequently shown to exist between one of the sites and a form of mental retardation.

The cause of the fragility at these sites is unknown. Because they represent points along the chromosome that are susceptible to breakage, these sites may indicate regions where the chromatin is not tightly coiled. Note that even though almost all studies of fragile sites have been carried out *in vitro* using mitotically dividing cells, clear associations have been established between several of these sites and the corresponding altered phenotype, including mental retardation and cancer.

Fragile-X Syndrome

Most fragile sites do not appear to be associated with any clinical syndrome. However, individuals bearing a folate-sensitive site on the X chromosome (**Figure 6–18**) may exhibit the **fragile-X syndrome** (**FXS**), the most common form of inherited mental retardation. This syndrome affects about 1 in 4000 males and 1 in 8000 females. All males bearing this X chromosome exhibit the syndrome, while about 60 percent of females bearing one affected chromosome exhibit the syndrome. In addition to mental retardation, affected males and females have characteristic long, narrow faces with protruding chins and enlarged ears.

A gene that spans the fragile site is now known to be responsible for this syndrome. Named *FMR1,* it is one of a growing number of genes that have been discovered in which a sequence of three nucleotides is repeated many times, expanding the size of the gene. Such **trinucleotide repeats** are also recognized in other human disorders, including Huntington disease and myotonic dystrophy. In *FMR1,* the trinucleotide sequence CGG is repeated in an untranslated area adjacent to the coding sequence of the gene (called the "upstream" region). The number of repeats varies immensely within the human population, and a high number correlates directly with expression of fragile-X syndrome. Normal individuals have between 6 and 54 repeats, whereas those with 55 to 230 repeats are considered "carriers" of the disorder. More than 230 repeats lead to expression of the syndrome.

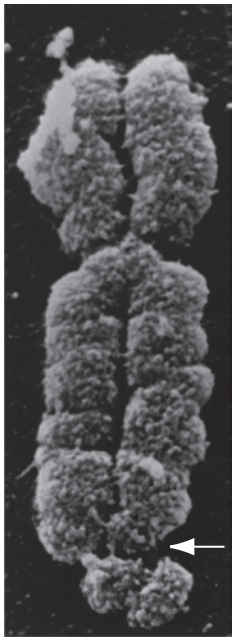


FIGURE 6–18 A fragile human X chromosome. The "gap" region, identified by the arrow, is associated with the fragile-X syndrome.

It is thought that, once the gene contains this increased number of repeats, it becomes chemically modified so that the bases within and around the repeats are methylated, an epigenetic process that inactivates the gene. The normal product of the *FMR1* gene is an RNA-binding protein, FMRP, known to be produced in the brain. Evidence is now accumulating that directly links the absence of the protein in the brain with the cognitive defects associated with the syndrome.

From a genetic standpoint, an interesting aspect of fragile-X syndrome is the instability of the number of CGG repeats. An individual with 6 to 54 repeats transmits a gene containing the same number of copies to his or her offspring. However, carrier individuals with 55 to 230 repeats, though not at risk to develop the syndrome, may transmit to their offspring a gene with an increased number of repeats. This number increases in future generations, demonstrating the phenomenon known as **genetic anticipation.** Once the threshold of 230 repeats is exceeded, retardation becomes more severe in each successive generation as the number of trinucleotide repeats increases. Interestingly, expansion from the carrier status (55 to 230 repeats) to the syndrome status (over 230 repeats) occurs only during the transmission of the gene by the maternal parent, not by the paternal parent. Thus, a "carrier" male may transmit a stable chromosome to his daughter, who may subsequently transmit an unstable chromosome with an increased number of repeats to her offspring. Their grandfather was the source of the original chromosome.

The Link between Fragile Sites and Cancer

While the study of the fragile-X syndrome first brought unstable chromosome regions to the attention of geneticists, a link between an autosomal fragile site and lung cancer was reported in 1996 by Carlo Croce, Kay Huebner, and their colleagues. They have subsequently postulated that the defect is associated with the formation of a variety of different tumor types. Croce and Huebner first showed that the *FHIT* gene (standing for *f*ragile *hi*stidine *t*riad),

located within the well-defined fragile site designated as *FRA3B* on the p arm of chromosome 3, is often altered or missing in cells taken from tumors of individuals with lung cancer. More extensive studies have now revealed that the normal protein product of this gene is absent in cells of many other cancers, including those of the esophagus, breast, cervix, liver, kidney, pancreas, colon, and stomach. Genes such as *FHIT* that are located within fragile regions undoubtedly have an increased susceptibility to mutations and deletions.

Subsequently, Muller Fabbri and Kay Huebner, working with others in Croce's lab, identified and studied another fragile site, with most interesting results. Found within the *FRA16D* site on chromosome 16 is the *WWOX* gene. Like the *FHIT* gene, it has been implicated in a range of human cancers. In particular, like *FHIT,* it has been found to be either lost or genetically silenced in the large majority of lung tumors, as well as in cancer tissue of the breast, ovary, prostate, bladder, esophagus, and pancreas. When the gene is present but silent, its DNA is thought to be heavily methylated, rendering it inactive. Furthermore, the active gene is also thought to behave as a tumor suppressor, providing a surveillance function by recognizing cancer cells and inducing apoptosis, effectively eliminating them before malignant tumors can be initiated.

ESSENTIAL POINT

Fragile sites in human mitotic chromosomes have sparked research interest because one such site on the X chromosome is associated with the most common form of inherited mental retardation, while other autosomal sites have been linked to various forms of cancer. ■

CASE STUDY | Changing the face of Down syndrome

Australian model Madeline Stuart hit the headlines when she took to the runway during the New York Fashion Week. The 18-year old is hailed as the face of GlossiGirl, a cosmetics company. She already has a collection of handbags named after her, with its proceeds donated to the National Down Syndrome Society. She has been nominated for the Pride of Australia award. Her aim is to inspire people all over the world with her slogan "I Can I Will". Why is any of these significant? Madeline Stuart has Down syndrome.

1. What would the attitude of the society towards people with Down syndrome have been 75 years ago? 200 years ago?
2. How are the examples of people like Madeline Stuart changing the attitude of the society towards people with Down syndrome?
3. What impact do social media have on this change?

INSIGHTS AND SOLUTIONS

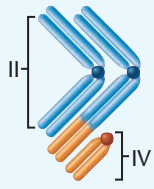
1. In a cross using maize that involves three genes, *a, b,* and *c,* a heterozygote (*abc*/+++) is testcrossed to *abc*/*abc.* Even though the three genes are separated along the chromosome, thus predicting that crossover gametes and the resultant phenotypes should be observed, only two phenotypes are recovered: *abc* and +++. In addition, the cross produced significantly fewer viable plants than expected. Can you propose why no other phenotypes were recovered and why the viability was reduced?

   Solution: One of the two chromosomes may contain an inversion that overlaps all three genes, effectively precluding the recovery of any "crossover" offspring. If this is a paracentric inversion and the genes are clearly separated (assuring that a significant number of crossovers occurs between them), then numerous acentric and dicentric chromosomes will form, resulting in the observed reduction in viability.

2. A male *Drosophila* from a wild-type stock is discovered to have only seven chromosomes, whereas normally $2n = 8$. Close examination reveals that one member of chromosome IV (the smallest chromosome) is attached to (translocated to) the distal end of chromosome II and is missing its centromere, thus accounting for the reduction in chromosome number.

   (a) Diagram all members of chromosomes II and IV during synapsis in meiosis I.

   Solution:

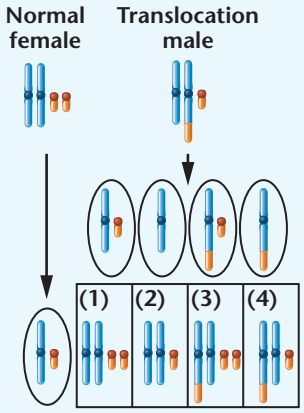


   (b) If this male mates with a female with a normal chromosome composition who is homozygous for the recessive chromosome IV mutation *eyeless* (*ey*), what chromosome compositions will occur in the offspring regarding chromosomes II and IV?

(continued)

Insights and Solutions—continued

Solution:



(c) Referring to the diagram in the solution to part (b), what phenotypic ratio will result regarding the presence of eyes, assuming all abnormal chromosome compositions survive?

Solution:

1. normal (heterozygous)
2. eyeless (monosomic, contains chromosome IV from mother)
3. normal (heterozygous; trisomic and may die)
4. normal (heterozygous; balanced translocation)

The final ratio is 3/4 normal: 1/4 eyeless.

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we focused on chromosomal mutations resulting from a change in number or arrangement of chromosomes. In our discussions, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions?
   (a) How do we know that the extra chromosome causing Down syndrome is usually maternal in origin?
   (b) How do we know that human aneuploidy for each of the 22 autosomes occurs at conception, even though most often human aneuploids do not survive embryonic or fetal development and thus are never observed at birth?
   (c) How do we know that specific mutant phenotypes are due to changes in chromosome number or structure?
   (d) How do we know that the mutant *Bar*-eye phenotype in *Drosophila* is due to a duplicated gene region rather than to a change in the nucleotide sequence of a gene?

CONCEPT QUESTION

2. Review the Chapter Concepts list on page 115. These all center on chromosome aberrations that create variations from the "normal" diploid genome. Write a short essay that discusses five altered phenotypes that result from specific chromosomal aberrations. ■

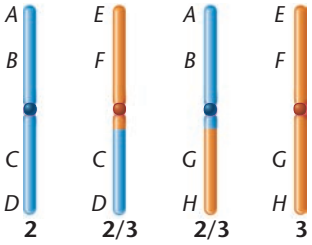
3. Define these pairs of terms, and distinguish between them.

   aneuploidy/euploidy
   monosomy/trisomy
   Patau syndrome/Edwards syndrome
   autopolyploidy/allopolyploidy
   autotetraploid/amphidiploid
   paracentric inversion/pericentric inversion

4. How can you explain that not all Down syndrome cases are due to nondisjunction?
5. Compare partial monosomy with haploinsufficiency.
6. What are the possible reasons behind translocations?
7. Why do human monosomics most often fail to survive prenatal development?
8. What advantages and disadvantages do polyploid plants have?
9. A couple goes through multiple miscarriages, and fetal karyotyping indicates the same trisomy every time. This trisomy is not compatible with life. Obviously, neither of the parents has this trisomy. Can you explain this situation?
10. What are inversion heterozygotes? How can meiotic pairing occur in these organisms? What will be the consequence?
11. Predict the genetic composition of gametes derived from tetrads of inversion heterozygotes where crossing over occurs within a pericentric inversion.
12. Human adult hemoglobin is a tetramer containing two alpha (α) and two beta (β) polypeptide chains. The a gene cluster on chromosome 16 and the β gene cluster on chromosome 11 share amino acid similarities such that 61 of the amino acids of the α-globin polypeptide (141 amino acids long) are shared in identical sequence with the β-globin polypeptide (146 amino acids long). How might one explain the existence of two polypeptides with partially shared function and structure on two different chromosomes? Include in your answer a link to Ohno's hypothesis regarding the origin of new genes during evolution.
13. The primrose, *Primula kewensis,* has 36 chromosomes that are similar in appearance to the chromosomes in two related species, *P. floribunda* ($2n = 18$) and *P. verticillata* ($2n = 18$). How could *P. kewensis* arise from these species? How would you describe *P. kewensis* in genetic terms?
14. Certain varieties of chrysanthemums contain 18, 36, 54, 72, and 90 chromosomes; all are multiples of a basic set of nine chromosomes. How would you describe these varieties genetically? What feature do the karyotypes of each variety share? A variety with 27 chromosomes has been discovered, but it is sterile. Why?
15. *Drosophila* may be monosomic for chromosome 4, yet remain fertile. Contrast the F_1 and F_2 results of the following crosses involving the recessive chromosome 4 trait, *bent* bristles: (a) monosomic IV, bent bristles × normal bristles; (b) monosomic IV, normal bristles × bent bristles.
16. Mendelian ratios are modified in crosses involving autotetraploids. Assume that one plant expresses the dominant trait

green seeds and is homozygous (*WWWW*). This plant is crossed to one with white seeds that is also homozygous (*wwww*). If only one dominant allele is sufficient to produce green seeds, predict the F_1 and F_2 results of such a cross. Assume that synapsis between chromosome pairs is random during meiosis.

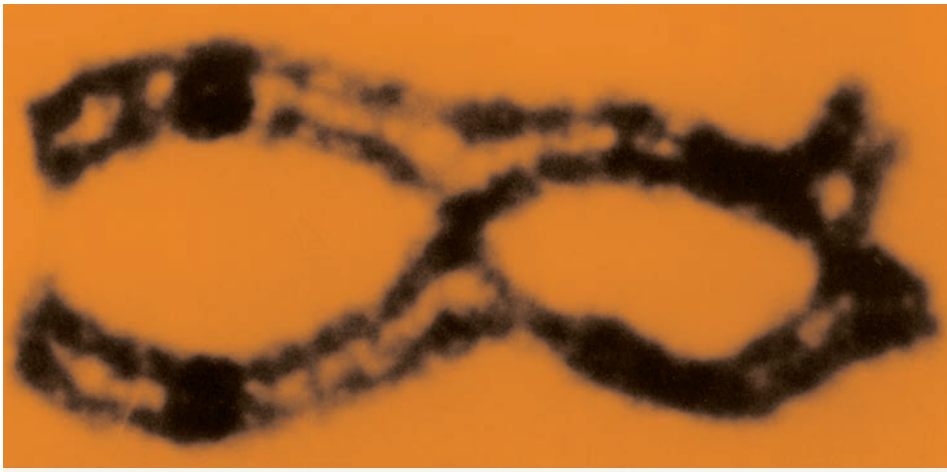
17. Having correctly established the F_2 ratio in Problem 16, predict the F_2 ratio of a "dihybrid" cross involving two independently assorting characteristics (e.g., P_1 = *WWWWAAAA* × *wwwwaaaa*).
18. In a cross between two varieties of corn, $gl_1gl_1Ws_3Ws_3$ (egg parent) × $Gl_1Gl_1ws_3ws_3$ (pollen parent), a triploid offspring was produced with the genetic constitution $Gl_1Gl_1gl_1Ws_3ws_3ws_3$. From which parent, egg or pollen, did the 2*n* gamete originate? Is another explanation possible? Explain.
19. A couple planning their family are aware that through the past three generations on the husband's side a substantial number of stillbirths have occurred and several malformed babies were born who died early in childhood. The wife has studied genetics and urges her husband to visit a genetic counseling clinic, where a complete karyotype-banding analysis is performed. Although the tests show that he has a normal complement of 46 chromosomes, banding analysis reveals that one member of the chromosome 1 pair (in group A) contains an inversion covering 70 percent of its length. The homolog of chromosome 1 and all other chromosomes show the normal banding sequence. (a) How would you explain the high incidence of past stillbirths? (b) What can you predict about the probability of abnormality/normality of their future children? (c) Would you advise the woman that she will have to bring each pregnancy to term to determine whether the fetus is normal? If not, what else can you suggest?
20. A woman who sought genetic counseling is found to be heterozygous for a chromosomal rearrangement between the second and third chromosomes. Her chromosomes, compared to those in a normal karyotype, are diagrammed here:



(a) What kind of chromosomal aberration is shown?
(b) Using a drawing, demonstrate how these chromosomes would pair during meiosis. Be sure to label the different segments of the chromosomes.
(c) This woman is phenotypically normal. Does this surprise you? Why or why not? Under what circumstances might you expect a phenotypic effect of such a rearrangement?

21. The woman in Problem 20 has had two miscarriages. She has come to you, an established genetic counselor, with these questions: (a) Is there a genetic explanation of her frequent miscarriages? (b) Should she abandon her attempts to have a child of her own? (c) If not, what is the chance that she could have a normal child? Provide an informed response to her concerns.
22. In a sample of 1000 patients with Down syndrome, a geneticist discovers that 95% of them are trisomic, while 5% have diploid number of chromosomes. Explain this discrepancy.
23. A boy with Klinefelter syndrome (47,XXY) is born to a mother who is phenotypically normal and a father who has the X-linked skin condition called anhidrotic ectodermal dysplasia. The mother's skin is completely normal with no signs of the skin abnormality. In contrast, her son has patches of normal skin and patches of abnormal skin. (a) Which parent contributed the abnormal gamete? (b) Using the appropriate genetic terminology, describe the meiotic mistake that occurred. Be sure to indicate in which division the mistake occurred. (c) Using the appropriate genetic terminology, explain the son's skin phenotype.
24. In a human genetic study, a family with five phenotypically normal children was investigated. Two children were "homozygous" for a Robertsonian translocation between chromosomes 19 and 20 (they contained two identical copies of the fused chromosome). They have only 44 chromosomes but a complete genetic complement. Three of the children were "heterozygous" for the translocation and contained 45 chromosomes, with one translocated chromosome plus a normal copy of both chromosomes 19 and 20. Two other pregnancies resulted in stillbirths. It was later discovered that the parents were first cousins. Based on this information, determine the chromosome compositions of the parents. What led to the stillbirths? Why was the discovery that the parents were first cousins a key piece of information in understanding the genetics of this family?
25. A couple has two children, of whom one has Turner syndrome and the other has Klinefelter syndrome. A genetic analysis of the parents reveals one to produce normal gametes. Which of the two parents is normal? What kind of abnormality would you predict in the gametes of the other parent?
26. A normal female is discovered with 45 chromosomes, one of which exhibits a Robertsonian translocation containing most of chromosomes 18 and 21. Discuss the possible outcomes in her offspring when her husband contains a normal karyotype.

7 Linkage and Chromosome Mapping in Eukaryotes



Chiasmata between synapsed homologs during the first meiotic prophase.

CHAPTER CONCEPTS

- Chromosomes in eukaryotes contain many genes whose locations are fixed along the length of the chromosomes.
- Unless separated by crossing over, alleles present on a chromosome segregate as a unit during gamete formation.
- Crossing over between homologs is a process of genetic recombination during meiosis that creates gametes with new combinations of alleles that enhance genetic variation within species.
- Crossing over between homologs serves as the basis for the construction of chromosome maps.
- While exchange occurs between sister chromatids during mitosis, no new recombinant chromatids are created.

Walter Sutton, along with Theodor Boveri, was instrumental in uniting the fields of cytology and genetics. As early as 1903, Sutton pointed out the likelihood that there must be many more "unit factors" than chromosomes in most organisms. Soon thereafter, genetics investigations revealed that certain genes segregate as if they were somehow joined or linked together. Further investigations showed that such genes are part of the same chromosome and may indeed be transmitted as a single unit. We now know that most chromosomes contain a very large number of genes. Those that are part of the same chromosome are said to be *linked* and to demonstrate **linkage** in genetic crosses.

Because the chromosome, not the gene, is the unit of transmission during meiosis, linked genes are not free to undergo independent assortment. Instead, the alleles at all loci of one chromosome should, in theory, be transmitted as a unit during gamete formation. However, in many instances this does not occur. During the first meiotic prophase, when homologs are paired or synapsed, a reciprocal exchange of chromosome segments can take place. This **crossing over** event results in the reshuffling, or **recombination,** of the alleles between homologs, and it always occurs during the tetrad stage.

The frequency of crossing over between any two loci on a single chromosome is proportional to the distance between them. Therefore, depending on which loci are being studied, the percentage of recombinant gametes varies. This correlation allows us to construct chromosome maps, which give the relative locations of genes on chromosomes.

In this chapter, we will discuss linkage, crossing over, and chromosome mapping in more detail.

7.1 Genes Linked on the Same Chromosome Segregate Together

A simplified overview of the major theme of this chapter is given in **Figure 7–1**, which contrasts the meiotic consequences of (a) independent assortment, (b) linkage *without* crossing over, and (c) linkage *with* crossing over. In **Figure 7–1(a)**, we see the results of independent assortment of two pairs of chromosomes, each containing one heterozygous gene pair. No linkage is exhibited. When a large number of meiotic events are observed, four genetically different gametes are formed in equal proportions, and each contains a different combination of alleles of the two genes.

Now let's compare these results with what occurs if the same genes are linked on the same chromosome. If no crossing over occurs between the two genes [**Figure 7–1(b)**], only two genetically different gametes are formed. Each gamete receives the alleles present on one homolog or the other, which is transmitted intact as the result of segregation. This case demonstrates *complete linkage,* which produces only **parental** or **noncrossover gametes.** The two parental gametes are formed in equal proportions. Though complete linkage between two genes seldom occurs, it is useful to consider the theoretical consequences of this concept.

Figure 7–1(c) shows the results of crossing over between two linked genes. As you can see, this crossover involves only two nonsister chromatids of the four chromatids present in the tetrad. This exchange generates two new allele combinations, called **recombinant** or **crossover gametes.** The two chromatids not involved in the exchange result in noncrossover gametes, like those in Figure 7–1(b). The frequency with which crossing over occurs between any two linked genes is generally proportional to the distance separating the respective loci along the chromosome. In theory, two randomly selected genes can be so close to each other that crossover events are too infrequent to be detected easily. As shown in Figure 7–1(b), this complete linkage produces only parental gametes. On the other hand, if a small but distinct distance separates two

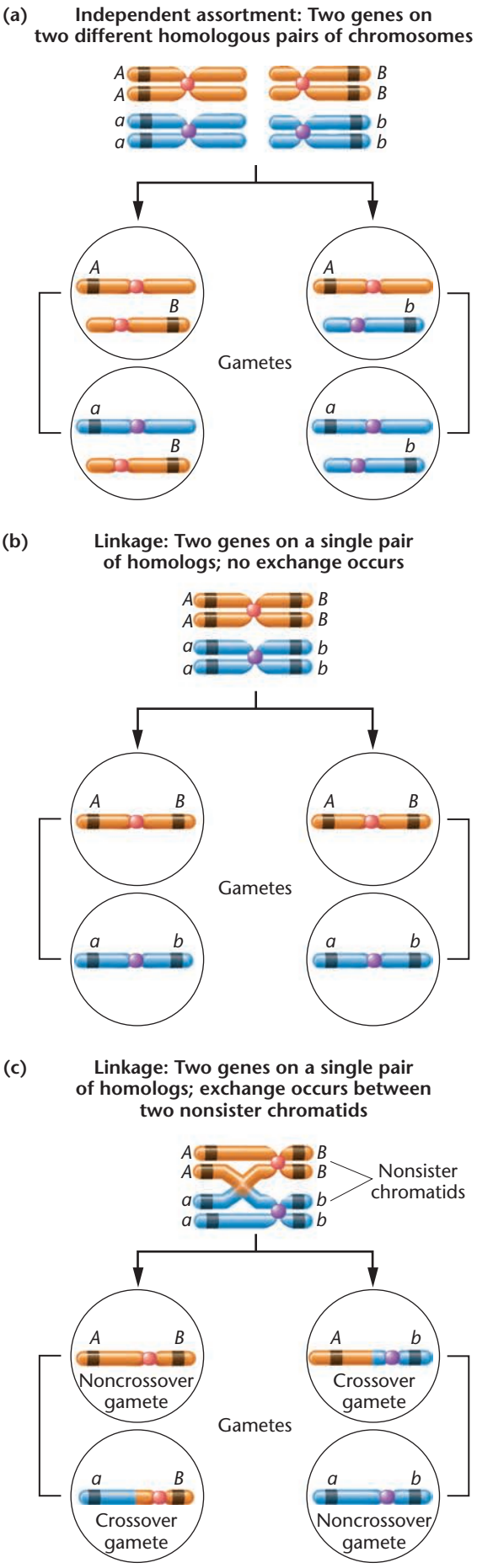


FIGURE 7–1 Results of gamete formation when two heterozygous genes are (a) on two different pairs of chromosomes; (b) on the same pair of homologs, but with no exchange occurring between them; and (c) on the same pair of homologs, but with an exchange occurring between two nonsister chromatids. Note in this and the following figures that members of homologous pairs of chromosomes are shown in two different colors. This convention was established in Chapter 2 (see, for example, Figure 2–7 and Figure 2–11).

genes, few recombinant and many parental gametes will be formed. As the distance between the two genes increases, the proportion of recombinant gametes increases and that of the parental gametes decreases.

As we will discuss later in this chapter, when the loci of two linked genes are far apart, the number of recombinant gametes approaches, but does not exceed, 50 percent. If 50 percent recombinants occur, the result is a 1:1:1:1 ratio of the four types (two parental and two recombinant gametes). In this case, transmission of two linked genes is indistinguishable from that of two unlinked, independently assorting genes. That is, the proportion of the four possible genotypes is identical, as shown in Figure 7–1(a) and (c).

The Linkage Ratio

If complete linkage exists between two genes because of their close proximity, and organisms heterozygous at both loci are mated, a unique F_2 phenotypic ratio results, which we designate the **linkage ratio.** To illustrate this ratio, let's consider a cross involving the closely linked, recessive, mutant genes *heavy* wing vein (*hv*) and *brown* eye (*bw*) in *Drosophila melanogaster* (**Figure 7–2**). The normal, wild-type alleles hv^+ and bw^+ are both dominant and result in thin wing veins and red eyes, respectively.

In this cross, flies with normal thin wing veins and mutant brown eyes are mated to flies with mutant heavy wing veins and normal red eyes. In more concise terms, brown-eyed flies are crossed with heavy-veined flies. If we extend the system of genetic symbols established in Chapter 4, linked genes are represented by placing their allele designations above and below a single or double horizontal line. Those above the line are located at loci on one homolog, and those below are located at the homologous loci on the other homolog. Thus, we represent the P_1 generation as follows:

$$P_1: \frac{hv^+\ bw}{hv^+\ bw} \times \frac{hv\ bw^+}{hv\ bw^+}$$

thin, brown heavy, red

These genes are located on an autosome, so no distinction between males and females is necessary.

In the F_1 generation, each fly receives one chromosome of each pair from each parent. All flies are heterozygous for both gene pairs and exhibit the dominant traits of thin wing veins and red eyes:

$$F_1: \frac{hv^+\ bw}{hv\ \ bw^+}$$

thin, red

As shown in **Figure 7–2(a)**, when the F_1 generation is interbred, each F_1 individual forms only parental gametes because of complete linkage. After fertilization, the F_2 generation is produced in a 1:2:1 phenotypic and genotypic ratio. One-fourth of this generation shows thin wing veins and brown eyes; one-half shows both wild-type traits, namely, thin wing veins and red eyes; and one-fourth shows heavy wing veins and red eyes. In more concise terms, the ratio is 1 heavy:2 wild:1 brown. Such a 1:2:1 ratio is characteristic of complete linkage. Complete linkage is usually observed only when genes are very close together and the number of progeny is relatively small.

Figure 7–2(b) demonstrates the results of a testcross with the F_1 flies. Such a cross produces a 1:1 ratio of thin, brown and heavy, red flies. Had the genes controlling these traits been incompletely linked or located on separate autosomes, the testcross would have produced four phenotypes rather than two.

When large numbers of mutant genes present in any given species are investigated, genes located on the same chromosome show evidence of linkage to one another. As a result, **linkage groups** can be established, one for each chromosome. In theory, the number of linkage groups should correspond to the haploid number of chromosomes. In diploid organisms in which large numbers of mutant genes are available for genetic study, this correlation has been confirmed.

NOW SOLVE THIS

7–1 Consider two hypothetical recessive autosomal genes *a* and *b,* where a heterozygote is testcrossed to a double-homozygous mutant. Predict the phenotypic ratios under the following conditions:

(a) *a* and *b* are located on separate autosomes.
(b) *a* and *b* are linked on the same autosome but are so far apart that a crossover always occurs between them.
(c) *a* and *b* are linked on the same autosome but are so close together that a crossover almost never occurs.

HINT: *This problem involves an understanding of linkage, crossing over, and independent assortment. The key to its solution is to be aware that results are indistinguishable when two genes are unlinked compared to the case where they are linked but so far apart that crossing over always intervenes between them during meiosis.*

ESSENTIAL POINT

Genes located on the same chromosome are said to be linked. Alleles of linked genes located close together on the same homolog are usually transmitted together during gamete formation.

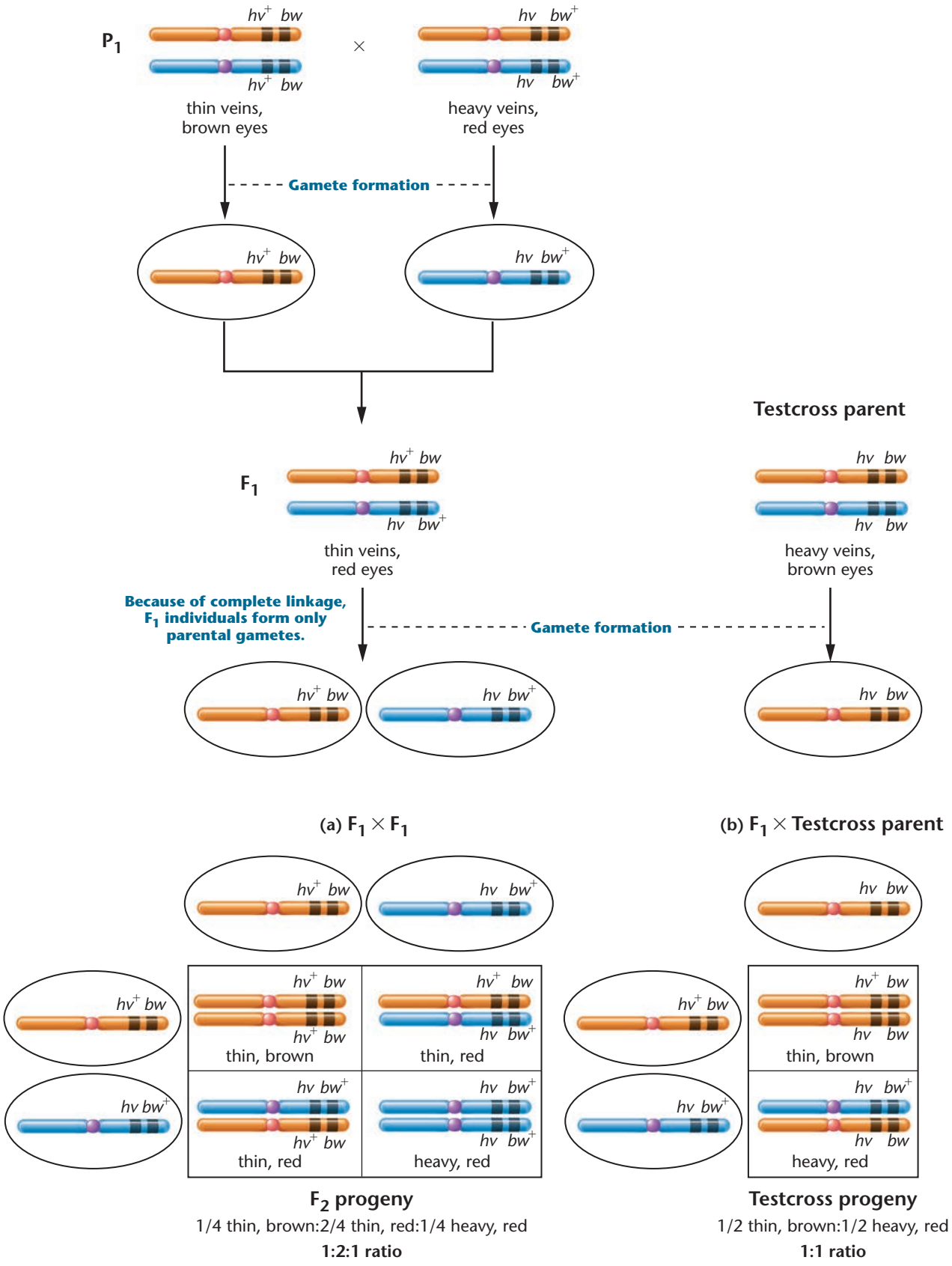


FIGURE 7–2 Results of a cross involving two genes located on the same chromosome and demonstrating complete linkage. (a) The F_2 results of the cross. (b) The results of a testcross involving the F_1 progeny.

7.2 Crossing Over Serves as the Basis of Determining the Distance between Genes during Mapping

It is highly improbable that two randomly selected genes linked on the same chromosome will be so close to one another along the chromosome that they demonstrate complete linkage. Instead, crosses involving two such genes almost always produce a percentage of offspring resulting from recombinant gametes. This percentage is variable and depends on the distance between the two genes along the chromosome. This phenomenon was first explained around 1910 by two *Drosophila* geneticists, Thomas H. Morgan and his undergraduate student, Alfred H. Sturtevant.

Morgan, Sturtevant, and Crossing Over

In his studies, Morgan investigated numerous *Drosophila* mutations located on the X chromosome. When he analyzed crosses involving only one trait, he deduced the mode of X-linked inheritance. However, when he made crosses involving two X-linked genes, his results were initially puzzling. For example, female flies expressing the mutant *yellow* body (*y*) and *white* eyes (*w*) alleles were crossed with wild-type males (gray bodies and red eyes). The F_1 females were wild type, while the F_1 males expressed both mutant traits. In the F_2 generation, the vast majority of the offspring showed the expected parental phenotypes—either yellow-bodied, white-eyed flies or wild-type flies (gray-bodied, red-eyed). However, the remaining flies, less than 1.0 percent, were either yellow-bodied with red eyes or gray-bodied with white eyes. It was as if the two mutant alleles had somehow separated from each other on the homolog during gamete formation in the F_1 female flies. This cross is illustrated in cross A of **Figure 7–3** , using data later compiled by Sturtevant.

When Morgan studied other X-linked genes, the same basic pattern was observed, but the proportion of the unexpected F_2 phenotypes differed. For example, in a cross involving the mutant *white* eye (*w*), *miniature* wing (*m*) alleles, the majority of the F_2 again showed the parental phenotypes, but a much higher proportion of the offspring appeared as if the mutant genes had separated during gamete formation. This is illustrated in cross B of Figure 7–3, again using data subsequently compiled by Sturtevant.

Morgan was faced with two questions: (1) What was the source of gene separation, and (2) why did the frequency of the apparent separation vary depending on the genes being studied? The answer he proposed for the first question was based on his knowledge of earlier cytological observations made by F.A. Janssens and others. Janssens had observed that synapsed homologous chromosomes in meiosis wrapped around each other, creating **chiasmata** (sing., *chiasma*) where points of overlap are evident (see the photo on p. 153). Morgan proposed that chiasmata could represent points of genetic exchange.

In the crosses shown in Figure 7–3, Morgan postulated that if an exchange occurs during gamete formation between the mutant genes on the two X chromosomes of the F_1 females, the unique phenotypes will occur. He suggested that such exchanges led to **recombinant gametes** in both the *yellow–white* cross and the *white–miniature* cross, in contrast to the **parental gametes** that have undergone no exchange. On the basis of this and other experiments, Morgan concluded that linked genes exist in a linear order along the chromosome and that a variable amount of exchange occurs between any two genes during gamete formation.

In answer to the second question, Morgan proposed that two genes located relatively close to each other along a chromosome are less likely to have a chiasma form between them than if the two genes are farther apart on the chromosome. Therefore, the closer two genes are, the less likely a genetic exchange will occur between them. Morgan was the first to propose the term **crossing over** to describe the physical exchange leading to recombination.

Sturtevant and Mapping

Morgan's student, Alfred H. Sturtevant, was the first to realize that his mentor's proposal could be used to map the sequence of linked genes. According to Sturtevant,

> *In a conversation with Morgan . . . I suddenly realized that the variations in strength of linkage, already attributed by Morgan to differences in the spatial separation of the genes, offered the possibility of determining sequences in the linear dimension of a chromosome. I went home and spent most of the night (to the neglect of my undergraduate homework) in producing the first chromosomal map.*

Sturtevant compiled data from numerous crosses made by Morgan and other geneticists involving recombination between the genes represented by the *yellow, white,* and *miniature* mutants. These data are shown in Figure 7–3. The following recombination between each pair of these three genes, published in Sturtevant's paper in 1913, is as follows:

(1)	*yellow–white*	0.5%
(2)	*white–miniature*	34.5%
(3)	*yellow–miniature*	35.4%

Because the sum of (1) and (2) approximately equals (3), Sturtevant suggested that the recombination frequencies between linked genes are additive. On this basis, he

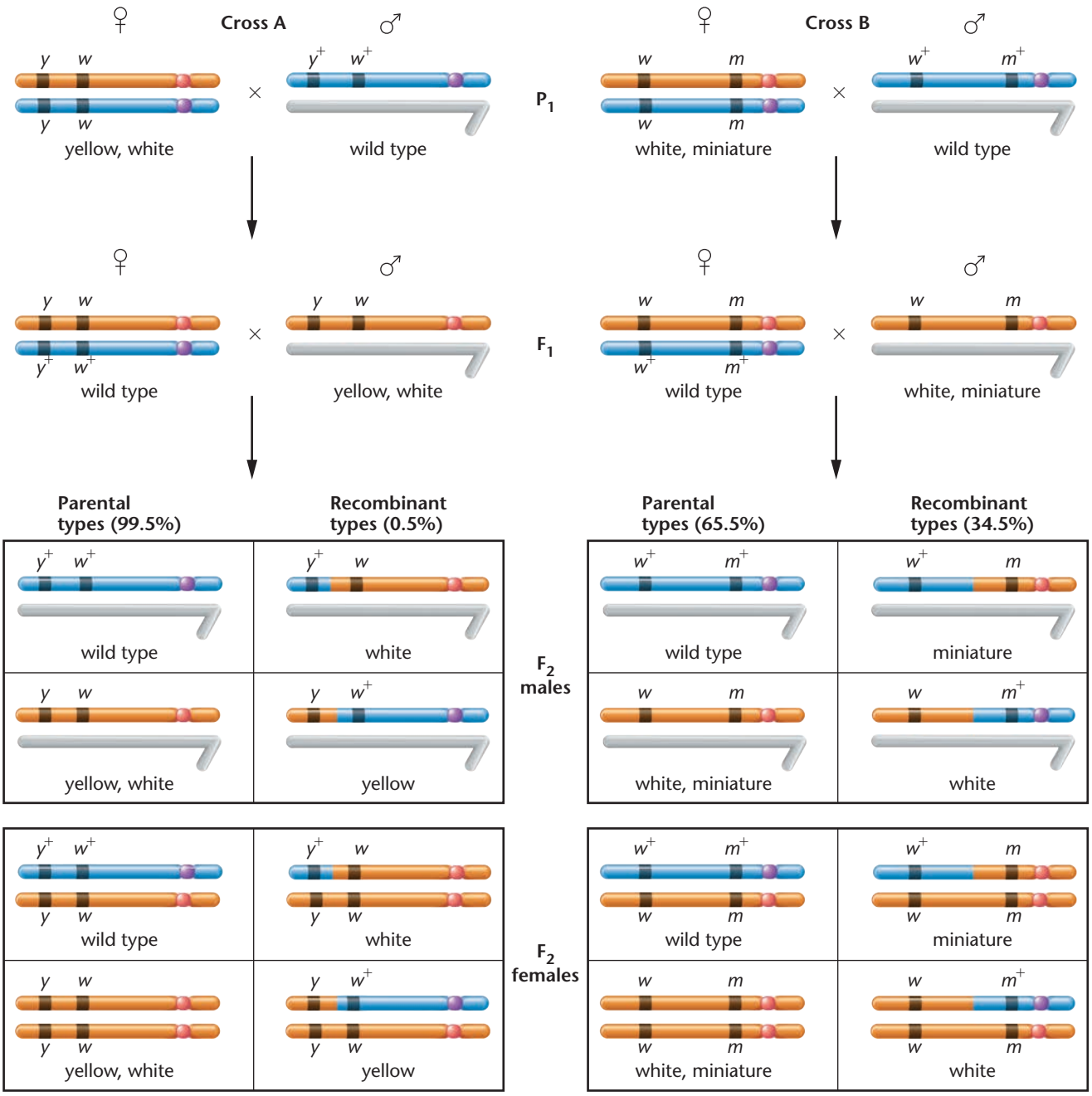


FIGURE 7–3 The F_1 and F_2 results of crosses involving the *yellow* (*y*), *white* (*w*) mutations (cross A), and the *white, miniature* (*m*) mutations (cross B), as compiled by Sturtevant. In cross A, 0.5 percent of the F_2 flies (males and females) demonstrate recombinant phenotypes, which express either *white* or *yellow*. In cross B, 34.5 percent of the F_2 flies (males and females) demonstrate recombinant phenotypes, which are either *miniature* or *white* mutants.

predicted that the order of the genes on the X chromosome is *yellow–white–miniature*. In arriving at this conclusion, he reasoned as follows: the *yellow* and *white* genes are apparently close to each other because the recombination frequency is low. However, both of these genes are much farther apart from *miniature* genes because the *white–miniature* and *yellow–miniature* combinations show larger recombination frequencies. Because *miniature* shows more recombination with *yellow* than with *white* (35.4 percent versus 34.5 percent), it follows that *white* is located between the other two genes, not outside of them.

Sturtevant knew from Morgan's work that the frequency of exchange could be used as an estimate of the distance between two genes or loci along the chromosome. He constructed a **chromosome map** of the three genes on the X chromosome, setting 1 map unit (mu) equal to 1 percent

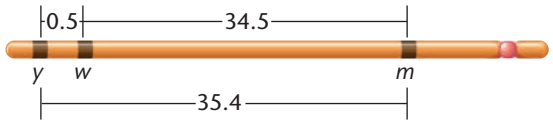


FIGURE 7–4 A map of the *yellow* (*y*), *white* (*w*), and *miniature* (*m*) genes on the X chromosome of *Drosophila melanogaster.* Each number represents the percentage of recombinant offspring produced in one of three crosses, each involving two different genes.

recombination between two genes.* In the preceding example, the distance between *yellow* and *white* is thus 0.5 mu, and the distance between *yellow* and *miniature* is 35.4 mu. It follows that the distance between *white* and *miniature* should be 35.4 − 0.5 = 34.9 mu. This estimate is close to the actual frequency of recombination between *white* and *miniature* (34.5 percent). The map for these three genes is shown in **Figure 7–4** . The fact that these do not add up perfectly is due to the imprecision of mapping experiments, particularly as the distance between genes increases.

In addition to these three genes, Sturtevant considered two other genes on the X chromosome and produced a more extensive map that included all five genes. He and a colleague, Calvin Bridges, soon began a search for autosomal linkage in *Drosophila.* By 1923, they had clearly shown that linkage and crossing over are not restricted to X-linked genes but can also be demonstrated with autosomes. During this work, they made another interesting observation. Crossing over in *Drosophila* was shown to occur only in females. The fact that no crossing over occurs in males made genetic mapping much less complex to analyze in *Drosophila.* However, crossing over does occur in both sexes in most other organisms.

Although many refinements in chromosome mapping have been developed since Sturtevant's initial work, his basic principles are considered to be correct. These principles are used to produce detailed chromosome maps of organisms for which large numbers of linked mutant genes are known. Sturtevant's findings are also historically significant to the broader field of genetics. In 1910, the **chromosomal theory of inheritance** was still widely disputed—even Morgan was skeptical of this theory before he conducted his experiments. Research has now firmly established that chromosomes contain genes in a linear order and that these genes are the equivalent of Mendel's unit factors.

Single Crossovers

Why should the relative distance between two loci influence the amount of recombination and crossing over observed between them? During meiosis, a limited number of crossover events occur in each tetrad. These recombinant events occur randomly along the length of the tetrad. Therefore, the closer two loci reside along the axis of the chromosome, the less likely any single-crossover event will occur between them. The same reasoning suggests that the farther apart two linked loci are, the more likely a random crossover event will occur between them.

In **Figure 7–5(a)**, a **single crossover** occurs between two nonsister chromatids but not between the two loci; therefore, the crossover is not detected because no recombinant gametes are produced. In **Figure 7–5(b)**, where two loci are quite far apart, a crossover does occur between them, yielding gametes in which the traits of interest are recombined.

When a single crossover occurs between two nonsister chromatids, the other two chromatids of the tetrad are not involved in this exchange and enter the gamete unchanged. Even if a single crossover occurs 100 percent of the time between two linked genes, recombination is subsequently observed in only 50 percent of the potential gametes formed. This concept is diagrammed in **Figure 7–6**. Theoretically, if we consider only single exchanges and observe

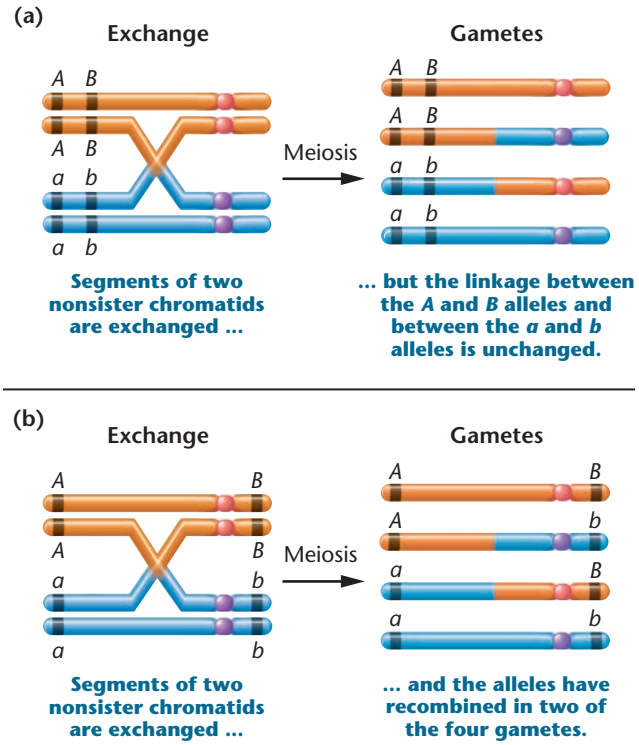


FIGURE 7–5 Two examples of a single crossover between two nonsister chromatids and the gametes subsequently produced. In (a) the exchange does not alter the linkage arrangement between the alleles of the two genes, only parental gametes are formed, and the exchange goes undetected. In (b) the exchange separates the alleles, resulting in recombinant gametes, which are detectable.

* In honor of Morgan's work, map units are often referred to as centiMorgans (cM).

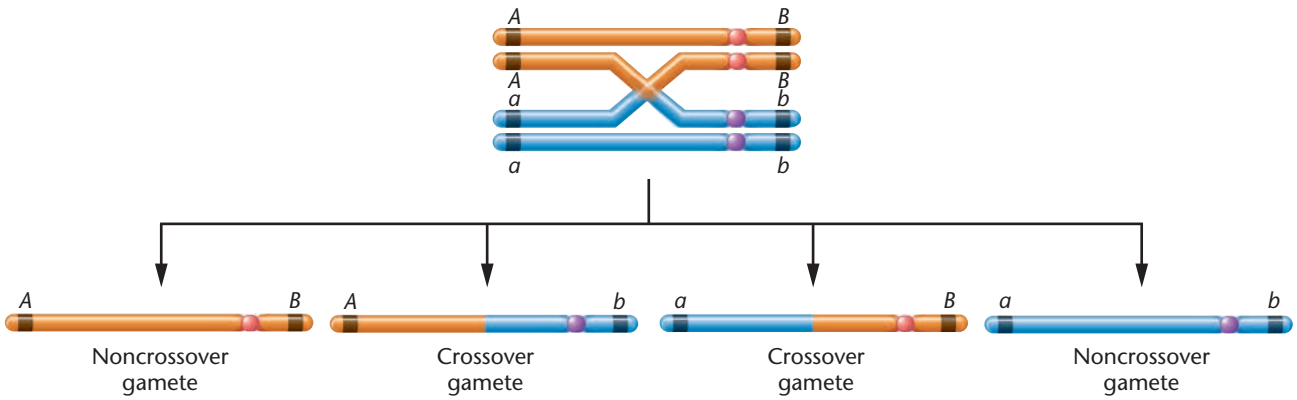


FIGURE 7–6 The consequences of a single exchange between two nonsister chromatids occurring in the tetrad stage. Two noncrossover (parental) and two crossover (recombinant) gametes are produced.

20 percent recombinant gametes, crossing over actually occurred in 40 percent of the tetrads. Under these conditions, the general rule is that the percentage of tetrads involved in an exchange between two genes is twice the percentage of recombinant gametes produced. Therefore, the theoretical limit of observed recombination due to crossing over is 50 percent.

When two linked genes are more than 50 mu apart, a crossover can theoretically be expected to occur between them in 100 percent of the tetrads. If this prediction were achieved, each tetrad would yield equal proportions of the four gametes shown in Figure 7–6, just as if the genes were on different chromosomes and assorting independently. However, this theoretical limit is seldom achieved.

ESSENTIAL POINT

Crossover frequency between linked genes during gamete formation is proportional to the distance between genes, providing the experimental basis for mapping the location of genes relative to one another along the chromosome. ■

7.3 Determining the Gene Sequence during Mapping Requires the Analysis of Multiple Crossovers

The study of single crossovers between two linked genes provides the basis of determining the *distance* between them. However, when many linked genes are studied, their *sequence* along the chromosome is more difficult to determine. Fortunately, the discovery that multiple exchanges occur between the chromatids of a tetrad has facilitated the process of producing more extensive chromosome maps. As we shall see next, when three or more linked genes are investigated simultaneously, it is possible to determine first the sequence of the genes and then the distances between them.

Multiple Crossovers

It is possible that in a single tetrad, two, three, or more exchanges will occur between nonsister chromatids as a result of several crossover events. Double exchanges of genetic material result from **double crossovers (DCOs),** as shown in Figure 7–7. For a double exchange to be studied, three gene pairs must be investigated, each heterozygous for two alleles. Before we determine the frequency of recombination among all three loci, let's review some simple probability calculations.

As we have seen, the probability of a single exchange occurring between the *A* and *B* or the *B* and *C* genes relates directly to the distance between the respective loci. The

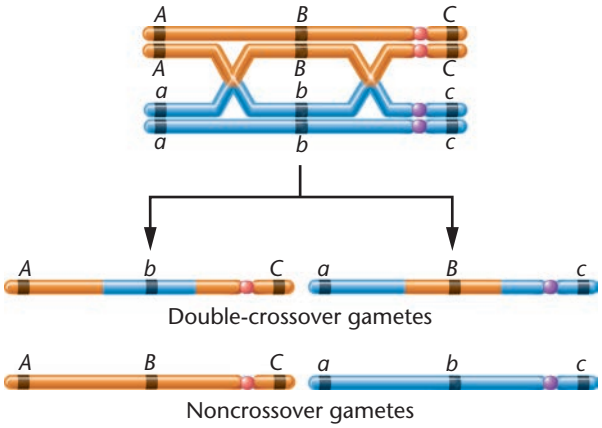


FIGURE 7–7 Consequences of a double exchange occurring between two nonsister chromatids. Because the exchanges involve only two chromatids, two noncrossover gametes and two double-crossover gametes are produced. The chapter opening photograph on p. 136 illustrates two chiasmata present in a tetrad isolated during the first meiotic prophase stage.

closer *A* is to *B* and *B* is to *C*, the less likely a single exchange will occur between either of the two sets of loci. In the case of a double crossover, two separate and independent events or exchanges must occur simultaneously. The mathematical probability of two independent events occurring simultaneously is equal to the product of the individual probabilities (the **product law**).

Suppose that crossover gametes resulting from single exchanges are recovered 20 percent of the time ($p = 0.20$) between *A* and *B*, and 30 percent of the time ($p = 0.30$) between *B* and *C*. The probability of recovering a double-crossover gamete arising from two exchanges (between *A* and *B*, and between *B* and *C*) is predicted to be (0.20) (0.30) = 0.06, or 6 percent. It is apparent from this calculation that the frequency of double-crossover gametes is always expected to be much lower than that of either single-crossover class of gametes.

If three genes are relatively close together along one chromosome, the expected frequency of double-crossover gametes is extremely low. For example, suppose the *A*–*B* distance in Figure 7–7 is 3 mu and the *B*–*C* distance is 2 mu. The expected double-crossover frequency is (0.03) (0.02) = 0.0006, or 0.06 percent. This translates to only 6 events in 10,000. Thus, in a mapping experiment where closely linked genes are involved, very large numbers of offspring are required to detect double-crossover events. In this example, it is unlikely that a double crossover will be observed even if 1000 offspring are examined. Thus, it is evident that if four or five genes are being mapped, even fewer triple and quadruple crossovers can be expected to occur.

NOW SOLVE THIS

7–2 With two pairs of genes involved (*P*/*p* and *Z*/*z*), a testcross (*ppzz*) with an organism of unknown genotype indicated that the gametes produced were in the following proportions

PZ, 42.4%; *Pz*, 6.9%; *pZ*, 7.1%; *pz*, 43.6%

Draw all possible conclusions from these data.

HINT: *This problem involves an understanding of the proportionality between crossover frequency and distance between genes. The key to its solution is to be aware that noncrossover and crossover gametes occur in reciprocal pairs of approximately equal proportions.*

ESSENTIAL POINT

Determining the sequence of genes in a three-point mapping experiment requires analysis of the double-crossover gametes, as reflected in the phenotype of the offspring receiving those gametes.

Three-Point Mapping in *Drosophila*

The information in the preceding section enables us to map three or more linked genes in a single cross. To illustrate the mapping process in its entirety, we examine two situations involving three linked genes in two quite different organisms.

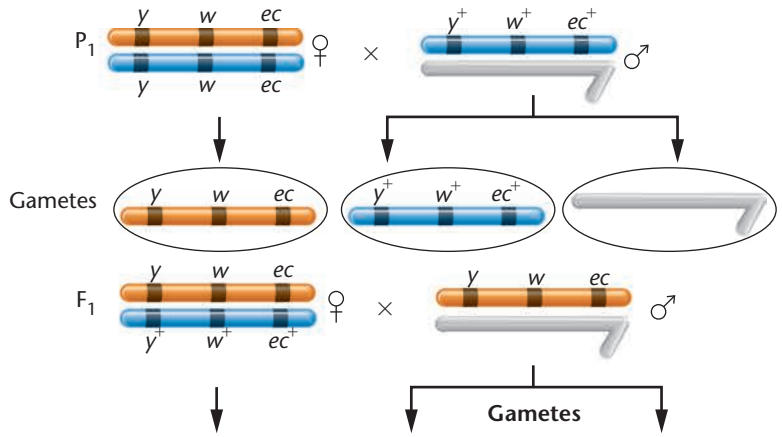
To execute a successful mapping cross, three criteria must be met:

1. The genotype of the organism producing the crossover gametes must be heterozygous at all loci under consideration.
2. The cross must be constructed so that genotypes of all gametes can be determined accurately by observing the phenotypes of the resulting offspring. This is necessary because the gametes and their genotypes can never be observed directly. To overcome this problem, each phenotypic class must reflect the genotype of the gametes of the parents producing it.
3. A sufficient number of offspring must be produced in the mapping experiment to recover a representative sample of all crossover classes.

These criteria are met in the three-point mapping cross from *Drosophila* shown in **Figure 7–8**. In this cross, three X-linked recessive mutant genes—*yellow* body color (*y*), *white* eye color (*w*), and *echinus* eye shape (*ec*)—are considered. To diagram the cross, *we must assume some theoretical sequence, even though we do not yet know if it is correct.* In Figure 7–8, we initially assume the sequence of the three genes to be *y*–*w*–*ec*. If this assumption is incorrect, our analysis will demonstrate this and reveal the correct sequence.

In the P_1 generation, males hemizygous for all three wild-type alleles are crossed to females that are homozygous for all three recessive mutant alleles. Therefore, the P_1 males are wild type with respect to body color, eye color, and eye shape. They are said to have a *wild-type phenotype.* The females, on the other hand, exhibit the three mutant traits—yellow body color, white eyes, and echinus eye shape.

This cross produces an F_1 generation consisting of females that are heterozygous at all three loci and males that, because of the Y chromosome, are hemizygous for the three mutant alleles. Phenotypically, all F_1 females are wild type, while all F_1 males are yellow, white, and echinus. The genotype of the F_1 females fulfills the first criterion for mapping; that is, it is heterozygous at the three loci and can serve as the source of recombinant gametes generated by crossing over. Note that because of the genotypes of the P_1 parents, all three mutant alleles in the F_1 female are on one homolog and all three wild-type alleles are on the other homolog. With other females, *other arrangements are possible that could produce a heterozygous genotype.* For example, a



Origin of female gametes	Gametes			F_2 phenotype	Observed number	Category, total, and percentage
NCO	1	y w ec	y w ec / y w ec	y w ec	4685	**Non-crossover**
	2	y^+ w^+ ec^+	y^+ w^+ ec^+ / y w ec	y^+ w^+ ec^+	4759	9444 94.44%
SCO	3	y w^+ ec^+	y w^+ ec^+ / y w ec	y w^+ ec^+	80	**Single crossover between *y* and *w***
	4	y^+ w ec	y^+ w ec / y w ec	y^+ w ec	70	150 1.50%
SCO	5	y w ec^+	y w ec^+ / y w ec	y w ec^+	193	**Single crossover between *w* and *ec***
	6	y^+ w^+ ec	y^+ w^+ ec / y w ec	y^+ w^+ ec	207	400 4.00%
DCO	7	y w^+ ec	y w^+ ec / y w ec	y w^+ ec	3	**Double crossover between *y* and *w* and between *w* and *ec***
	8	y^+ w ec^+	y^+ w ec^+ / y w ec	y^+ w ec^+	3	6 0.06%

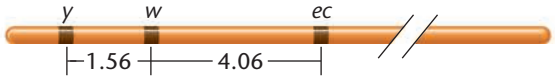


FIGURE 7–8 A three-point mapping cross involving the *yellow* (*y* or y^+), *white* (*w* or w^+), and *echinus* (*ec* or ec^+) genes in *Drosophila melanogaster*. NCO, SCO, and DCO refer to noncrossover, single-crossover, and double-crossover groups, respectively. Centromeres are not drawn on the chromosomes, and only two nonsister chromatids are initially shown in the left-hand column.

heterozygous female could have the *y* and *ec* mutant alleles on one homolog and the *w* allele on the other. This would occur if, in the P_1 cross, one parent was yellow, echinus and the other parent was white.

In our cross, the second criterion is met by virtue of the gametes formed by the F_1 males. Every gamete contains either an X chromosome bearing the three mutant alleles or a Y chromosome, which is genetically inert for the three loci being considered. Whichever type participates in fertilization, the genotype of the gamete produced by the F_1 female will be expressed phenotypically in the F_2 male and female offspring derived from it. Thus, all F_1 noncrossover and crossover gametes can be detected by observing the F_2 phenotypes.

With these two criteria met, we can now construct a chromosome map from the crosses shown in Figure 7–8. First, we determine which F_2 phenotypes correspond to the various noncrossover and crossover categories. To determine the noncrossover F_2 phenotypes, we must identify individuals derived from the parental gametes formed by the F_1 female. Each such gamete contains an X chromosome *unaffected by crossing over.* As a result of segregation, approximately equal proportions of the two types of gametes and, subsequently, the F_2 phenotypes, are produced. Because they derive from a heterozygote, the genotypes of the two parental gametes and the resultant F_2 phenotypes complement one another. For example, if one is wild type, the other is completely mutant. This is the case in the cross being considered. In other situations, if one chromosome shows one mutant allele, the second chromosome shows the other two mutant alleles, and so on. They are therefore called **reciprocal classes** of gametes and phenotypes.

The two noncrossover phenotypes are most easily recognized because *they exist in the greatest proportion.* Figure 7–8 shows that gametes 1 and 2 are present in the greatest numbers. Therefore, flies that express yellow, white, and echinus phenotypes and flies that are normal (or wild type) for all three characters constitute the noncrossover category and represent 94.44 percent of the F_2 offspring.

The second category that can be easily detected is represented by the double-crossover phenotypes. Because of their low probability of occurrence, *they must be present in the least numbers.* Remember that this group represents two independent but simultaneous single-crossover events. Two reciprocal phenotypes can be identified: gamete 7, which shows the mutant traits yellow, echinus but normal eye color; and gamete 8, which shows the mutant trait white but normal body color and eye shape. Together these double-crossover phenotypes constitute only 0.06 percent of the F_2 offspring.

The remaining four phenotypic classes represent two categories resulting from single crossovers. Gametes 3 and 4, reciprocal phenotypes produced by single-crossover events occurring between the *yellow* and *white* loci, are equal to 1.50 percent of the F_2 offspring; gametes 5 and 6, constituting 4.00 percent of the F_2 offspring, represent the reciprocal phenotypes resulting from single-crossover events occurring between the *white* and *echinus* loci.

The map distances separating the three loci can now be calculated. The distance between *y* and *w* or between *w* and *ec* is equal to the percentage of all detectable exchanges occurring between them. For any two genes under consideration, this includes all appropriate single crossovers as well as all double crossovers. *The latter are included because they represent two simultaneous single crossovers.* For the *y* and *w* genes, this includes gametes 3, 4, 7, and 8, totaling 1.50% + 0.06%, or 1.56 mu. Similarly, the distance between *w* and *ec* is equal to the percentage of offspring resulting from an exchange between these two loci: gametes 5, 6, 7, and 8, totaling 4.00% + 0.06%, or 4.06 mu. The map of these three loci on the X chromosome is shown at the bottom of Figure 7–8.

Determining the Gene Sequence

In the preceding example, the sequence (or order) of the three genes along the chromosome was assumed to be *y–w–ec*. Our analysis shows this sequence to be consistent with the data. However, in most mapping experiments the gene sequence is not known, and this constitutes another variable in the analysis. In our example, had the gene sequence been unknown, it could have been determined using a straightforward method.

This method is based on the fact that there are only three possible arrangements, each containing one of the three genes between the other two:

(I)	*w–y–ec*	(*y* in the middle)
(II)	*y–ec–w*	(*ec* in the middle)
(III)	*y–w–ec*	(*w* in the middle)

Use the following steps during your analysis to determine the gene order:

1. Assuming any one of the three orders, first determine the *arrangement of alleles* along each homolog of the heterozygous parent giving rise to noncrossover and crossover gametes (the F_1 female in our example).
2. Determine whether a double-crossover event occurring within that arrangement will produce the *observed double-crossover phenotypes.* Remember that these phenotypes occur least frequently and are easily identified.
3. If this order does not produce the predicted phenotypes, try each of the other two orders. One must work!

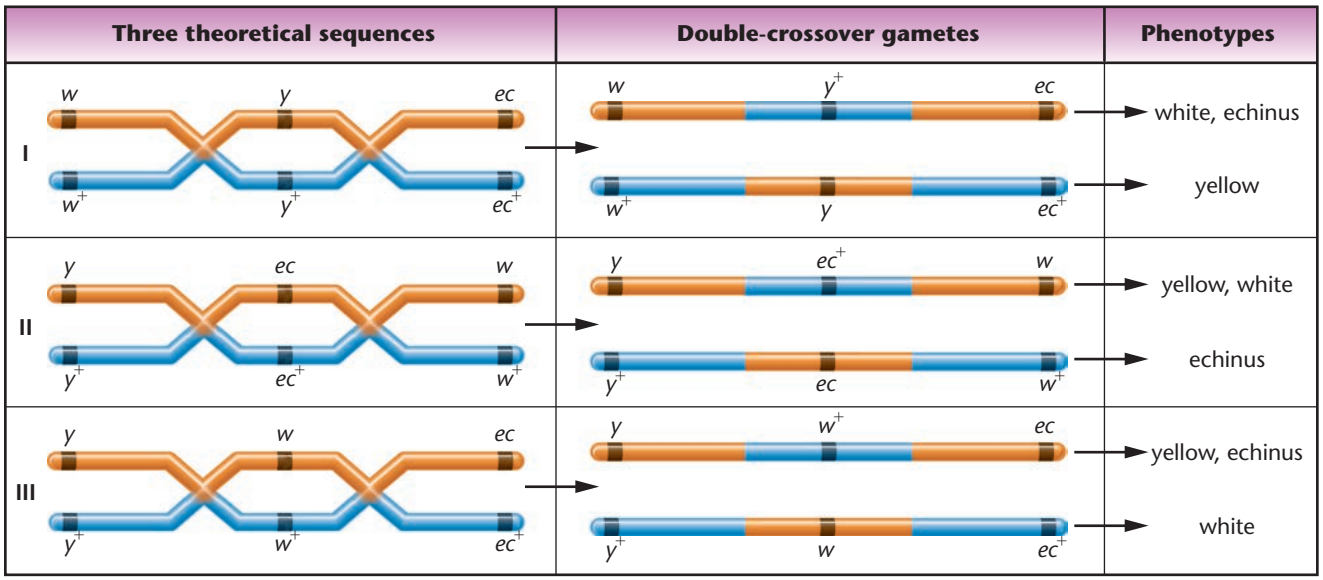


FIGURE 7–9 The three possible sequences of the *white, yellow*, and *echinus* genes, the results of a double crossover in each case, and the resulting phenotypes produced in a testcross. For simplicity, the two noncrossover chromatids of each tetrad are omitted.

In **Figure 7–9**, the above steps are applied to each of the three possible arrangements (I, II, and III above). A full analysis can proceed as follows:

1. Assuming that *y* is between *w* and *ec,* arrangement I of alleles along the homologs of the F_1 heterozygote is

$$\frac{w \quad y \quad ec}{w^+ \quad y^+ \quad ec^+}$$

We know this because of the way in which the P_1 generation was crossed: The P_1 female contributes an X chromosome bearing the *w, y,* and *ec* alleles, while the P_1 male contributes an X chromosome bearing the w^+, y^+, and ec^+ alleles.

2. A double crossover within that arrangement yields the following gametes

$$\underline{w \quad y^+ \quad ec} \quad \text{and} \quad \underline{w^+ \quad y \quad ec^+}$$

Following fertilization, if *y* is in the middle, the F_2 double-crossover phenotypes will correspond to these gametic genotypes, yielding offspring that express the white, echinus phenotype and offspring that express the yellow phenotype. Instead, determination of the actual double-crossover phenotypes reveals them to be yellow, echinus flies and white flies. *Therefore, our assumed order is incorrect.*

3. If we consider arrangement II with the ec/ec^+ alleles in the middle or arrangement III with the w/w^+ alleles in the middle

$$\text{(II)}\ \frac{w \quad ec \quad w}{y^+ \quad ec^+ \quad w^+} \quad \text{or} \quad \text{(III)}\ \frac{y \quad w \quad ec}{y^+ \quad w^+ \quad ec^+}$$

we see that arrangement II again provides *predicted* double-crossover phenotypes that *do not* correspond to the *actual* (observed) double-crossover phenotypes. The predicted phenotypes are yellow, white flies and echinus flies in the F_2 generation. *Therefore, this order is also incorrect.* However, arrangement III produces the observed phenotypes—yellow, echinus flies and white flies. *Therefore, this arrangement, with the w gene in the middle, is correct.*

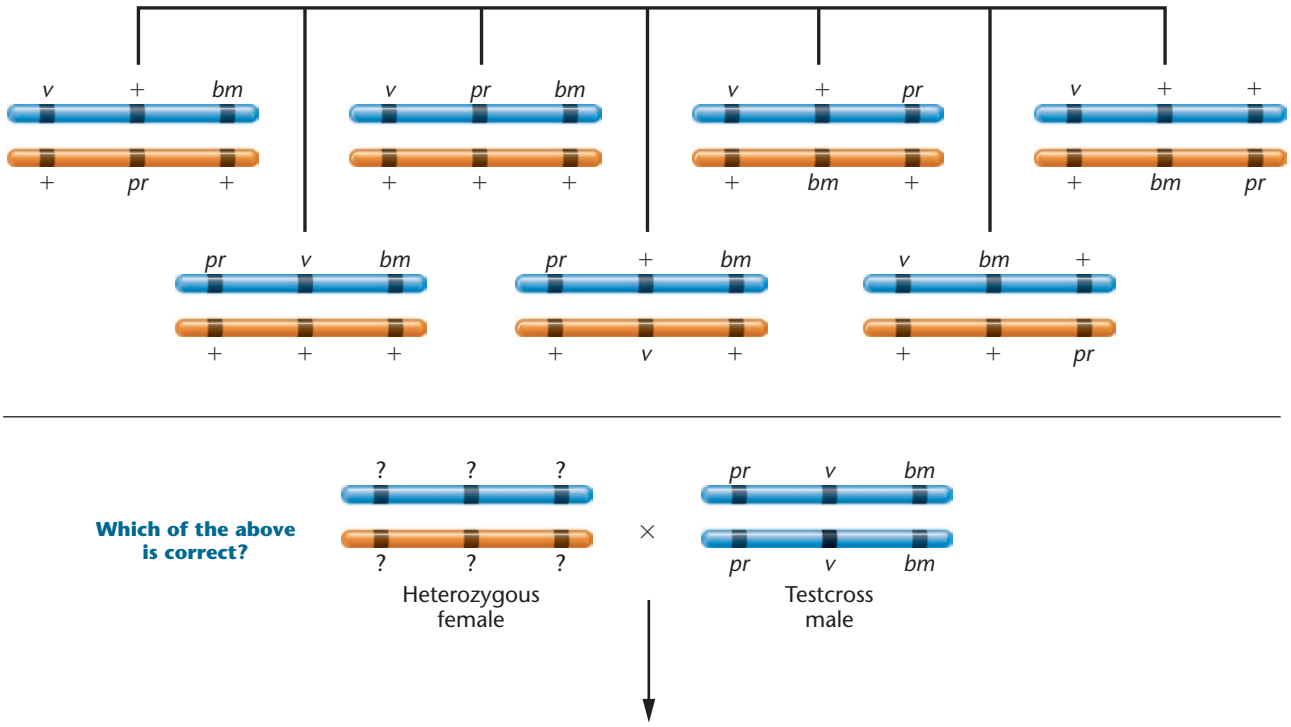
To summarize, first determine the arrangement of alleles on the homologs of the heterozygote yielding the crossover gametes by locating the reciprocal noncrossover phenotypes. Then, test each of three possible orders to determine which yields the observed double-crossover phenotypes—*the one that does so represents the correct order.*

Solving an Autosomal Mapping Problem

Having established the basic principles of chromosome mapping, we will now consider a related problem in maize (corn). This analysis differs from the preceding example in two ways. First, the previous mapping cross involved X-linked genes. Here, we consider autosomal genes. Second, in the discussion of this cross we have changed our use of symbols, as first suggested in Chapter 4. Instead of using the gene symbols and superscripts (e.g., bm^+, v^+, and pr^+), we simply use + to denote each wild-type allele. This system is easier to manipulate but requires a better understanding of mapping procedures.

When we look at three autosomally linked genes in maize, the experimental cross must still meet the same three criteria we established for the X-linked genes in

(a) Some possible allele arrangements and gene sequences in a heterozygous female



(b) Actual results of mapping cross*

Phenotypes of offspring	Number	Total and percentage	Exchange classification
+ *v* *bm*	230	467 42.1%	Noncrossover (NCO)
pr + +	237		
+ + *bm*	82	161 14.5%	Single crossover (SCO)
pr *v* +	79		
+ *v* +	200	395 35.6%	Single crossover (SCO)
pr + *bm*	195		
pr *v* *bm*	44	86 7.8%	Double crossover (DCO)
+ + +	42		

* The sequence *pr* – *v* – *bm* may or may not be correct.

FIGURE 7–10 (a) Some possible allele arrangements and gene sequences in a heterozygous female. The data from a three-point mapping cross, depicted in (b), where the female is testcrossed, provide the basis for determining which combination of arrangement and sequence is correct. [See Figure 7–11(d).]

Drosophila: (1) One parent must be heterozygous for all traits under consideration; (2) the gametic genotypes produced by the heterozygote must be apparent from observing the phenotypes of the offspring; and (3) a sufficient sample size must be available for complete analysis.

In maize, the recessive mutant genes *brown* midrib (*bm*), *virescent* seedling (*v*), and *purple* aleurone (*pr*) are linked on chromosome 5. Assume that a female plant is known to be heterozygous for all three traits, but we do not know (1) the arrangement of the mutant alleles on the maternal and paternal homologs of this heterozygote, (2) the sequence of genes, or (3) the map distances between the genes. What genotype must the male plant have to allow successful mapping? To meet the second criterion, the male must be homozygous for all three recessive mutant alleles. Otherwise, offspring of this cross showing a given phenotype might represent more than one genotype, making accurate mapping impossible.

Figure 7–10 diagrams this cross. As shown, we know neither the arrangement of alleles nor the sequence of loci

in the heterozygous female. Several possibilities are shown, but we have yet to determine which is correct. We don't know the sequence in the testcross male parent either, and so we must designate it randomly. Note that we have initially placed *v* in the middle. *This may or may not be correct.*

The offspring are arranged in groups of two for each pair of reciprocal phenotypic classes. The two members of each reciprocal class are derived from no crossing over (NCO), one of two possible single-crossover events (SCO), or a double crossover (DCO).

To solve this problem, refer to Figures 7–10 and 7–11 as you consider the following questions.

1. *What is the correct heterozygous arrangement of alleles in the female parent?*

   Determine the two noncrossover classes, those that occur with the highest frequency. In this case, they are $+\ v\ bm$ and $pr\ +\ +$. Therefore, the alleles on the homologs of the female parent must be arranged as shown in **Figure 7–11(a)**. These homologs segregate into gametes, unaffected by any recombination event. Any other arrangement of alleles will not yield the observed noncrossover classes. (Remember that $+\ v\ bm$ is equivalent to $pr^+\ v\ bm$ and that $pr\ +\ +$ is equivalent to $pr\ v^+\ bm^+$.)

2. *What is the correct sequence of genes?*

   We know that the arrangement of alleles is

   $$\frac{+\quad v\quad bm}{pr\quad +\quad +}$$

   But is the gene sequence correct? That is, will a double-crossover event yield the observed double-crossover phenotypes after fertilization? *Observation shows that it will not* [**Figure 7–11(b)**]. Now try the other two orders [**Figure 7–11(c)** and **(d)**] *maintaining the same arrangement of alleles:*

   $$\frac{+\quad bm\quad v}{pr\quad +\quad +} \quad \text{or} \quad \frac{v\quad +\quad bm}{+\quad pr\quad +}$$

   Only the order on the right yields the observed double-crossover gametes [Figure 7–11(d)]. Therefore, the *pr* gene is in the middle. From this point on, work the problem using this arrangement and sequence, with the *pr* locus in the middle.

3. *What is the distance between each pair of genes?*

   Having established the sequence of loci as *v*–*pr*–*bm*, we can determine the distance between *v* and *pr* and between *pr* and *bm*. Remember that the map distance between two genes is calculated on the basis of all detectable recombination events occurring between them. This includes both single- and double-crossover events.

Figure 7–11(e) shows that the phenotypes $v\ pr\ +$ and $+\ +\ bm$ result from single crossovers between the *v* and *pr* loci, accounting for 14.5 percent of the offspring [according to data in Figure 7–10(b)]. By adding the percentage of double crossovers (7.8 percent) to the number obtained for single crossovers, the total distance between the *v* and *pr* loci is calculated to be 22.3 mu.

Figure 7–11(f) shows that the phenotypes $v\ +\ +$ and $+\ pr\ bm$ result from single crossovers between the *pr* and *bm* loci, totaling 35.6 percent. Added to the double crossovers (7.8 percent), the distance between *pr* and *bm* is calculated to be 43.4 mu. The final map for all three genes in this example is shown in **Figure 7–11(g)**.

NOW SOLVE THIS

7–3 In *Drosophila*, a heterozygous female for the X-linked recessive traits *a*, *b*, and *c* was crossed to a male that phenotypically expressed *a*, *b*, and *c*. The offspring occurred in the following phenotypic ratios.

+	*b*	*c*	460
a	+	+	450
a	*b*	*c*	32
+	+	+	38
a	+	*c*	11
+	*b*	+	9

No other phenotypes were observed.

(a) What is the genotypic arrangement of the alleles of these genes on the X chromosome of the female?
(b) Determine the correct sequence and construct a map of these genes on the X chromosome.
(c) What progeny phenotypes are missing? Why?

HINT: *This problem involves a three-point mapping experiment where only six phenotypic categories are observed, even though eight categories are typical of such a cross. The key to its solution is to be aware that if the distances between the loci are relatively small, the sample size may be too small for the predicted number of double crossovers to be recovered, even though reciprocal pairs of single crossovers are seen. You should write the missing gametes down as double crossovers and record zeros for their frequency of appearance.*

7.4 As the Distance between Two Genes Increases, Mapping Estimates Become More Inaccurate

So far, we have assumed that crossover frequencies are directly proportional to the distance between any two loci along the chromosome. However, it is not always possible

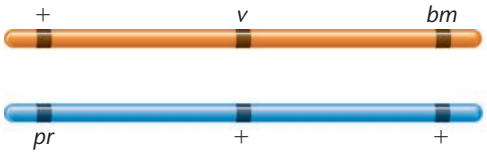
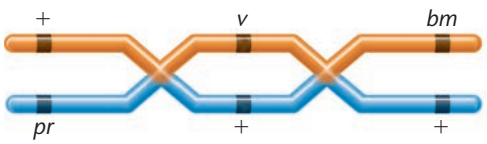
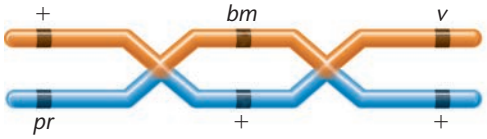
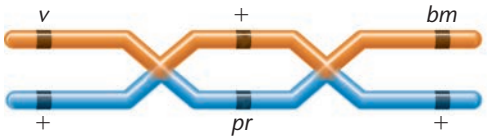
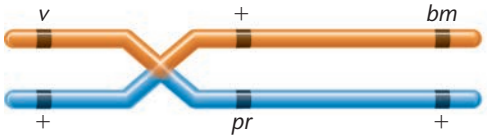
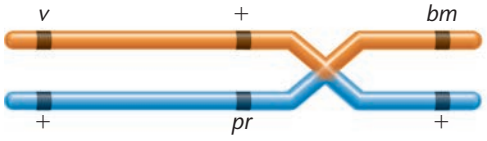
Possible allele arrangements and sequences	Testcross phenotypes	Explanation
(a) 	+ v bm and pr + +	Noncrossover phenotypes provide the basis for determining the correct arrangement of alleles on homologs
(b) 	+ + bm and pr v +	Expected double-crossover phenotypes if *v* is in the middle
(c) 	+ + v and pr bm +	Expected double-crossover phenotypes if *bm* is in the middle
(d) 	v pr bm and + + +	Expected double-crossover phenotypes if *pr* is in the middle **(This is the *actual situation*.)**
(e) 	v pr + and + + bm	Given that (a) and (d) are correct, single-crossover phenotypes when exchange occurs between *v* and *pr*
(f) 	v + + and + pr bm	Given that (a) and (d) are correct, single-crossover phenotypes when exchange occurs between *pr* and *bm*
(g) **Final map** 		

FIGURE 7–11 Steps utilized in producing a map of the three genes in the cross in Figure 7–10, where neither the arrangement of alleles nor the sequence of genes in the heterozygous female parent is known.

to detect all crossover events. A case in point is a double exchange that occurs between the two loci in question. As shown in **Figure 7–12(a)**, if a double exchange occurs, the original arrangement of alleles on each nonsister homolog is recovered. Therefore, even though crossing over has occurred, it is impossible to detect. This phenomenon is true for all even-numbered exchanges between two loci.

Furthermore, as a result of complications posed by **multiple-strand exchanges,** mapping determinations usually underestimate the actual distance between two genes. The farther apart two genes are, the greater the probability that undetected crossovers will occur. While the discrepancy is minimal for two genes relatively close together, the degree of inaccuracy increases as the distance increases, as shown in the graph of map distance versus recombination frequency in **Figure 7–12(b)**. There, the theoretical frequency where a direct correlation between recombination and map distance exists is contrasted with the actual frequency observed as the distance between two genes increases. The most accurate maps are constructed from experiments where genes are relatively close together.

Interference and the Coefficient of Coincidence

As shown in our maize example, we can predict the expected frequency of multiple exchanges, such as double crossovers, once the distance between genes is established.

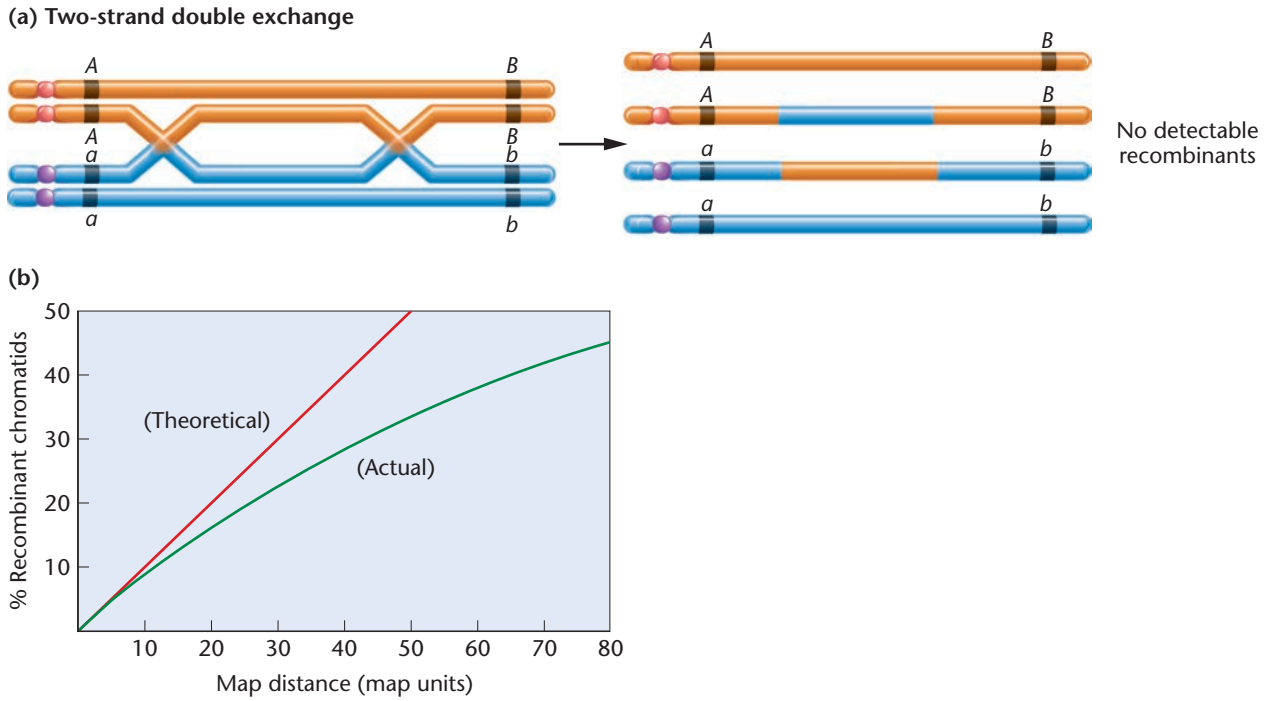


FIGURE 7–12 (a) A double crossover is undetected because no rearrangement of alleles occurs. (b) The theoretical and actual percentage of recombinant chromatids versus map distance. The straight line shows the theoretical relationship if a direct correlation between recombination and map distance exists. The curved line is the actual relationship derived from studies of *Drosophila, Neurospora,* and *Zea mays.*

For example, in the maize cross, the distance between *v* and *pr* is 22.3 mu, and the distance between *pr* and *bm* is 43.4 mu. If the two single crossovers that make up a double crossover occur independently of one another, we can calculate the expected frequency of double crossovers (DCO_{exp}):

$$DCO_{exp} = (0.223) \times (0.434) = 0.097 = 9.7\%$$

Often in mapping experiments, the observed DCO frequency is less than the expected number of DCOs. In the maize cross, for example, only 7.8 percent DCOs are observed when 9.7 percent are expected. **Interference (*I*)**, the phenomenon through which a crossover event in one region of the chromosome inhibits a second event in nearby regions, causes this reduction.

To quantify the disparities that result from interference, we calculate the **coefficient of coincidence (*C*)**:

$$C = \frac{\text{Observed DCO}}{\text{Expected DCO}}$$

In the maize cross, we have

$$C = \frac{0.078}{0.097} = 0.804$$

Once we have found *C,* we can quantify interference using the simple equation

$$I = 1 - C$$

In the maize cross, we have

$$I = 1.000 - 0.804 = 0.196$$

If interference is complete and no double crossovers occur, then $I = 1.0$. If fewer DCOs than expected occur, I is a positive number and positive interference has occurred. If more DCOs than expected occur, I is a negative number and negative interference has occurred. In the maize example, I is a positive number (0.196), indicating that 19.6 percent fewer double crossovers occurred than expected.

Positive interference is most often the rule in eukaryotic systems. In general, the closer genes are to one another along the chromosome, the more positive interference occurs. In fact, interference in *Drosophila* is often complete within a distance of 10 mu, and no multiple crossovers are recovered. This observation suggests that physical constraints preventing the formation of closely aligned chiasmata contribute to interference. This interpretation is consistent with the finding that interference decreases as the genes in question are located farther apart. In the maize cross in Figures 7–10 and 7–11, the three genes are relatively far apart, and 80 percent of the expected double crossovers are observed.

ESSENTIAL POINT

Interference describes the extent to which a crossover in one region of a chromosome influences the occurrence of a crossover in an adjacent region of the chromosome and is quantified by calculating the coefficient of coincidence (*C*). ■

EVOLVING CONCEPT OF THE GENE

Based on the gene-mapping studies in *Drosophila* and many other organisms from the 1920s through the mid-1950s, geneticists regarded genes as hereditary units organized in a specific sequence along chromosomes, between which recombination could occur. Genes were thus viewed as indivisible "beads on a string." ■

7.5 Chromosome Mapping Is Now Possible Using DNA Markers and Annotated Computer Databases

Although traditional methods based on recombination analysis have produced detailed chromosomal maps in several organisms, such maps in other organisms (including humans) that do not lend themselves to such studies are greatly limited. Fortunately, the development of technology allowing direct analysis of DNA has greatly enhanced mapping in those organisms. We will address this topic using humans as an example.

Progress has initially relied on the discovery of **DNA markers** that have been identified during recombinant DNA and genomic studies. These markers are short segments of DNA whose sequence and location are known, making them useful *landmarks* for mapping purposes. The analysis of human genes in relation to these markers has extended our knowledge of the location within the genome of countless genes, which is the ultimate goal of mapping.

The earliest examples are the DNA markers referred to as **restriction fragment length polymorphisms** (**RFLPs**) (see Chapter 19) and **microsatellites** (see Chapter 11). RFLPS are polymorphic sites generated when specific DNA sequences are recognized and cut by restriction enzymes. Microsatellites are short repetitive sequences that are found throughout the genome, and they vary in the number of repeats at any given site. For example, the two-nucleotide sequence CA is repeated 5–50 times per site [$(CA)_n$] and appears throughout the genome approximately every 10,000 bases, on average. Microsatellites may be identified not only by the number of repeats but by the DNA sequences that flank them. More recently, variation in single nucleotides (called **single-nucleotide polymorphisms** or **SNPs**) has been utilized. Found throughout the genome, up to several million of these variations may be screened for an association with a disease or trait of interest, thus providing geneticists with a means to identify and locate related genes.

Cystic fibrosis offers an early example of a gene located by using DNA markers. It is a life-shortening autosomal recessive exocrine disorder resulting in excessive, thick mucus that impedes the function of organs such as the lung and pancreas. After scientists established that the gene causing this disorder is located on chromosome 7, they were then able to pinpoint its exact location on the long arm (the q arm) of that chromosome.

In 2007, using SNPs as DNA markers, associations between 24 genomic locations were established with seven common human diseases: *Type 1* (insulin dependent) and *Type 2 diabetes*, *Crohn disease* (inflammatory bowel disease), *hypertension*, *coronary artery disease*, *bipolar* (manic-depressive) *disorder*, and *rheumatoid arthritis.* In each case, an inherited susceptibility effect was mapped to a specific location on a specific chromosome within the genome. In some cases, this either confirmed or led to the identification of a specific gene involved in the cause of the disease.

During the past 15 years or so, dramatic improvements in DNA sequencing technology have resulted in a proliferation of **sequence maps** for humans and many other species. Sequence maps provide the finest level of mapping detail because they pinpoint the nucleotide sequence of genes (and noncoding sequences) on a chromosome. The Human Genome Project resulted in sequence maps for all human chromosomes, providing an incredible level of detail about human gene sequences, the specific location of genes on a chromosome, and the proximity of genes and noncoding sequences to each other, among other details. For instance, when human chromosome sequences were analyzed by software programs, an approach called **bioinformatics,** to be discussed later in the text (see Chapter 19), geneticists could utilize such data to map possible protein-coding sequences in the genome. This led to the identification of thousands of potential genes that were previously unknown.

The many Human Genome Project databases that are now available make it possible to map genes along a human chromosome in base-pair distances rather than recombination frequency. This distinguishes what is referred to as a *physical map* of the genome from the *genetic maps* described above. When the genome sequence of a species is available, mapping by linkage or other genetic mapping approaches becomes obsolete.

ESSENTIAL POINT

Human linkage studies have been enhanced by the use of newly discovered molecular DNA markers. ■

7.6 Other Aspects of Genetic Exchange

Careful analysis of crossing over during gamete formation allows us to construct chromosome maps in many organisms. However, we should not lose sight of the real biological significance of crossing over, which is to generate genetic variation in gametes and, subsequently, in the offspring derived from the resultant eggs and sperm. Many unanswered questions remain, which we consider next.

Crossing Over—A Physical Exchange between Chromatids

Once genetic mapping was understood, it was of great interest to investigate the relationship between chiasmata observed in meiotic prophase I and crossing over. Are chiasmata visible manifestations of crossover events? If so, then crossing over in higher organisms appears to result from an actual physical exchange between homologous chromosomes. That this is the case was demonstrated independently in the 1930s by Harriet Creighton and Barbara McClintock in *Zea mays* and by Curt Stern in *Drosophila.*

Since the experiments are similar, we will consider only the work with maize. Creighton and McClintock studied two linked genes on chromosome 9. At one locus, the alleles *colorless* (*c*) and *colored* (*C*) control endosperm coloration. At the other locus, the alleles *starchy* (*Wx*) and *waxy* (*wx*) control the carbohydrate characteristics of the endosperm. The maize plant studied is heterozygous at both loci. The key to this experiment is that one of the homologs contains two unique cytological markers. The markers consist of a densely stained knob at one end of the chromosome and a translocated piece of another chromosome (8) at the other end. The arrangements of alleles and cytological markers can be detected cytologically and are shown in **Figure 7–13** .

Creighton and McClintock crossed this plant to one homozygous for the *colored* allele (*c*) and heterozygous for the endosperm alleles. They obtained a variety of different phenotypes in the offspring, but they were most interested in a crossover result involving the chromosome with the unique cytological markers. They examined the chromosomes of this plant with the colorless, waxy phenotype (Case I in Figure 7–13) for the presence of the cytological markers. If physical exchange between homologs accompanies genetic crossing over, the translocated chromosome will still be present, but the knob will not—this is exactly what happened. In a second plant (Case II), the phenotype colored, starchy should result from either nonrecombinant gametes or crossing over. Some of the plants then ought to contain chromosomes with the dense knob but not the translocated chromosome. This condition was also found, and the conclusion that a physical exchange takes place was again supported. Along with Stern's findings with *Drosophila,* this work clearly established that crossing over has a cytological basis.

ESSENTIAL POINT

Cytological investigations of both maize and *Drosophila* reveal that crossing over involves a physical exchange of segments between nonsister chromatids. ■

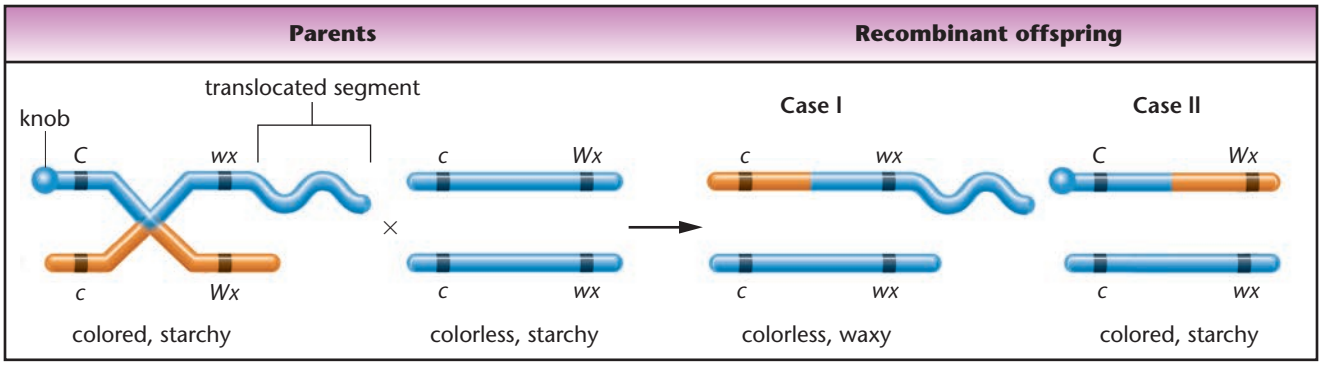


FIGURE 7–13 The phenotypes and chromosome compositions of parents and recombinant offspring in Creighton and McClintock's experiment in maize. The knob and translocated segment served as cytological markers, which established that crossing over involves an actual exchange of chromosome arms.

Sister Chromatid Exchanges between Mitotic Chromosomes

Considering that crossing over occurs between synapsed homologs in meiosis, we might ask whether a similar physical exchange occurs between homologs during mitosis. While homologous chromosomes do not usually pair up or synapse in somatic cells (*Drosophila* is an exception), each individual chromosome in prophase and metaphase of mitosis consists of two identical sister chromatids, joined at a common centromere. Surprisingly, several experimental approaches have demonstrated that reciprocal exchanges similar to crossing over occur between sister chromatids. These **sister chromatid exchanges (SCEs)** do not produce new allelic combinations, but evidence is accumulating that attaches significance to these events.

Identification and study of SCEs are facilitated by several modern staining techniques. In one technique, cells replicate for two generations in the presence of the thymidine analog **bromodeoxyuridine (BrdU).** Following two rounds of replication, each pair of sister chromatids has one member with one strand of DNA "labeled" with BrdU and the other member with both strands labeled with BrdU. Using a differential stain, chromatids with the analog in both strands stain less brightly than chromatids with BrdU in only one strand. As a result, SCEs are readily detectable if they occur. In **Figure 7–14**, numerous instances of SCE events are clearly evident. These sister chromatids are sometimes referred to as **harlequin chromosomes** because of their patchlike appearance.

The significance of SCEs is still uncertain, but several observations have generated great interest in this phenomenon. We know, for example, that agents that induce chromosome damage (viruses, X rays, ultraviolet light, and certain chemical mutagens) increase the frequency of SCEs. The frequency of SCEs is also elevated in **Bloom syndrome,** a human disorder caused by a mutation in the *BLM* gene on chromosome 15. This rare, recessively inherited disease is characterized by prenatal and postnatal retardation of growth, a great sensitivity of the facial skin to the sun, immune deficiency, a predisposition to malignant and benign tumors, and abnormal behavior patterns. The chromosomes from cultured leukocytes, bone marrow cells, and fibroblasts derived from homozygotes are very fragile and unstable compared to those of homozygous and heterozygous normal individuals. Increased breaks and rearrangements between nonhomologous chromosomes are observed in addition to excessive amounts of sister chromatid exchanges. Work by James German and colleagues suggests that the *BLM* gene encodes an enzyme called **DNA helicase**, which is best known for its role in DNA replication (see Chapter 10).

ESSENTIAL POINT

Recombination events between sister chromatids in mitosis, referred to as sister chromatid exchanges (SCEs), occur at an elevated frequency in the human disorder, Bloom syndrome. ■

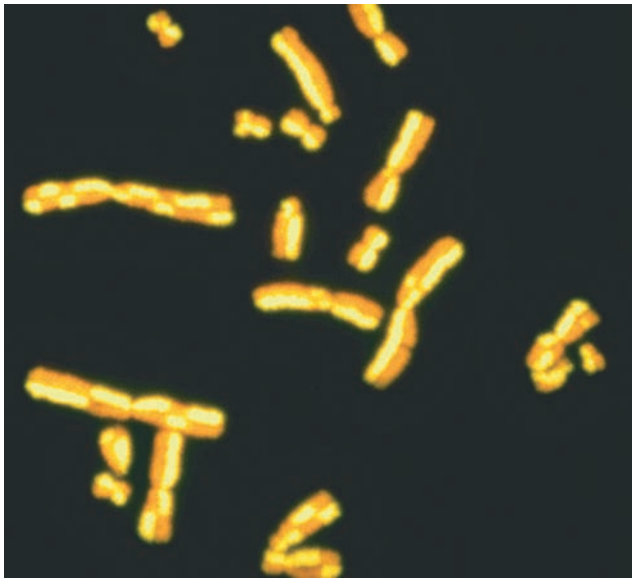
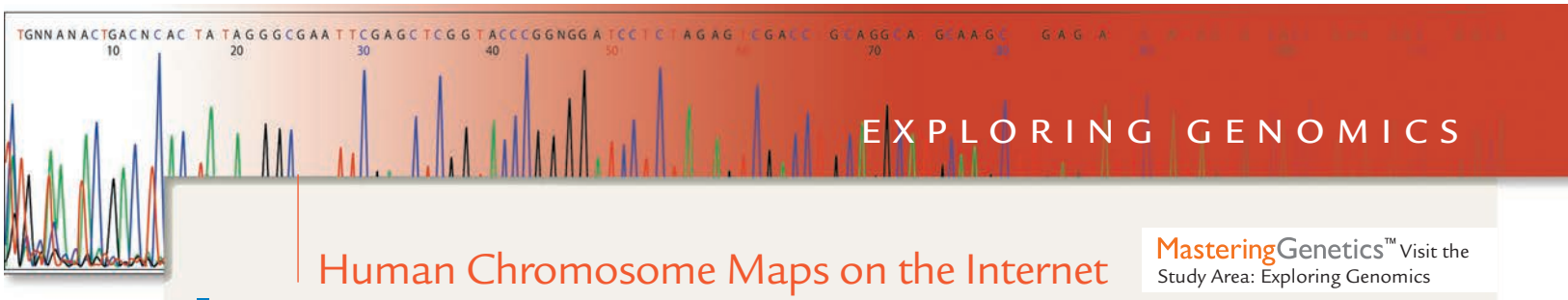


FIGURE 7–14 Demonstration of sister chromatid exchanges (SCEs) in mitotic chromosomes. Sometimes called harlequin chromosomes because of the alternating patterns they exhibit, sister chromatids containing the thymidine analog BrdU are seen to fluoresce *less* brightly where they contain the analog in both DNA strands than when they contain the analog in only one strand. These chromosomes were stained with 33258-Hoechst reagent and acridine orange and then viewed under fluorescence microscopy.



Human Chromosome Maps on the Internet

MasteringGenetics™ Visit the Study Area: Exploring Genomics

In this chapter we discussed how recombination data can be analyzed to develop chromosome maps based on linkage. Chromosome maps are increasingly being developed for many species using genomics techniques. As a result of the Human Genome Project, maps of human chromosomes are now freely available on the Internet. In this exercise we will explore the **National Center for Biotechnology Information (NCBI) Genes and Disease** Web site to learn more about human chromosome maps.

NCBI Genes and Disease

Here we explore the Genes and Disease site, which presents human chromosome maps that show the locations of specific disease genes.

1. Access the Genes and Disease site at http://www.ncbi.nlm.nih.gov/books/NBK22183/
2. Under contents, click on "chromosome map" to see a page with an image of a karyotype of human chromosomes. Click on a chromosome in the chromosome map image, scroll down the page to view a chromosome or click on a chromosome listed on the right side of the page. For example, click on chromosome 7. Notice that the number of genes on the chromosomes and the number of base pairs the chromosome contains are displayed above the image.
3. Look again at chromosome 7 and click the "MapViewer" link for the chromosome (just above the drawing) to learn more about a gene of interest.
4. Visit the **NCBI Map Viewer** homepage (http://www.ncbi.nlm.nih.gov/projects/mapview/) for an excellent database containing chromosome maps for a wide variety of different organisms.

CASE STUDY | Links to autism

As parents of an autistic child, a couple decided that entering a research study would not only educate them about their son's condition, but also help further research into this complex, behaviorally defined disorder. In an interview, researchers explained to the parents that autism results from the action of hundreds of genes and that no single gene accounts for more than a small percentage of cases. Recent studies have identified 18 genes that have a higher likelihood of involvement, referred to as candidate genes; three of these, on chromosomes 2, 7, and 14, are regarded as very strong candidate genes. Generally unaware of the principles of basic genetics, the couple asked a number of interesting questions. If you were the interviewer, how would you respond to them?

1. How might identification of a "candidate" gene be helpful in treating autism?
2. Because there are hundreds of genes involved, is it possible that our children might inherit several or even many mutant versions of these genes?
3. With such a complex genetic condition that may also depend on environmental factors, is there a way to calculate the risk that our next child will be autistic?
4. Is prenatal diagnosis for autism possible?

INSIGHTS AND SOLUTIONS

1. In rabbits, black color (*B*) is dominant to brown (*b*), while full color (*C*) is dominant to *chinchilla* (c^{ch}). The genes controlling these traits are linked. Rabbits that are heterozygous for both traits and express black, full color are crossed to rabbits that express brown, chinchilla with the following results:

31	brown, chinchilla	34	black, full
16	brown, full	19	black, chinchilla

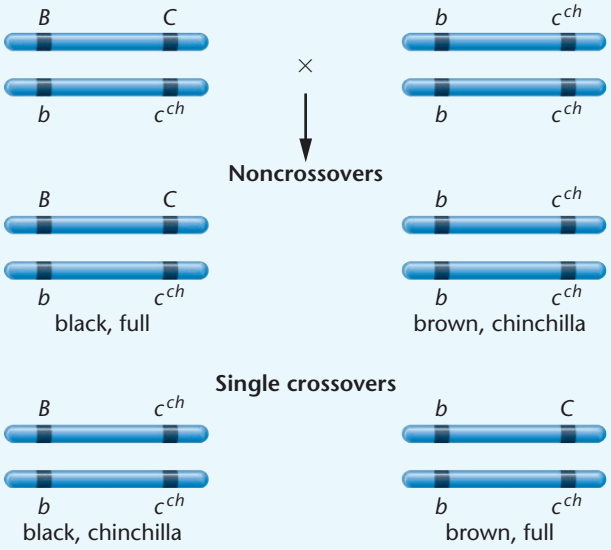
Determine the arrangement of alleles in the heterozygous parents and the map distance between the two genes.

Solution: This is a two-point map problem, where the two most prevalent reciprocal phenotypes are the noncrossovers. The less frequent reciprocal phenotypes arise from a single crossover. The arrangement of alleles is derived from the noncrossover phenotypes because they enter gametes intact.

(continued)

Insights and Solutions—continued

The single crossovers give rise to 35 of 100 offspring (35 percent). Therefore, the distance between the two genes is 35 mu.



2. Examine the a set of 3 point mapping data from *Drosophila* involving two dominant mutations [*Stubble* (*Sb*) and *Lyra* (*Ly*)] and one recessive mutation [*bright* (*br*)]. Identify the categories of data below (1–8) that are essential in establishing: (a) the arrangement of alleles in the crossover parent; (b) the sequence of the three genes along the chromosome; and (c) the interlocus distance between the genes.

Phenotype				Number
(1)	*Ly*	*Sb*	*br*	404
(2)	+	+	+	422
(3)	*Ly*	+	+	18
(4)	+	*Sb*	*br*	16
(5)	*Ly*	+	*br*	75
(6)	+	*Sb*	+	59
(7)	*Ly*	*Sb*	+	4
(8)	+	+	*br*	2
				Total = 1000

Solution:

a. Categories 1 and 2, which represent the noncrossover gametes.
b. Categories 7 and 8, which represent the double crossover gametes.
c. Categories 3, 4, 5, 6, 7, and 8, all of which arise from crossover events.

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we focused on linkage, chromosomal mapping, and many associated phenomena. In the process, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions?
   (a) How was it established experimentally that the frequency of recombination (crossing over) between two genes is related to the distance between them along the chromosome?
   (b) How do we know that specific genes are linked on a single chromosome, in contrast to being located on separate chromosomes?
   (c) How do we know that crossing over results from a physical exchange between chromatids?
   (d) How do we know that sister chromatids undergo recombination during mitosis?

CONCEPT QUESTION

2. Review the Chapter Concepts list on page 136. Most of these center on the process of crossing over between linked genes. Write a short essay that discusses how crossing over can be detected and how the resultant data provide the basis of chromosome mapping. ■
3. Describe the cytological observation that suggests that crossing over occurs during the first meiotic prophase.
4. Why does more crossing over occur between two distantly linked genes than between two genes that are very close together on the same chromosome?
5. Why is a 50 percent recovery of single-crossover products the upper limit, even when crossing over *always* occurs between two linked genes?
6. Why are double-crossover events expected less frequently than single-crossover events?
7. Explain the meaning of the term *interference*.
8. What three essential criteria must be met in order to execute a successful mapping cross?
9. The genes *dumpy* wings (*dp*), *clot* eyes (*cl*), and *apterous* wings (*ap*) are linked on chromosome II of *Drosophila*. In a series of two-point mapping crosses, the genetic distances shown below were determined. What is the sequence of the three genes?

dp–ap	42
dp–cl	3
ap–cl	39

10. Colored aleurone in the kernels of corn is due to the dominant allele *R*. The recessive allele *r*, when homozygous, produces colorless aleurone. The plant color (not kernel color) is controlled by another gene with two alleles, *Y* and *y*. The dominant *Y* allele results in green color, whereas the homozygous presence of the recessive *y* allele causes the plant to appear yellow. In a testcross between a plant of unknown genotype and phenotype and a plant that is homozygous recessive for both traits, the following progeny were obtained:

colored, green	88
colored, yellow	12
colorless, green	8
colorless, yellow	92

Explain how these results were obtained by determining the exact genotype and phenotype of the unknown plant, including the precise association of the two genes on the homologs (i.e., the arrangement).

11. Phenotypically wild F_1 female *Drosophila*, whose mothers had light eyes (*lt*) and fathers had straw (*stw*) bristles, produced the following offspring when crossed with homozygous light-straw males:

Phenotype	*Number*
light-straw	22
wild	18
light	990
straw	970
Total	2000

Compute the map distance between the *light* and *straw* loci.

12. In a series of two-point map crosses involving five genes located on chromosome II in *Drosophila,* the following recombinant (single-crossover) frequencies were observed:

pr—adp	29
pr—vg	13
pr—c	21
pr—b	6
adp—b	35
adp—c	8
adp—vg	16
vg—b	19
vg—c	8
c—b	27

(a) If the *adp* gene is present near the end of chromosome II (locus 83), construct a map of these genes.
(b) In another set of experiments, a sixth gene (*d*) was tested against *b* and *pr,* and the results were $d - b = 17\%$ and $d - pr = 23\%$. Predict the results of two-point maps between *d* and *c, d* and *vg,* and *d* and *adp.*

13. Assume that investigators crossed a strain of flies carrying the dominant eye mutation Lobe on the second chromosome with a strain homozygous for the second chromosome recessive mutations smooth abdomen and straw body. The F_1 Lobe females were then backcrossed with homozygous smooth abdomen, straw body males, and the following phenotypes were observed:

smooth abdomen, straw body	820
Lobe	780
smooth abdomen, Lobe	42
straw body	58
smooth abdomen	148
Lobe, straw body	152

(a) Give the gene order and map units between these three loci.
(b) What is the coefficient of coincidence?

14. In *Drosophila,* a cross was made between females expressing the three X-linked recessive traits, *scute* bristles (*sc*), *sable* body (*s*), and *vermilion* eyes (*v*), and wild-type males. All females were wild type in the F_1, while all males expressed all three mutant traits. The cross was carried to the F_2 generation and 1000 offspring were counted, with the results shown in the following table. No determination of sex was made in the F_2 data. (a) Using proper nomenclature, determine the genotypes of the P_1 and F_1 parents. (b) Determine the sequence of the three genes and the map distance between them. (c) Are there more or fewer double crossovers than expected? (d) Calculate the coefficient of coincidence; does this represent positive or negative interference?

Phenotype			Offspring
sc	*s*	*v*	314
+	+	+	280
+	*s*	*v*	150
sc	+	+	156
sc	+	*v*	46
+	*s*	+	30
sc	*s*	+	10
+	+	*v*	14

15. A cross in *Drosophila* involved the recessive, X-linked genes *yellow* body (*y*), *white* eyes (*w*), and *cut* wings (*ct*). A yellow-bodied, white-eyed female with normal wings was crossed to a male whose eyes and body were normal, but whose wings were cut. The F_1 females were wild type for all three traits, while the F_1 males expressed the yellow-body, white-eye traits. The cross was carried to F_2 progeny, and only male offspring were tallied. On the basis of the data shown here, a genetic map was constructed. (a) Diagram the genotypes of the F_1 parents. (b) Construct a map, assuming that *w* is at locus 1.5 on the X chromosome. (c) Were any double-crossover offspring expected? (d) Could the F_2 female offspring be used to construct the map? Why or why not?

Phenotype			Male Offspring
y	+	*ct*	9
+	*w*	+	6
y	*w*	*ct*	90
+	+	+	95
+	+	*ct*	424
y	*w*	+	376
y	+	+	0
+	*w*	*ct*	0

16. *Drosophila melanogaster* has one pair of sex chromosomes (XX or XY) and three autosomes (chromosomes II, III, and IV). A genetics student discovered a male fly with very short (*sh*) legs. Using this male, the student was able to establish a pure-breeding stock of this mutant and found that it was recessive. She then incorporated the mutant into a stock containing the recessive gene *black* (*b,* body color, located on chromosome II) and the recessive gene *pink* (*p,* eye color, located on chromosome III). A female from the homozygous black, pink, short

stock was then mated to a wild-type male. The F_1 males of this cross were all wild type and were then backcrossed to the homozygous *b, p, sh* females. The F_2 results appeared as shown in the following table, and no other phenotypes were observed. (a) Based on these results, the student was able to assign *sh* to a linkage group (a chromosome). Determine which chromosome, and include step-by-step reasoning. (b) The student repeated the experiment, making the reciprocal cross: F_1 females backcrossed to homozygous *b, p, sh* males. She observed that 85 percent of the offspring fell into the given classes, but that 15 percent of the offspring were equally divided among *b*+*p*, *b*++, +*shp*, and +*sh*+ phenotypic males and females. How can these results be explained, and what information can be derived from these data?

Phenotype	Female	Male
wild	63	59
pink*	58	65
black, short	55	51
black, pink, short	69	60

*Pink indicates that the other two traits are wild type (normal). Similarly, black, short offspring are wild type for eye color.

17. Three loci, *mitochondrial malate dehydrogenase* that forms *a* and *b* (*MDHa, MDHb*), *glucouronidase* that forms *1* and *2* (*GUS1*, *GUS2*), and a *histone* gene that forms + and − (*H*+, *H*–), are located on chromosome #7 in humans. Assume that the *MDH* locus is at position 35, *GUS* at position 45, and *H* at position 75. A female whose mother was homozygous for *MDHa*, *GUS2*, and *H*+ and whose father was homozygous for *MDHb, GUS1*, and *H*– produces a sample of 1000 egg cells. Give the genotypes and expected numbers of the various types of cells she would produce. Assume no chromosomal interference.
18. A backcross was set up between two homozygous laboratory mouse strains *A* and *B*, with the F_1 backcrossed to *B*. The F_2 were typed using SNPs *x* and *y*, which varied between strains *A* and *B* (x^A, x^B, y^A, y^B). Out of 100 mice, 38 were $x^A y^A$, 40 were $x^B y^B$, 11 were $x^A y^B$, and 11 were $x^B y^A$. What is the genetic distance between SNPs *x* and *y*?
19. A female of genotype

$$\frac{a \quad b \quad c}{+ \quad + \quad +}$$

produces 100 meiotic tetrads. Of these, 68 show no crossover events. Of the remaining 32, 20 show a crossover between *a* and *b,* 10 show a crossover between *b* and *c,* and 2 show a double crossover between *a* and *b* and between *b* and *c.* Of the 400 gametes produced, how many of each of the eight different genotypes will be produced? Assuming the order *a*–*b*–*c* and the allele arrangement shown above, what is the map distance between these loci?
20. In a plant, fruit color is either red or yellow, and fruit shape is either oval or long. Red and oval are the dominant traits. Two plants, both heterozygous for these traits, were testcrossed, with the results shown in the following table. Determine the location of the genes relative to one another and the genotypes of the two parental plants.

	Progeny	
Phenotype	Plant A	Plant B
red, long	46	4
yellow, oval	44	6
red, oval	5	43
yellow, long	5	47
Total	100	100

21. In the fruit fly, *Drosophila melanogaster*, a spineless (no wing bristles) female fly is mated to a male that is claret (dark eyes) and hairless (no thoracic bristles). Phenotypically wild-type F_1 female progeny were mated to fully homozygous (mutant) males, and the following progeny (1000 total) were observed:

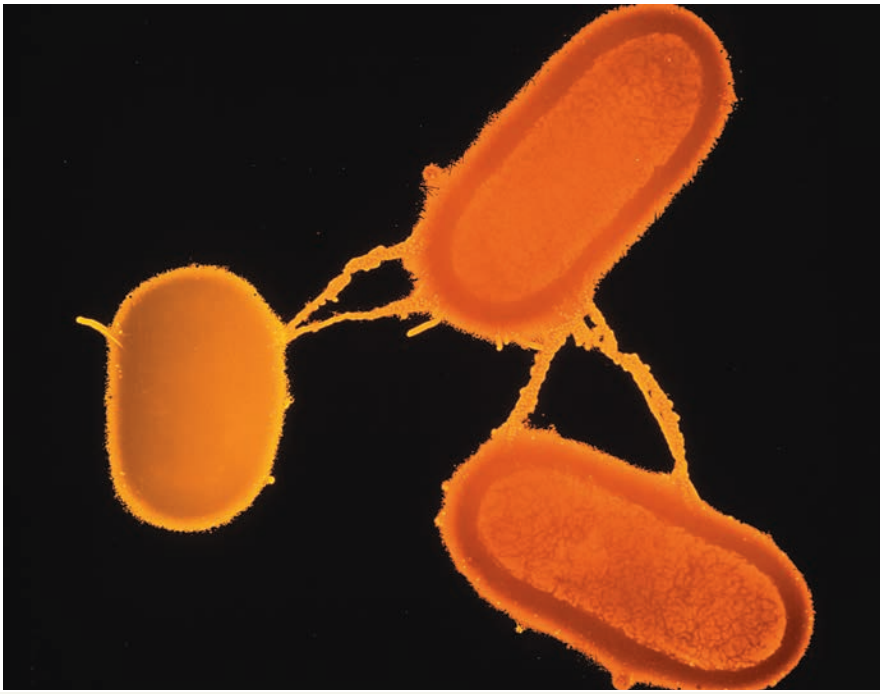
Phenotypes	*Number Observed*
spineless	321
wild	38
claret, spineless	130
claret	18
claret, hairless	309
hairless, claret, spineless	32
hairless	140
hairless, spineless	12

(a) Which gene is in the middle?
(b) With respect to the three genes mentioned in the problem, what are the genotypes of the homozygous parents used in making the phenotypically wild F_1 heterozygote?
(c) What are the map distances between the three genes? A correct formula with the values "plugged in" for each distance will be sufficient.
(d) What is the coefficient of coincidence? A correct formula with the values "plugged in" will be sufficient.
22. An organism of the genotype *AaBbCc* was testcrossed to a triply recessive organism (*aabbcc*). The genotypes of the progeny are in the following table.

AaBbCc	20	*AaBbcc*	20
aabbCc	20	*aabbcc*	20
AabbCc	5	*Aabbcc*	5
aaBbCc	5	*aaBbcc*	5

(a) Assuming simple dominance and recessiveness in each gene pair, if these three genes were all assorting independently, how many genotypic and phenotypic classes would result in the offspring, and in what proportion?
(b) Answer part (a) again, assuming the three genes are so tightly linked on a single chromosome that no crossover gametes were recovered in the sample of offspring.
(c) What can you conclude from the actual data about the location of the three genes in relation to one another?
23. Based on our discussion of the potential inaccuracy of mapping (see Figure 7–12), would you revise your answer to Problem 23(c)? If so, how?
24. In Creighton and McClintock's experiment demonstrating that crossing over involves physical exchange between chromosomes (see Section 7.6), explain the importance of the cytological markers (the translocated segment and the chromosome knob) in the experimental rationale.
25. Explain why restriction fragment length polymorphisms and microsatellites are important landmarks for mapping purposes.
26. Are sister chromatid exchanges effective in producing genetic variability in an individual? in the offspring of individuals?

8 Genetic Analysis and Mapping in Bacteria and Bacteriophages



An electron micrograph showing the sex pilus between two conjugating *Escherichia coli* cells.

CHAPTER CONCEPTS

- Bacterial genomes are most often contained in a single circular chromosome.
- Bacteria have developed numerous ways in which they can exchange and recombine genetic information between individual cells, including conjugation, transformation, and transduction.
- The ability to undergo conjugation and to transfer a portion or all of the bacterial chromosome from one cell to another is governed by the presence of genetic information contained in the DNA of a "fertility," or F factor.
- The F factor can exist autonomously in the bacterial cytoplasm as a plasmid, or it can integrate into the bacterial chromosome, where it facilitates the transfer of the host chromosome to the recipient cell, leading to genetic recombination.
- Genetic recombination during conjugation provides a means of mapping bacterial genes.
- Bacteriophages are viruses that have bacteria as their hosts. During infection of the bacterial host, bacteriophage DNA is injected into the host cell, where it is replicated and directs the reproduction of the bacteriophage.
- Rarely, following infection, bacteriophage DNA integrates into the host chromosome, becoming a prophage, where it is replicated along with the bacterial DNA.

In this chapter, we shift from consideration of mapping genetic information in eukaryotes to discussion of the analysis and mapping of genes in **bacteria** (prokaryotes) and **bacteriophages,** viruses that use bacteria as their hosts. The study of bacteria and bacteriophages has been essential to the accumulation of knowledge in many areas of genetic study. For example, much of what we know about molecular genetics, recombinational phenomena, and gene structure was initially derived from experimental work with them. Furthermore, our extensive knowledge of bacteria and their resident plasmids has led to their widespread use in DNA cloning and other recombinant DNA studies.

Bacteria and their viruses are especially useful research organisms in genetics for several reasons. They have extremely short reproductive cycles—literally hundreds of generations, giving rise to billions of genetically identical bacteria or phages, can be produced in short periods of time. Furthermore, they can be studied in pure cultures. That is, a single species or mutant strain of bacteria or one type of virus can be isolated and investigated independently of other similar organisms.

In this chapter, we focus on genetic recombination and chromosome mapping. Complex processes have evolved in bacteria and bacteriophages that facilitate the transfer of genetic information between individual cells within populations. As we shall see, these processes are the basis for the chromosome mapping analysis that forms the cornerstone of molecular genetic investigations of bacteria and the viruses that invade them.

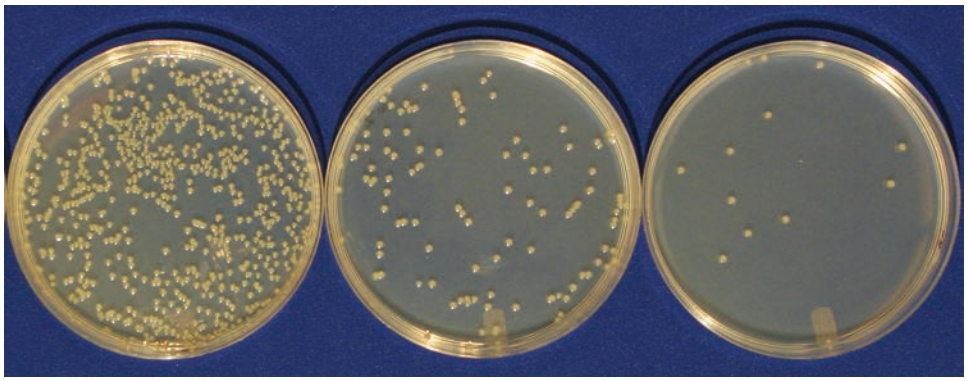


FIGURE 8–1 Results of the serial dilution technique and subsequent culture of bacteria. Each dilution varies by a factor of 10. Each colony is derived from a single bacterial cell.

8.1 Bacteria Mutate Spontaneously and Are Easily Cultured

It has long been known that pure cultures of bacteria give rise to cells that exhibit heritable variation, particularly with respect to growth under unique environmental conditions. Mutant cells that arise spontaneously in otherwise pure cultures can be isolated and established independently from the parent strain by using established selection techniques. As a result, mutations for almost any desired characteristic can now be isolated. Because bacteria and viruses usually contain only a single chromosome and are therefore haploid, all mutations are expressed directly in the descendants of mutant cells, adding to the ease with which these microorganisms can be studied.

Bacteria are grown in a liquid culture medium or in a petri dish on a semisolid agar surface. If the nutrient components of the growth medium are simple and consist only of an organic carbon source (such as glucose or lactose) and a variety of ions, including Na^+, K^+, Mg^{2+}, Ca^{2+}, and NH_4^+, present as inorganic salts, it is called **minimal medium.** To grow on such a medium, a bacterium must be able to synthesize all essential organic compounds (e.g., amino acids, purines, pyrimidines, vitamins, and fatty acids). A bacterium that can accomplish this remarkable biosynthetic feat—one that we ourselves cannot duplicate—is a **prototroph.** It is said to be wild-type for all growth requirements. On the other hand, if a bacterium loses the ability to synthesize one or more organic components through mutation, it is an **auxotroph.** For example, if a bacterium loses the ability to make histidine, then this amino acid must be added as a supplement to the minimal medium for growth to occur. The resulting bacterium is designated as an *his*$^-$ auxotroph, in contrast to its prototrophic *his*$^+$ counterpart.

To study bacterial growth quantitatively, an inoculum of bacteria is placed in liquid culture medium. Cells grown in liquid medium can be quantified by transferring them to a semisolid medium in a petri dish. Following incubation and many divisions, each cell gives rise to a visible colony on the surface of the medium. If the number of colonies is too great to count, then a series of successive dilutions (a technique called **serial dilution**) of the original liquid culture is made and plated, until the colony number is reduced to the point where it can be counted **(Figure 8–1)**. This technique allows the number of bacteria present in the original culture to be calculated.

As an example, let's assume that the three dishes in Figure 8–1 represent serial dilutions of 10^{-3}, 10^{-4}, and 10^{-5} (from left to right). We need only select the dish in which the number of colonies can be counted accurately. Assuming that a 1-mL sample was used, and because each colony arose from a single bacterium, the number of colonies multiplied by the dilution factor represents the number of bacteria in each milliliter of the initial inoculum used to start the serial dilutions. In Figure 8–1, the rightmost dish has 12 colonies. The dilution factor for a 10^{-5} dilution is 10^5. Therefore, the initial number of bacteria was 12×10^5 per mL.

8.2 Genetic Recombination Occurs in Bacteria

Development of techniques that allowed the identification and study of bacterial mutations led to detailed investigations of the transfer of genetic information between individual organisms. As we shall see, as with meiotic crossing over in eukaryotes, the process of genetic recombination in bacteria provided the basis for the development of chromosome mapping methodology. It is important to note at the outset of our discussion that the term *genetic recombination,* as applied to bacteria, refers to the *replacement* of one or more genes present in the chromosome of one cell with those from the chromosome of a genetically distinct cell. While this is somewhat different from our use of the term in eukaryotes—where it describes *crossing over resulting in a reciprocal exchange*—the overall effect is the same: Genetic information is transferred, and it results in an altered genotype.

We will discuss three processes that result in the transfer of genetic information from one bacterium to another: *conjugation*, *transformation,* and *transduction*. Collectively, knowledge of these processes has helped us understand the origin of genetic variation between members of the same bacterial species, and in some cases, between members of different species. When transfer of genetic information occurs between generations of the same species, the term **vertical gene transfer** applies. When transfer occurs between unrelated cells the term **horizontal gene transfer** is used. The horizontal gene transfer process has played a significant role in the evolution of bacteria. Often, the genes discovered to be involved in horizontal transfer are those that also confer survival advantages to the recipient species. For example, one species may transfer antibiotic resistance genes to another species. Or genes conferring enhanced pathogenicity may be transferred. Thus, the potential for such transfer is a major concern in the medical community. In addition, horizontal gene transfer has been a major factor in the process of speciation in bacteria. Many, if not most, bacterial species have been the recipient of genes from other species.

Conjugation in Bacteria: The Discovery of F^+ and F^- Strains

Studies of bacterial recombination began in 1946, when Joshua Lederberg and Edward Tatum showed that bacteria undergo **conjugation,** a process by which genetic information from one bacterium is transferred to and recombined with that of another bacterium. Their initial experiments were performed with two multiple auxotrophs (nutritional mutants) of *E. coli* strain K12. As shown in **Figure 8–2**, strain A required methionine (met) and biotin (bio) in order to grow, whereas strain B required threonine (thr), leucine (leu), and thiamine (thi). Neither strain would grow on minimal medium. The two strains were first grown separately in supplemented media, and then cells from both were mixed and grown together for several more generations. They were then plated on minimal medium. Any cells that grew on minimal medium were prototrophs. It is highly improbable that any of the cells containing two or three mutant genes would undergo spontaneous mutation simultaneously at two or three independent locations to become wild-type cells. Therefore, the researchers assumed that any prototrophs recovered must have arisen as a result of some form of genetic exchange and recombination between the two mutant strains.

In this experiment, prototrophs were recovered at a rate of $1/10^7$ (or 10^{-7}) cells plated. The controls for this experiment involved separate plating of cells from strains A and B on minimal medium. No prototrophs were recovered. Based on these observations, Lederberg and Tatum proposed that genetic exchange had occurred. Lederberg and Tatum's findings were soon followed by numerous experiments that elucidated the genetic basis of conjugation. It quickly became evident that

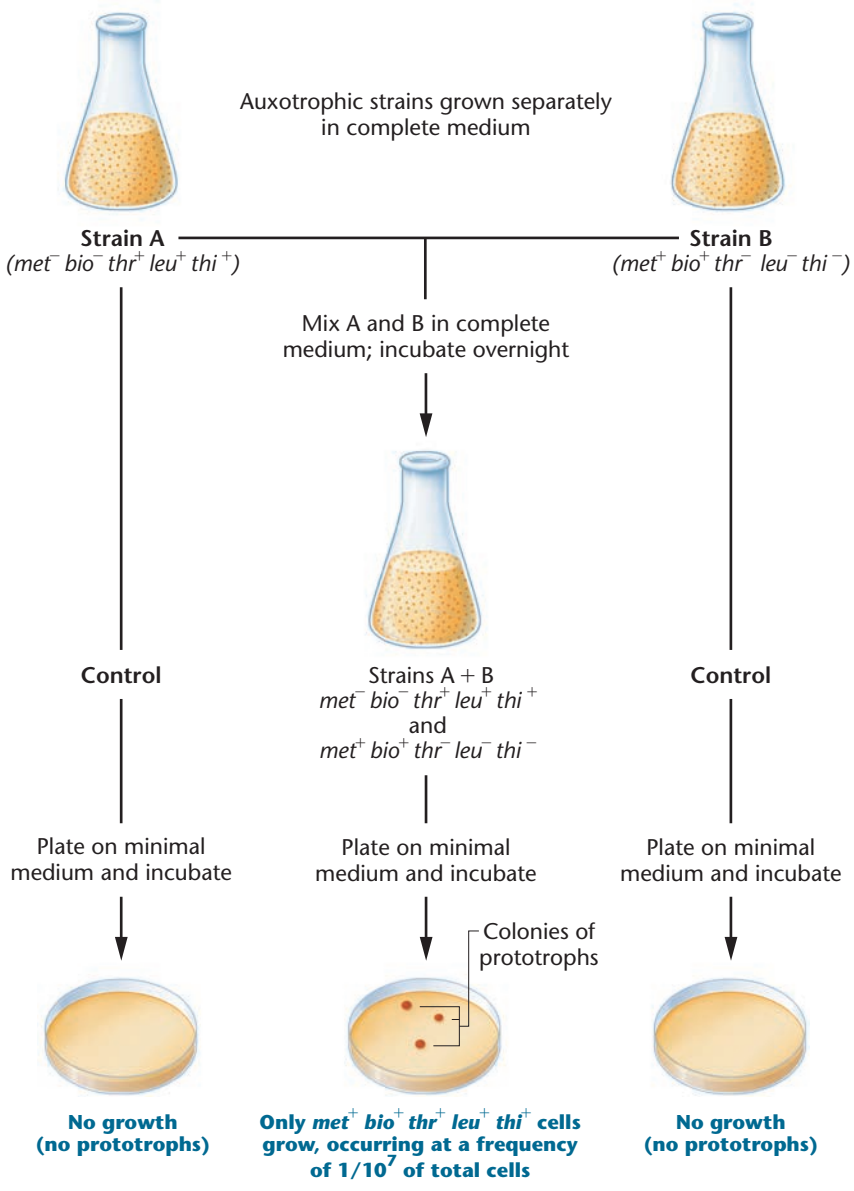


FIGURE 8–2 Genetic recombination of two auxotrophic strains producing prototrophs. Neither auxotroph grows on minimal medium, but prototrophs do, suggesting that genetic recombination has occurred.

different strains of bacteria are involved in a unidirectional transfer of genetic material. When cells serve as donors of parts of their chromosomes, they are designated as **F^+ cells** (F for "fertility"). Recipient bacteria receive the donor chromosome material (now known to be DNA), and recombine it with part of their own chromosome. They are designated as **F^- cells.**

Experimentation subsequently established that cell contact is essential for chromosome transfer to occur. Support for this concept was provided by Bernard Davis, who designed the Davis U-tube for growing F^+ and F^- cells shown in **Figure 8–3**. At the base of the tube is a sintered glass filter with a pore size that allows passage of the liquid medium but that is too small to allow the passage of bacteria. The F^+ cells are placed on one side of the filter, and F^- cells on the other side. The medium is moved back and forth across the filter so that the cells share a common medium during bacterial incubation. When Davis plated samples from both sides of the tube on minimal medium, no prototrophs were found, and he logically concluded that *physical contact between cells of the two strains is essential to genetic recombination.* We now know that this physical interaction is the initial step in the process of conjugation established by a structure called the **F pilus** (or **sex pilus;** pl. pili). Bacteria often have many pili, which are tubular extensions of the cell. After contact is initiated between mating pairs (see the chapter opening photograph on page 159), chromosome transfer is then possible.

Later evidence established that F^+ cells contain a **fertility factor (F factor)** that confers the ability to donate part of their chromosome during conjugation. Experiments by Joshua and Esther Lederberg and by William Hayes and Luca Cavalli-Sforza showed that certain conditions eliminate the F factor in otherwise fertile cells. However, if these "infertile" cells are then grown with fertile donor cells, the F factor is regained.

The conclusion that the F factor is a mobile element is further supported by the observation that, following conjugation and genetic recombination, recipient cells always become F^+. Thus, in addition to the *rare* cases of gene transfer from the bacterial chromosome (genetic recombination), the F factor itself is passed to *all* recipient cells. On this basis, the initial cross of Lederberg and Tatum (see Figure 8–2) can be interpreted as follow:

Strain A		Strain B
F^+	×	F^-
Donor		Recipient

Characterization of the F factor confirmed these conclusions. Like the bacterial chromosome, though distinct from it, the F factor has been shown to consist of a circular, double-stranded DNA molecule, equivalent to about

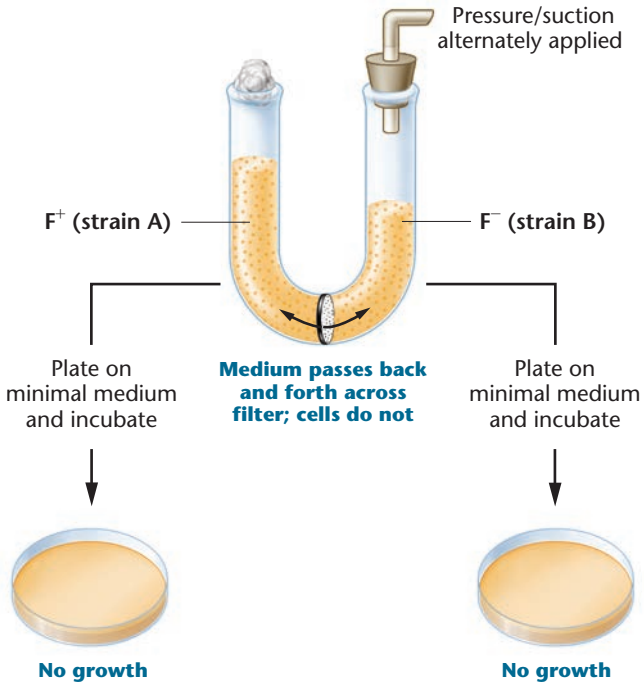


FIGURE 8–3 When strain A and B auxotrophs are grown in a common medium but separated by a filter, as in this Davis U-tube apparatus, no genetic recombination occurs and no prototrophs are produced.

2 percent of the bacterial chromosome (about 100,000 nucleotide pairs). There are 19 genes contained within the F factor whose products are involved in the transfer of genetic information, excluding those involved in the formation of the sex pilus.

Geneticists believe that transfer of the F factor during conjugation involves separation of the two strands of its double helix and movement of one of the two strands into the recipient cell. Both strands, one moving across the conjugation tube and one remaining in the donor cell, are replicated. The result is that both the donor *and* the recipient cells become F^+. This process is diagrammed in **Figure 8–4**.

To summarize, an *E. coli* cell may or may not contain the F factor. When this factor is present, the cell is able to form a sex pilus and potentially serve as a donor of genetic information. During conjugation, a copy of the F factor is almost always transferred from the F^+ cell to the F^- recipient, converting the recipient to the F^+ state. The question remained as to exactly why such a low proportion of cells involved in these matings (10^{-7}) also results in genetic recombination. The answer awaited further experimentation.

As you soon shall see, the F factor is in reality an autonomous genetic unit called a *plasmid.* However, in our historical coverage of its discovery, we will continue to refer to it as a factor.

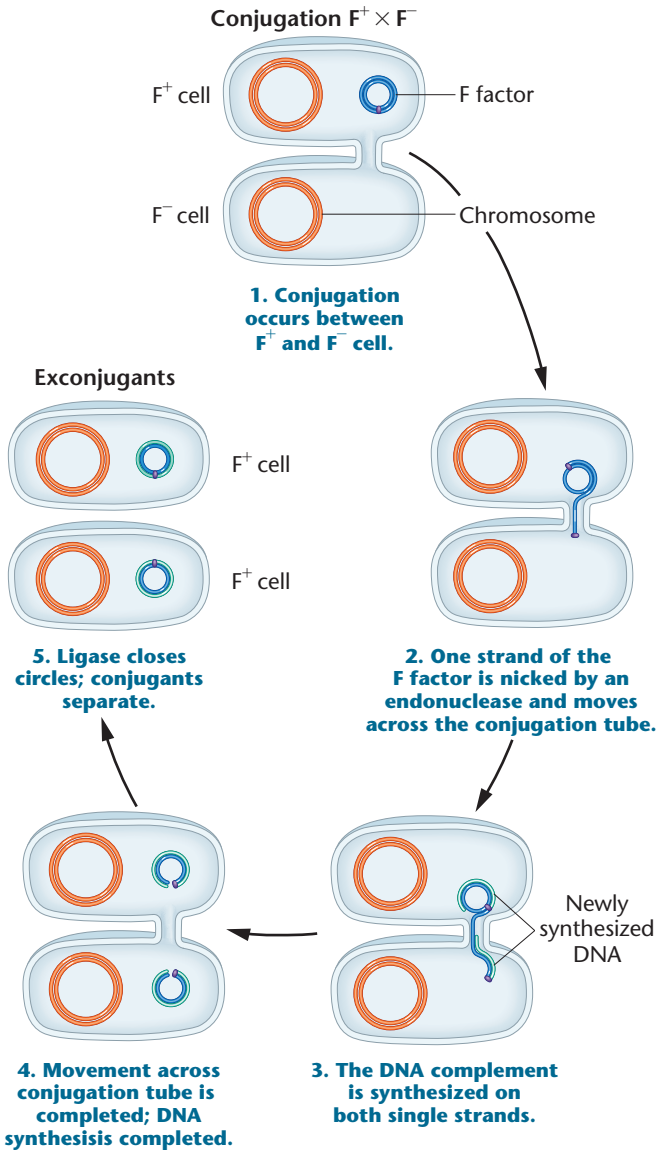


FIGURE 8–4 An $F^+ \times F^-$ mating demonstrating how the recipient F^- cell converts to F^+. During conjugation, the DNA of the F factor is replicated (Steps 2–4), with one new copy entering the recipient cell, converting it to F^+. Newly replicated DNA is depicted by a lighter shade of blue as the F factor is transferred.

ESSENTIAL POINT

Conjugation may be initiated by a bacterium housing a plasmid called the F factor in its cytoplasm, making it a donor (F^+) cell. Following conjugation, the recipient (F^-) cell receives a copy of the F factor and is converted to the F^+ status. ■

Hfr Bacteria and Chromosome Mapping

Subsequent discoveries not only clarified how genetic recombination occurs but also defined a mechanism by which the *E. coli* chromosome could be mapped. Let's address chromosome mapping first.

In 1950, Cavalli-Sforza treated an F^+ strain of *E. coli* K12 with nitrogen mustard, a chemical known to induce mutations. From these treated cells, he recovered a genetically altered strain of donor bacteria that underwent recombination at a rate of $1/10^4$ (or 10^{-4}), 1000 times more frequently than the original F^+ strains. In 1953, Hayes isolated another strain that demonstrated an elevated frequency. Both strains were designated **Hfr,** for **high-frequency recombination.** Because Hfr cells behave as donors, they are a special class of F^+ cells.

Another important difference was noted between Hfr strains and the original F^+ strains. If the donor is from an Hfr strain, recipient cells, though sometimes displaying genetic recombination, almost never become Hfr; that is, they remain F^-. In comparison, then,

$$F^+ \times F^- \longrightarrow F^+ \text{ (low rate of recombination)}$$

$$\text{Hfr} \times F^- \longrightarrow F^- \text{ (higher rate of recombination)}$$

Perhaps the most significant characteristic of Hfr strains is the *nature of recombination.* In any given strain, certain genes are more frequently recombined than others, and some not at all. This *nonrandom* pattern was shown to vary between Hfr strains. Although these results were puzzling, Hayes interpreted them to mean that some physiological alteration of the F factor had occurred, resulting in the production of Hfr strains of *E. coli.*

In the mid-1950s, experimentation by Ellie Wollman and François Jacob elucidated the difference between Hfr and F^+ strains and showed how Hfr strains allow genetic mapping of the *E. coli* chromosome. In their experiments, Hfr and F^- strains with suitable marker genes were mixed, and recombination of specific genes was assayed at different times. To accomplish this, a culture containing a mixture of an Hfr and an F^- strain was first incubated, and samples were removed at various intervals and placed in a blender. The shear forces in the blender separated conjugating bacteria so that the transfer of the chromosome was terminated. The cells were then assayed for genetic recombination.

This process, called the **interrupted mating technique,** demonstrated that specific genes of a given Hfr strain were transferred and recombined sooner than others. The graph in **Figure 8–5** illustrates this point. During the first 8 minutes after the two strains were mixed, no genetic recombination was detected. At about 10 minutes, recombination of the azi^R gene was detected, but no transfer of the ton^s, lac^+, or gal^+ genes was noted. By 15 minutes, 50 percent of the recombinants were azi^R, and 15 percent were ton^s; but none were lac^+ or gal^+. Within 20 minutes, the lac^+ was found among the recombinants; and within 25 minutes, gal^+ was also being transferred. Wollman and Jacob had demonstrated *an*

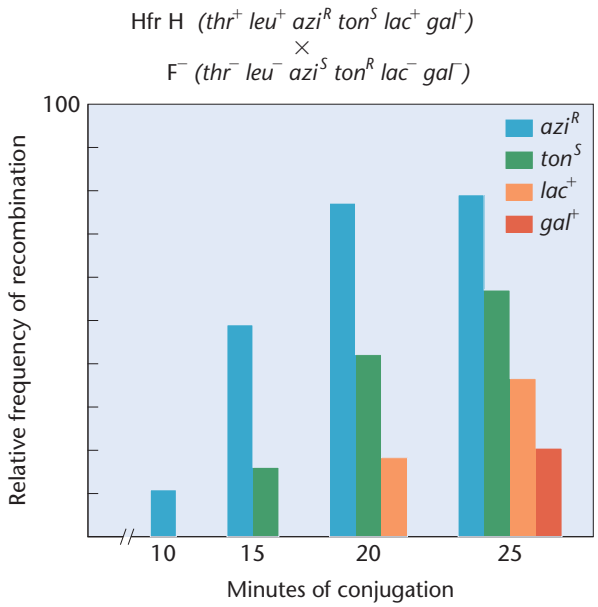


FIGURE 8–5 The progressive transfer during conjugation of various genes from a specific Hfr strain of *E. coli* to an F⁻ strain. Certain genes (*azi* and *ton*) transfer more quickly than others and recombine more frequently. Others (*lac* and *gal*) take longer to transfer, and recombinants are found at a lower frequency.

ordered transfer of genes that correlated with the length of time conjugation proceeded.

It appeared that the chromosome of the Hfr bacterium was transferred linearly and that the gene order and

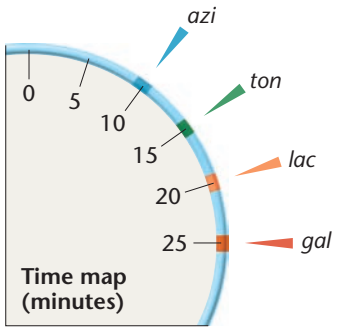


FIGURE 8–6 A time map of the genes studied in the experiment depicted in Figure 8–5.

distance between genes, as measured in minutes, could be predicted from such experiments (**Figure 8–6**). This process, sometimes referred to as **time mapping,** served as the basis for the first genetic map of the *E. coli* chromosome. Minutes in bacterial mapping are similar to map units in eukaryotes.

Wollman and Jacob then repeated the same type of experiment with other Hfr strains, obtaining similar results with one important difference. Although genes were always transferred linearly with time, as in their original experiment, the order in which genes entered seemed to vary from Hfr strain to Hfr strain [**Figure 8–7(a)**]. When they reexamined the entry rate of genes, and thus the genetic maps for each strain, a definite pattern emerged. The major

(a)

Hfr strain	(earliest) Order of transfer (latest)
H	*thr* – *leu* – *azi* – *ton* – *pro* – *lac* – *gal* – *thi*
1	*leu* – *thr* – *thi* – *gal* – *lac* – *pro* – *ton* – *azi*
2	*pro* – *ton* – *azi* – *leu* – *thr* – *thi* – *gal* – *lac*
7	*ton* – *azi* – *leu* – *thr* – *thi* – *gal* – *lac* – *pro*

(b)

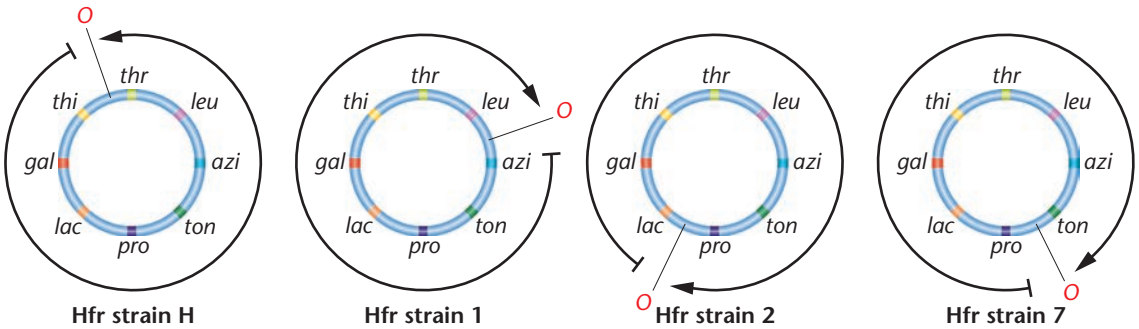


FIGURE 8–7 (a) The order of gene transfer in four Hfr strains, suggesting that the *E. coli* chromosome is circular. (b) The point where transfer originates (*O*) is identified in each strain. The origin is determined by the point of integration into the chromosome of the F factor, and the direction of transfer is determined by the orientation of the F factor as it integrates. The arrowheads indicate the points of initial transfer.

difference between each strain was simply the point of origin (*O*) and the direction in which entry proceeded from that point [**Figure 8–7(b)**].

To explain these results, Wollman and Jacob postulated that the *E. coli* chromosome is circular (a closed circle, with no free ends). If the point of origin (*O*) varies from strain to strain, a different sequence of genes will be transferred in each case. But what determines *O*? They proposed that *in various Hfr strains, the F factor integrates into the chromosome at different points and that its position determines the site of O.* A case of integration is shown in step 1 of **Figure 8–8**. During conjugation between an Hfr and an F^- cell, the position of the F factor determines the initial point of transfer (steps 2 and 3). Those genes adjacent to *O* are transferred first, and the F factor becomes the last part that can be transferred (step 4). However, conjugation rarely, if ever, lasts long enough to allow the entire chromosome to pass across the conjugation tube (step 5). *This proposal explains why most recipient cells, when mated with Hfr cells, remain F^-.*

Figure 8–8 also depicts the way in which the two strands making up a DNA molecule behave during transfer, allowing for the entry of one strand of DNA into the recipient (see step 3). Following replication, the entering DNA now has the potential to recombine with its homologous region of the host chromosome. The DNA strand that remains in the donor also undergoes replication.

Use of the interrupted mating technique with different Hfr strains allowed researchers to map the entire *E. coli* chromosome. Mapped in time units, strain K12 (or *E. coli* K12) was shown to be 100 minutes long. While modern genome analysis of the *E. coli* chromosome has now established the presence of just over 4000 protein-coding sequences, this original mapping procedure established the location of approximately 1000 genes.

ESSENTIAL POINT

When the F factor is integrated into the donor cell chromosome (making it Hfr), the donor chromosome moves unidirectionally into the recipient, initiating recombination and providing the basis for time mapping of the bacterial chromosome. ■

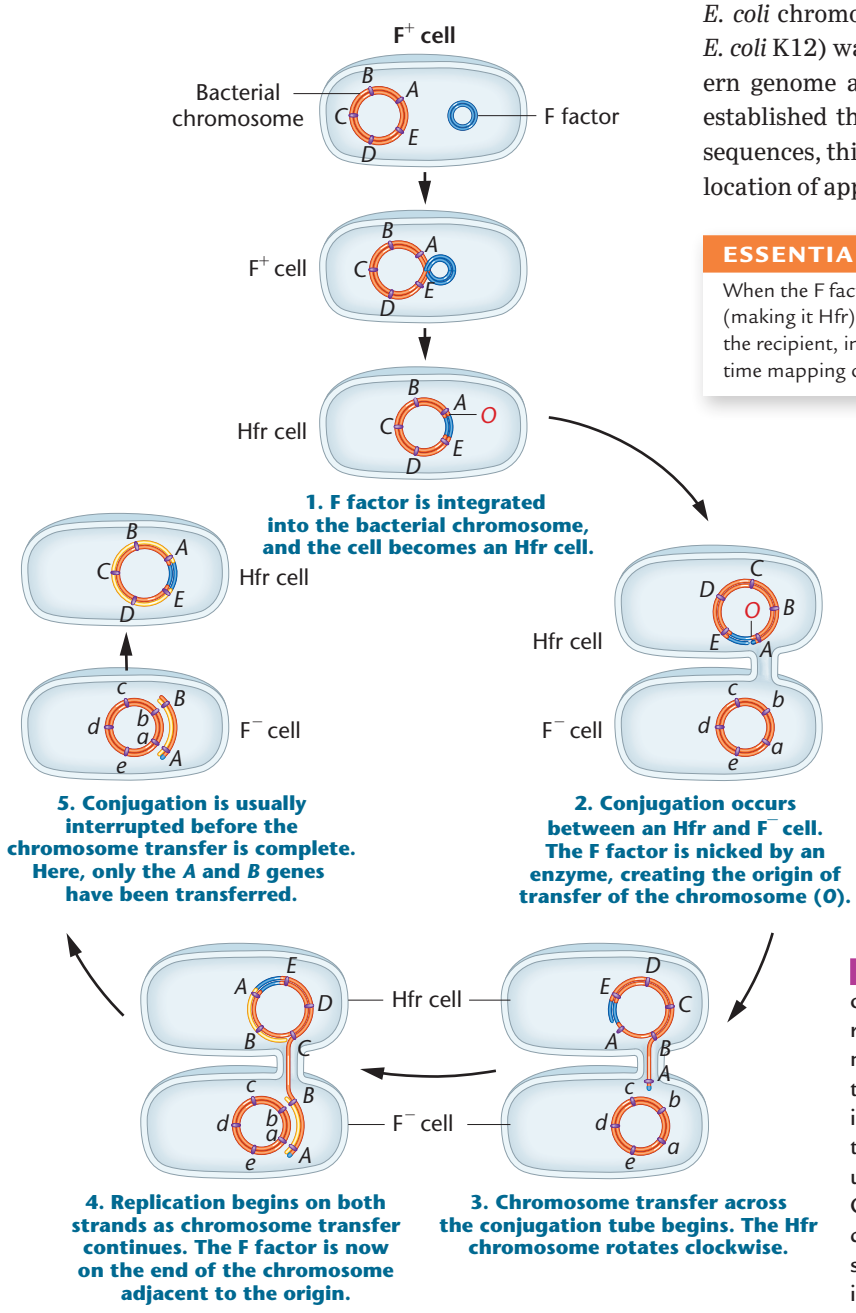


FIGURE 8–8 Conversion of F^+ to an Hfr state occurs by integrating the F factor into the bacterial chromosome. The point of integration determines the origin (*O*) of transfer. During conjugation, an enzyme nicks the F factor, now integrated into the host chromosome, initiating transfer of the chromosome at that point. Conjugation is usually interrupted prior to complete transfer. Only the *A* and *B* genes are transferred to the F^- cell, which may recombine with the host chromosome. Newly replicated DNA of the chromosome is depicted by a lighter shade of orange.

NOW SOLVE THIS

8–1 When the interrupted mating technique was used with five different strains of Hfr bacteria, the following orders of gene entry and recombination were observed. On the basis of these data, draw a map of the bacterial chromosome. Do the data support the concept of circularity?

Hfr Strain	Order				
1	*T*	*C*	*H*	*R*	*O*
2	*H*	*R*	*O*	*M*	*B*
3	*M*	*O*	*R*	*H*	*C*
4	*M*	*B*	*A*	*K*	*T*
5	*C*	*T*	*K*	*A*	*B*

■ **HINT:** *This problem involves an understanding of how the bacterial chromosome is transferred during conjugation, leading to recombination and providing data for mapping. The key to its solution is to understand that chromosome transfer is strain-specific and depends on where in the chromosome, and in which orientation, the F factor has integrated.*

Recombination in $F^+ \times F^-$ Matings: A Reexamination

The preceding experiment helped geneticists better understand how genetic recombination occurs during $F^+ \times F^-$ matings. Recall that recombination occurs much less frequently than in Hfr $\times$ F^- matings and that random gene transfer is involved. The current belief is that when F^+ and F^- cells are mixed, conjugation occurs readily and each F^- cell involved in conjugation with an F^+ cell receives a copy of the F factor, *but no genetic recombination occurs.* However, at an extremely low frequency in a population of F^+ cells, the F factor integrates spontaneously from the cytoplasm to a random point in the bacterial chromosome, converting the F^+ cell to the Hfr state, as we saw in Figure 8–8. Therefore, in $F^+ \times F^-$ matings, the extremely low frequency of genetic recombination (10^{-7}) is attributed to the rare, newly formed Hfr cells, which then undergo conjugation with F^- cells. Because the point of integration of the F factor is random, the gene or genes that are transferred by any newly formed Hfr donor *will also appear to be random within the larger F^+/F^- population.* The recipient bacterium will appear as a recombinant but will remain F^-. If it subsequently undergoes conjugation with an F^+ cell, it will then be converted to F^+.

The F′ State and Merozygotes

In 1959, during experiments with Hfr strains of *E. coli,* Edward Adelberg discovered that the F factor could lose its integrated status, causing the cell to revert to the F^+ state (**Figure 8–9**, step 1). When this occurs, the F factor frequently carries several adjacent bacterial genes along with it (step 2). Adelberg labeled this condition **F′** to distinguish it from F^+ and Hfr. F′, like Hfr, is thus another special case of F^+, but this conversion is from Hfr to F′.

The presence of bacterial genes within a cytoplasmic F factor creates an interesting situation. An F′ bacterium behaves like an F^+ cell by initiating conjugation with F^- cells (Figure 8–9, step 3). When this occurs, the F factor, containing chromosomal genes, is transferred to the F^- cell (step 4). As a result, whatever chromosomal genes are part of the F factor are now present as duplicates in the recipient cell (step 5) because the recipient still has a complete chromosome. This creates a partially diploid cell called a **merozygote.** Pure cultures of F′ merozygotes can be established. They have been extremely useful in the study of bacterial genetics, particularly in genetic regulation.

8.3 Rec Proteins Are Essential to Bacterial Recombination

Once researchers established that a unidirectional transfer of DNA occurs between bacteria, they became interested in determining how the actual recombination event occurs in the recipient cell. Just how does the donor DNA replace the homologous region in the recipient chromosome? As with many systems, the biochemical mechanism by which recombination occurs was deciphered through genetic studies. Major insights were gained as a result of the isolation of a group of mutations that impaired the process of recombination and led to the discovery of *rec* (for recombination) genes.

The first relevant observation in this case involved a series of mutant genes labeled *recA, recB, recC,* and *recD.* The first mutant gene, *recA,* diminished genetic recombination in bacteria 1000-fold, nearly eliminating it altogether; each of the other *rec* mutations reduced recombination by about 100 times. Clearly, the normal wild-type products of these genes play some essential role in the process of recombination.

Researchers looked for, and subsequently isolated, several functional gene products present in normal cells but missing in *rec* mutant cells and showed that they played a role in genetic recombination. The first product is called the **RecA protein.*** This protein plays an important

* Note that the names of bacterial genes use lowercase letters and are italicized, while the names of the corresponding gene products begin with capital letters and are not italicized. For example, the *recA* gene encodes the RecA protein.

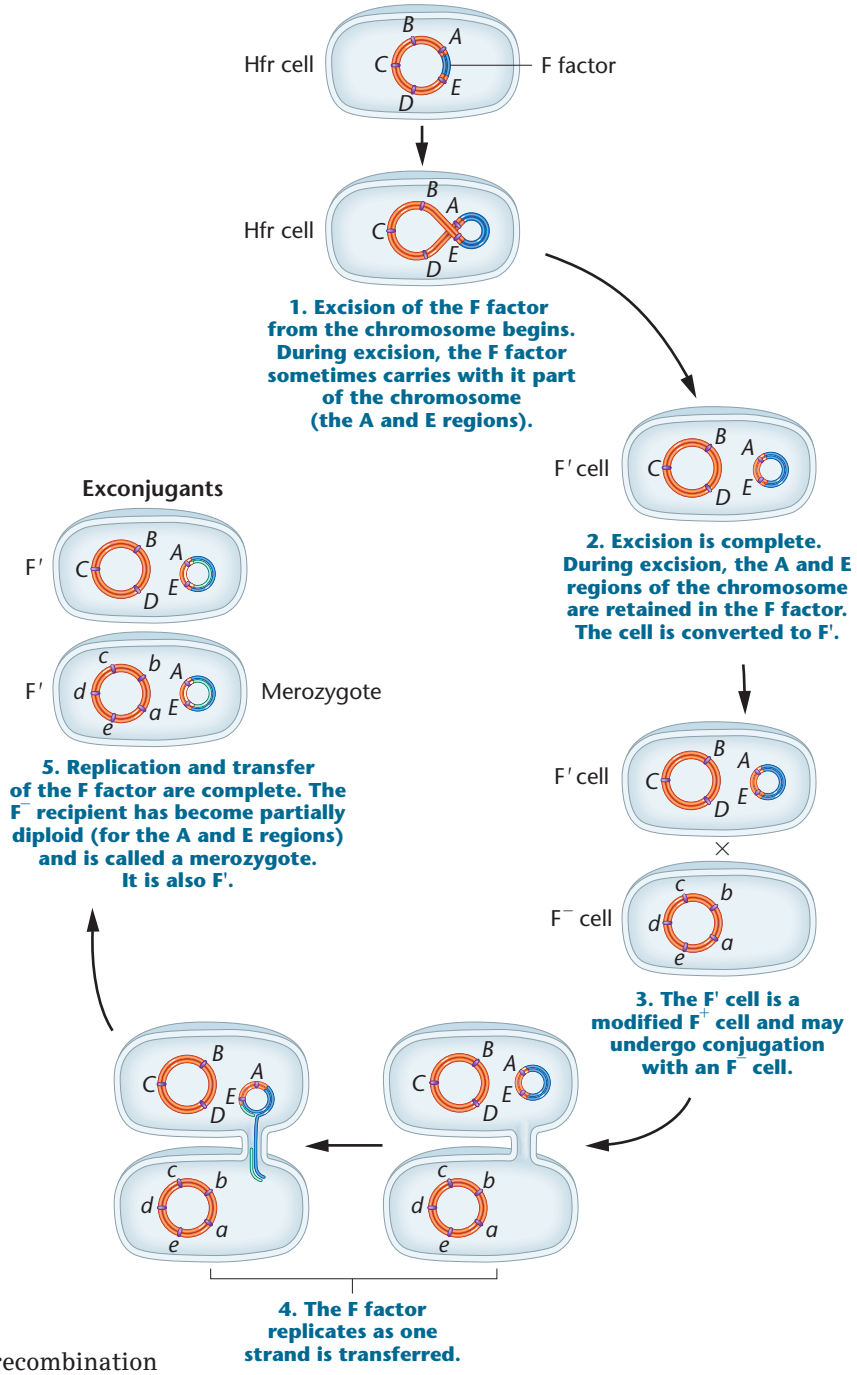


FIGURE 8–9 Conversion of an Hfr bacterium to F′ and its subsequent mating with an F⁻ cell. The conversion occurs when the F factor loses its integrated status. During excision from the chromosome, it carries with it one or more chromosomal genes (*A* and *E*). Following conjugation with an F⁻ cell, the recipient cell becomes partially diploid and is called a merozygote; it also behaves as an F⁺ donor cell.

role in recombination involving either a single-stranded DNA molecule or the linear end of a double-stranded DNA molecule that has unwound. As it turns out, **single-strand displacement** is a common form of recombination in many bacterial species. When double-stranded DNA enters a recipient cell, one strand is often degraded, leaving the complementary strand as the only source of recombination. This strand must find its homologous region along the host chromosome, and once it does, RecA facilitates recombination.

The second related gene product is a more complex protein called the **RecBCD protein,** an enzyme consisting of polypeptide subunits encoded by three other *rec* genes. This protein is important when double-stranded DNA serves as the source of genetic recombination. RecBCD unwinds the helix, facilitating recombination that involves RecA. These discoveries have extended our knowledge of the process of recombination considerably and underscore the value of isolating mutations, establishing their phenotypes, and determining the biological role of the normal, wild-type genes. The model of recombination based on the *rec* discoveries also applies to eukaryotes: Eukaryotic proteins similar to RecA have been isolated and studied.

8.4 The F Factor Is an Example of a Plasmid

The preceding sections introduced the extrachromosomal heredity unit required for conjugation called the F factor. When it exists autonomously in the bacterial cytoplasm, the F factor is composed of a double-stranded closed circle of DNA. These characteristics place the F factor in the more general category of genetic structures called **plasmids** [Figure 8–10(a)]. These structures contain one or more genes and often, quite a few. Their replication depends on the same enzymes that replicate the chromosome of the host cell, and they are distributed to daughter cells along with the host chromosome during cell division.

Plasmids are generally classified according to the genetic information specified by their DNA. The F factor plasmid

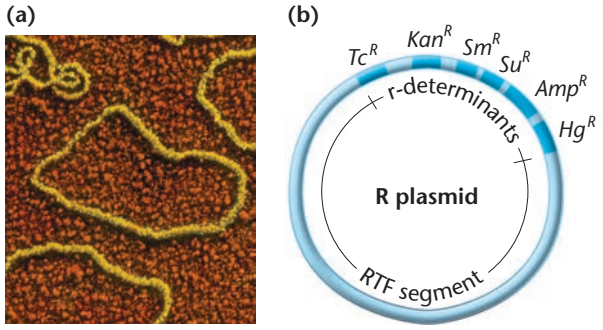


FIGURE 8–10 (a) Electron micrograph of a plasmid isolated from *E. coli*. (b) An R plasmid containing resistance transfer factors (RTFs) and multiple r-determinants (Tc, tetracycline; Kan, kanamycin; Sm, streptomycin; Su, sulfonamide; Amp, ampicillin; and Hg, mercury).

confers fertility and contains the genes essential for sex pilus formation, on which genetic recombination depends. Other examples of plasmids include the R and Col plasmids.

Most **R plasmids** consist of two components: the **resistance transfer factor (RTF)** and one or more **r-determinants** [**Figure 8–10(b)**]. The RTF encodes genetic information essential to transferring the plasmid between bacteria, and the r-determinants are genes that confer resistance to antibiotics or mercury. While RTFs are similar in a variety of plasmids from different bacterial species, r-determinants are specific for resistance to one class of antibiotic and vary widely. Resistance to tetracycline, streptomycin, ampicillin, sulfonamide, kanamycin, or chloramphenicol is most frequently encountered. Sometimes several r-determinants occur in a single plasmid, conferring multiple resistance to several antibiotics [Figure 8–10(b)]. Bacteria bearing these plasmids are of great medical significance not only because of their multiple resistance but because of the ease with which the plasmids can be transferred to other bacteria.

The first known case of such a plasmid occurred in Japan in the 1950s in the bacterium *Shigella*, which causes dysentery. In hospitals, bacteria were isolated that were resistant to as many as five of the above antibiotics. Obviously, this phenomenon represents a major health threat. Fortunately, a bacterial cell sometimes contains r-determinant plasmids but no RTF. Although such a cell is resistant, it cannot transfer the genetic information for resistance to recipient cells. The most commonly studied plasmids, however, contain the RTF as well as one or more r-determinants.

The **Col plasmid,** ColE1 (derived from *E. coli*), is clearly distinct from the R plasmid. It encodes one or more proteins that are highly toxic to bacterial strains that do not harbor the same plasmid. These proteins, called **colicins,** can kill neighboring bacteria, and bacteria that carry the plasmid are said to be *colicinogenic*. Present in 10 to 20 copies per cell, a gene in the Col plasmid encodes an immunity protein that protects the host cell from the toxin. Unlike an R plasmid, the Col plasmid is not usually transmissible to other cells.

Interest in plasmids has increased dramatically because of their role in recombinant DNA research. As we will see in Chapter 17, specific genes from any source can be inserted into a plasmid, which may then be inserted into a bacterial cell. As the altered cell replicates its DNA and undergoes division, the foreign gene is also replicated, thus cloning the foreign genes.

ESSENTIAL POINT

Plasmids, such as the F factor, are autonomously replicating DNA molecules found in the bacterial cytoplasm, sometimes containing unique genes conferring antibiotic resistance as well as the genes necessary for plasmid transfer during conjugation. ■

8.5 Transformation Is Another Process Leading to Genetic Recombination in Bacteria

Transformation provides another mechanism for recombining genetic information in some bacteria. Small pieces of extracellular DNA are taken up by a living bacterium, potentially leading to a stable genetic change in the recipient cell. We discuss transformation in this chapter because in those bacterial species where it occurs, the process can be used to map bacterial genes, though in a more limited way than conjugation. Transformation has also played a central role in experiments proving that DNA is the genetic material.

The process of transformation consists of numerous steps divided into two categories: (1) entry of DNA into a recipient cell and (2) recombination of the donor DNA with its homologous region in the recipient chromosome. In a population of bacterial cells, only those in the particular physiological state of **competence** take up DNA. Entry is thought to occur at a limited number of receptor sites on the surface of the bacterial cell. Passage into the cell is an active process that requires energy and specific transport molecules. This model is supported by the fact that substances that inhibit energy production or protein synthesis in the recipient cell also inhibit the transformation process.

During entry, one of the two strands of the double helix is digested by nucleases, leaving only a single strand to participate in transformation. The surviving strand of DNA then aligns with its complementary region of the bacterial chromosome. In a process involving several

enzymes, the segment replaces its counterpart in the chromosome, which is excised and degraded. For recombination to be detected, the transforming DNA must be derived from a different strain of bacteria that bears some genetic variation, such as a mutation. Once it is integrated into the chromosome, the recombinant region contains one host strand (present originally) and one mutant strand. Because these strands are from different sources, this helical region is referred to as a **heteroduplex.** Following one round of DNA replication, one chromosome is restored to its original configuration, and the other contains the mutant gene. Following cell division, one untransformed cell (nonmutant) and one transformed cell (mutant) are produced.

Transformation and Linked Genes

In early transformation studies, the most effective exogenous DNA contained 10,000–20,000 nucleotide pairs, a length sufficient to encode several genes.* Genes adjacent to or very close to one another on the bacterial chromosome can be carried on a single segment of this size. Consequently, a single transfer event can result in the cotransformation of several genes simultaneously. Genes that are close enough to each other to be cotransformed are *linked*. In contrast to *linkage groups* in eukaryotes, which consist of all genes on a single chromosome, note that here *linkage* refers to the proximity of genes that permits cotransformation (i.e., the genes are next to, or close to, one another).

If two genes are not linked, simultaneous transformation occurs only as a result of two independent events involving two distinct segments of DNA. As in double crossovers in eukaryotes, the probability of two independent events occurring simultaneously is equal to the product of the individual probabilities. Thus, the frequency of two unlinked genes being transformed simultaneously is much lower than if they are linked.

NOW SOLVE THIS

8–2 In a transformation experiment involving a recipient bacterial strain of genotype a^-b^-, the following results were obtained. What can you conclude about the location of the *a* and *b* genes relative to each other?

	Transformants (%)		
Transforming DNA	a^+b^-	a^-b^+	a^+b^+
a^+b^+	3.1	1.2	0.04
a^+b^- and a^-b^+	2.4	1.4	0.03

■ **HINT:** *This problem involves an understanding of how transformation can be used to determine if bacterial genes are closely "linked." You are asked to predict the location of two genes relative to one another. The key to its solution is to understand that cotransformation (of two genes) occurs according to the laws of probability. Two "unlinked" genes are transformed only as a result of two independent events. In such a case, the probability of that occurrence is equal to the product of the individual probabilities.*

* Today, we know that a 2000 nucleotide pair length of DNA is highly effective in gene cloning experiments.

ESSENTIAL POINT

Transformation in bacteria, which does not require cell-to-cell contact, involves exogenous DNA that enters a recipient bacterium and recombines with the host's chromosome. Linkage mapping of closely aligned genes is possible during the analysis of transformation. ■

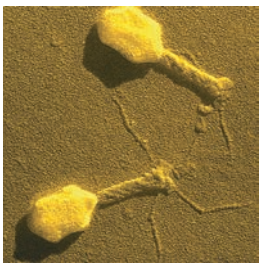
8.6 Bacteriophages Are Bacterial Viruses

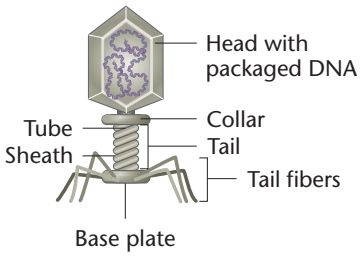
Bacteriophages, or **phages** as they are commonly known, are viruses that have bacteria as their hosts. During their reproduction, phages can be involved in still another mode of bacterial genetic recombination called *transduction.* To understand this process, we must consider the genetics of bacteriophages, which themselves undergo recombination.

A great deal of genetic research has been done using bacteriophages as a model system, making them a worthy subject of discussion. In this section, we will first examine the structure and life cycle of one type of bacteriophage. We then discuss how these phages are studied during their infection of bacteria. Finally, we contrast two possible modes of behavior once the initial phage infection occurs. This information is background for our discussion of *transduction* and *bacteriophage recombination.*

Phage T4: Structure and Life Cycle

Bacteriophage T4 is one of a group of related bacterial viruses referred to as T-even phages. It exhibits the intricate structure shown in **Figure 8–11**. The phage T4's genetic material (DNA) is contained within an icosahedral (a polyhedron with 20 faces) protein coat, making up the head of the virus. The DNA is sufficient in quantity to encode more than 150 average-sized genes. The head is connected to a tail that contains a collar and a contractile sheath surrounding a central core. Tail fibers, which protrude from





Mature T4 phage

FIGURE 8–11 The structure of bacteriophage T4 includes an icosahedral head filled with DNA, a tail consisting of a collar, tube, sheath, base plate, and tail fibers. During assembly, the tail components are added to the head, and then tail fibers are added.

the tail, contain binding sites in their tips that specifically recognize unique areas of the outer surface of the cell wall of the bacterial host, *E. coli.*

The life cycle of phage T4 (**Figure 8–12**) is initiated when the virus binds by adsorption to the bacterial host cell. Then, an ATP-driven contraction of the tail sheath causes the central core to penetrate the cell wall. The DNA in the head is extruded, and it moves across the cell membrane into the bacterial cytoplasm. Within minutes, all bacterial DNA, RNA, and protein synthesis in the host cell is inhibited, and synthesis of viral molecules begins. At the same time, degradation of the host DNA is initiated.

A period of intensive viral gene activity characterizes infection. Initially, phage DNA replication occurs, leading to a pool of viral DNA molecules. Then, the components of the head, tail, and tail fibers are synthesized. The assembly of mature viruses is a complex process that has been well studied by William Wood, Robert Edgar, and others. Three sequential pathways occur: (1) DNA packaging as the viral heads are assembled, (2) tail assembly, and (3) tail fiber assembly. Once DNA is packaged into the head, it combines with the tail components, to which tail fibers are added. Total construction is a combination of self-assembly and enzyme-directed processes.

When approximately 200 new viruses have been constructed, the bacterial cell is ruptured by the action of the enzyme lysozyme (a phage gene product), and the mature phages are released from the host cell. The new phages infect other available bacterial cells, and the process repeats itself over and over again.

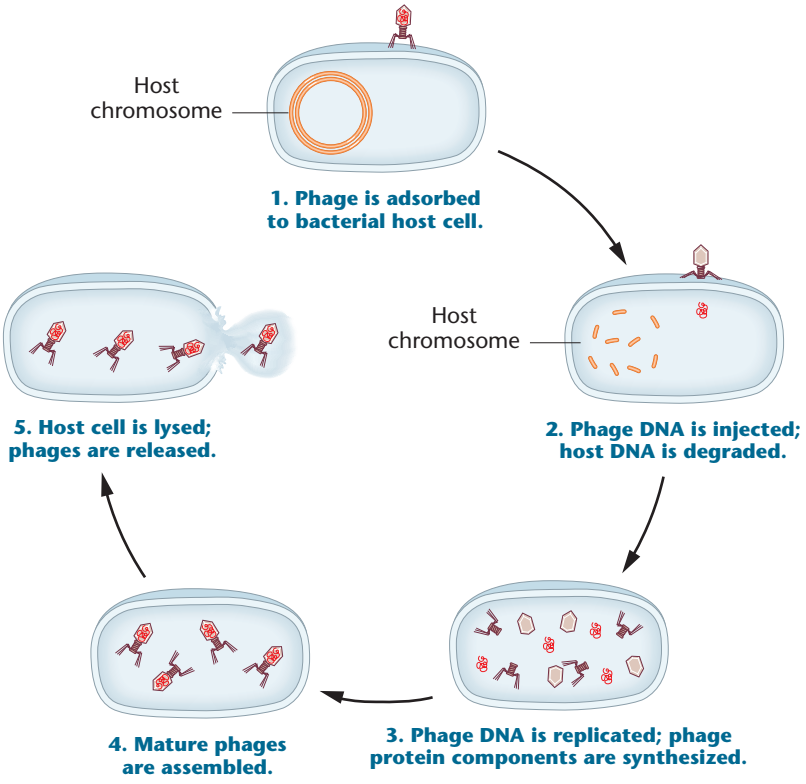


FIGURE 8–12 Life cycle of bacteriophage T4.

The Plaque Assay

Bacteriophages and other viruses have played a critical role in our understanding of molecular genetics. During infection of bacteria, enormous quantities of bacteriophages can be obtained for investigation. Often, over 10^{10} viruses are produced per milliliter of culture medium. Many genetic studies rely on the ability to quantify the number of phages produced following infection under specific culture conditions. The **plaque assay** is a routinely used technique, which is invaluable in mutational and recombinational studies of bacteriophages.

This assay is shown in **Figure 8–13**, where actual plaque morphology is also illustrated. A serial dilution of the original virally infected bacterial culture is performed first. Then, a 0.1-mL sample (an *aliquot*) from a dilution is added to melted nutrient agar (about 3 mL) into which a few drops of a healthy bacterial culture have been added. The solution is then poured evenly over a base of solid nutrient agar in a petri dish and allowed to solidify before incubation. A clear area called a **plaque** occurs wherever a single virus initially infected one bacterium in the culture (the lawn) that has grown up during incubation. The plaque represents clones of the single infecting bacteriophage, created as reproduction cycles are repeated. If the dilution factor is too low, the plaques are plentiful, and they will fuse, lysing the entire lawn—which has occurred in the 10^{-3} dilution of Figure 8–13. On the other hand, if

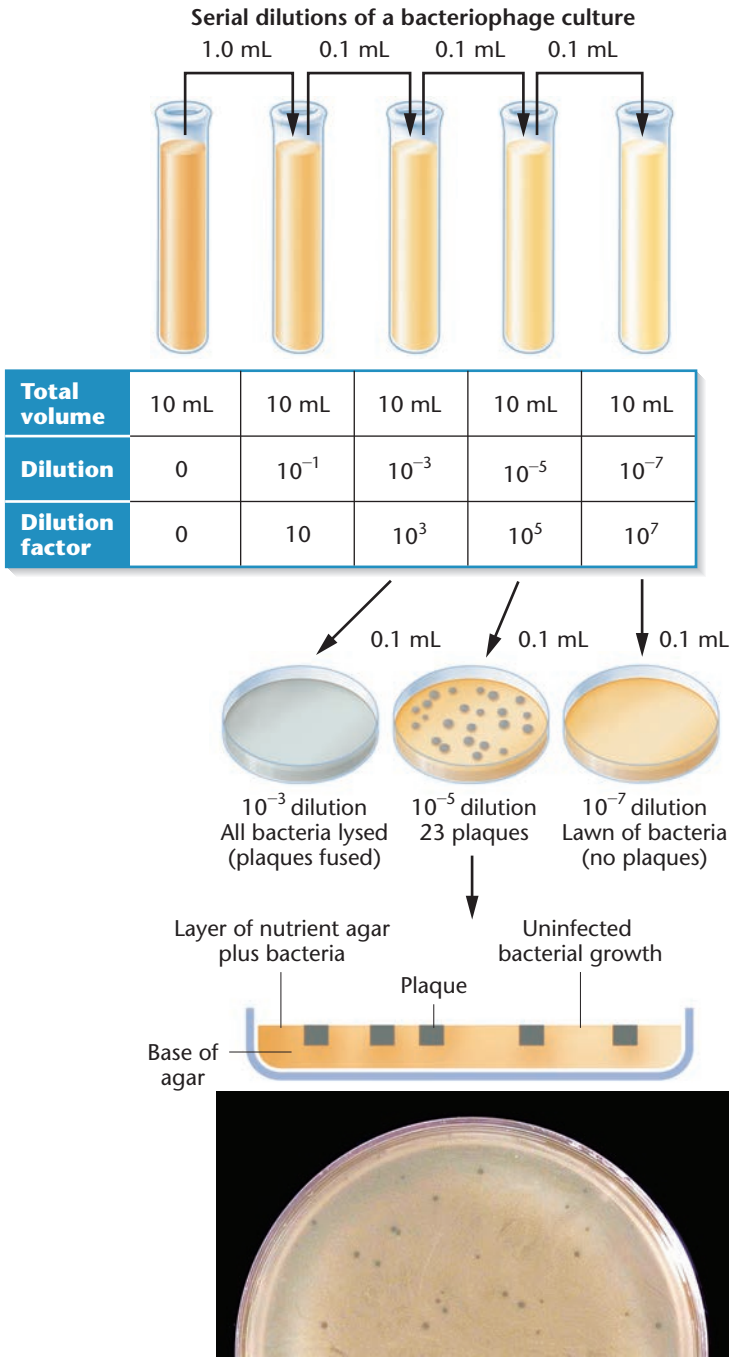


FIGURE 8–13 A plaque assay for bacteriophage analysis. Serial dilutions of a bacterial culture infected with bacteriophages are first made. Then three of the dilutions (10^{-3}, 10^{-5}, and 10^{-7}) are analyzed using the plaque assay technique. Each plaque represents the initial infection of one bacterial cell by one bacteriophage. In the 10^{-3} dilution, so many phages are present that all bacteria are lysed. In the 10^{-5} dilution, 23 plaques are produced. In the 10^{-7} dilution, the dilution factor is so great that no phages are present in the 0.1-mL sample, and thus no plaques form. From the 0.1-mL sample of the 10^{-5} dilution, the original bacteriophage density is calculated to be $(230/\text{mL}) \times (10^5)$ phages/mL (23×10^6, or 2.3×10^7). The photograph shows phage plaques on a lawn of *E. coli*.

the dilution factor is increased, plaques can be counted and the density of viruses in the initial culture can be estimated as

$$\text{initial phage density} = (\text{plaque number}/\text{mL}) \times (\text{dilution factor})$$

Using the results shown in Figure 8–13, 23 phage plaques are derived from the 0.1-mL aliquot of the 10^{-5} dilution. Therefore, we estimate that there are 230 phages/mL *at this dilution* (since the initial aliquot was 0.1 mL). The initial phage density in the undiluted sample, factoring in the 10^{-5} dilution, is then calculated as

$$\begin{aligned}\text{initial phage density} &= (230/\text{mL}) \times (10^5) \\ &= 230 \times 10^5/\text{mL}\end{aligned}$$

Because this figure is derived from the 10^{-5} dilution, we can also estimate that there will be only 0.23 phage/0.1 mL in the 10^{-7} dilution. Thus, when 0.1 mL from this tube is assayed, it is predicted that no phage particles will be present. This possibility is borne out in Figure 8–13, which depicts an intact lawn of bacteria lacking any plaques. The dilution factor is simply too great.

ESSENTIAL POINT

Bacteriophages (viruses that infect bacteria) demonstrate a well-defined life cycle where they reproduce within the host cell and can be studied using the plaque assay. ■

Lysogeny

Infection of a bacterium by a virus does not always result in viral reproduction and lysis. As early as the 1920s, it was known that some viruses can enter a bacterial cell and coexist with it. The precise molecular basis of this relationship is now well understood. Upon entry, the viral DNA is integrated into the bacterial chromosome instead of replicating in the bacterial cytoplasm, a step that characterizes the developmental stage referred to as **lysogeny.** Subsequently, each time the bacterial chromosome is replicated, the viral DNA is also replicated and passed to daughter bacterial cells following division. No new viruses are produced, and no lysis of the bacterial cell occurs. However, under certain stimuli, such as chemical or ultraviolet light treatment, the viral DNA loses its integrated status and initiates replication, phage reproduction, and lysis of the bacterium.

Several terms are used to describe this relationship. The viral DNA that integrates into the bacterial chromosome is called a **prophage.** Viruses that either lyse the cell

or behave as a prophage are **temperate phages.** Those that only lyse the cell are referred to as **virulent phages.** A bacterium harboring a prophage is said to be **lysogenic;** that is, it is capable of being lysed as a result of induced viral reproduction. The viral DNA, which can either replicate in the bacterial cytoplasm or become integrated into the bacterial chromosome, is thus classified as an **episome,** meaning a genetic molecule that can replicate either in the cytoplasm of a cell or as part of its chromosome.

ESSENTIAL POINT

Bacteriophages can be lytic, meaning they infect the host cell, reproduce, and then lyse it, or in contrast, they can lysogenize the host cell, where they infect it and integrate their DNA into the host chromosome, but do not reproduce. ■

8.7 Transduction Is Virus-Mediated Bacterial DNA Transfer

In 1952, Norton Zinder and Joshua Lederberg were investigating possible recombination in the bacterium *Salmonella typhimurium.* Although they recovered prototrophs from mixed cultures of two different auxotrophic strains, investigation revealed that recombination was occurring in a manner different from that attributable to the presence of an F factor, as in *E. coli.* What they had discovered was a process of bacterial recombination mediated by bacteriophages and now called **transduction.**

The Lederberg–Zinder Experiment

Lederberg and Zinder mixed the *Salmonella* auxotrophic strains LA-22 and LA-2 together, and when the mixture was plated on minimal medium, they recovered prototrophic cells. The LA-22 strain was unable to synthesize the amino acids phenylalanine and tryptophan (phe^-, trp^-), and LA-2 could not synthesize the amino acids methionine and histidine (met^-, his^-). Prototrophs (phe^+, trp^+, met^+, his^+) were recovered at a rate of about $1/10^5$ (10^{-5}) cells.

Although these observations at first suggested that the recombination involved was the type observed earlier in conjugative strains of *E. coli,* experiments using the Davis U-tube soon showed otherwise (**Figure 8–14**). The two auxotrophic strains were separated by a sintered glass filter, thus preventing cell contact but allowing growth to occur in a common medium. Surprisingly, when samples were removed from both sides of the filter and plated independently on minimal medium, prototrophs *were* recovered, but only from the side of the tube containing LA-22 bacteria. Recall that if conjugation were responsible, the conditions in the Davis U-tube would be expected to *prevent* recombination altogether (see Figure 8–3).

Since LA-2 cells appeared to be the source of the new genetic information (phe^+ and trp^+), how that information crossed the filter from the LA-2 cells to the LA-22 cells, allowing recombination to occur, was a mystery. The unknown source was designated simply as a *filterable agent* (*FA*).

Three observations were used to identify the FA:

1. The FA was produced by the LA-2 cells only when they were grown in association with LA-22 cells. If LA-2 cells were grown independently and that culture medium was then added to LA-22 cells, recombination did not occur. Therefore, LA-22 cells play some role in the production of FA by LA-2 cells and do so only when they share a common growth medium.
2. The addition of DNase, which enzymatically digests DNA, did not render the FA ineffective. Therefore, the FA is not naked DNA, ruling out transformation.
3. The FA could not pass across the filter of the Davis U-tube when the pore size was reduced below the size of bacteriophages.

Aided by these observations and aware that temperate phages can lysogenize *Salmonella,* researchers proposed that the genetic recombination event is mediated by bacteriophage P22, present initially as a prophage

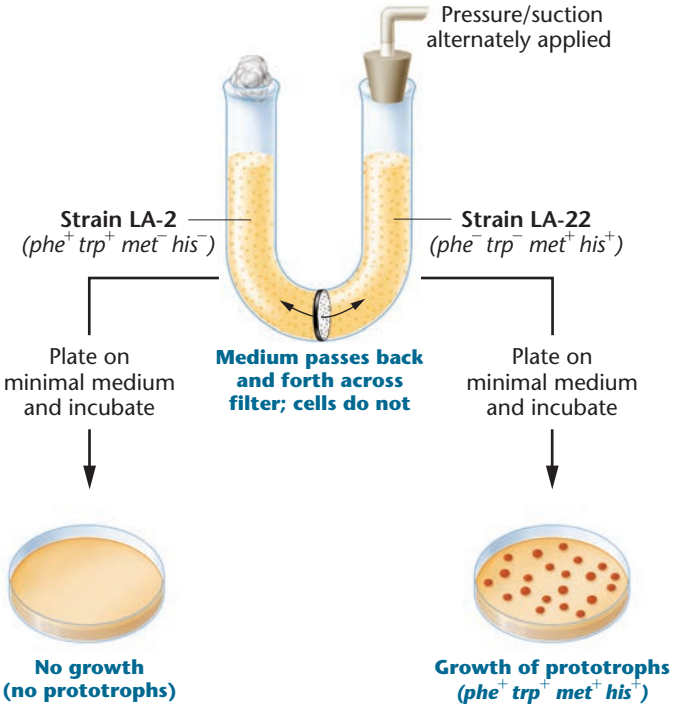


FIGURE 8–14 The Lederberg–Zinder experiment using *Salmonella*. After placing two auxotrophic strains on opposite sides of a Davis U-tube, Lederberg and Zinder recovered prototrophs from the side with the LA-22 strain, but not from the side containing the LA-2 strain.

in the chromosome of the LA-22 *Salmonella* cells. They hypothesized that P22 prophages rarely enter the vegetative or lytic phase, reproduce, and are released by the LA-22 cells. Such P22 phages, being much smaller than a bacterium, then cross the filter of the U-tube and subsequently infect and lyse some of the LA-2 cells. In the process of lysis of LA-2, these P22 phages occasionally package a region of the LA-2 chromosome in their heads. If this region contains the phe^+ and trp^+ genes and the phages subsequently pass back across the filter and infect LA-22 cells, these newly lysogenized cells will behave as prototrophs. This process of transduction, whereby bacterial recombination is mediated by bacteriophage P22, is diagrammed in **Figure 8–15**.

Transduction and Mapping

Like transformation, transduction was used in linkage and mapping studies of the bacterial chromosome. The fragment of bacterial DNA involved in a transduction event is large enough to include numerous genes. As a result, two genes that closely align (are linked) on the bacterial chromosome can be simultaneously transduced, a process called **cotransduction.** Two genes that are not close enough to one another along the chromosome to be included on a single DNA fragment require two independent events to be transduced into a single cell. Since this occurs with a much lower probability than cotransduction, linkage can be determined.

By concentrating on two or three linked genes, transduction studies can also determine the precise order of these genes. The closer linked genes are to each other, the greater the frequency of cotransduction. Mapping studies involving three closely aligned genes can thus be executed, and the analysis of such an experiment is predicated on the same rationale underlying other mapping techniques.

ESSENTIAL POINT

Transduction is virus-mediated bacterial DNA transfer and can be used to map phage genes. ■

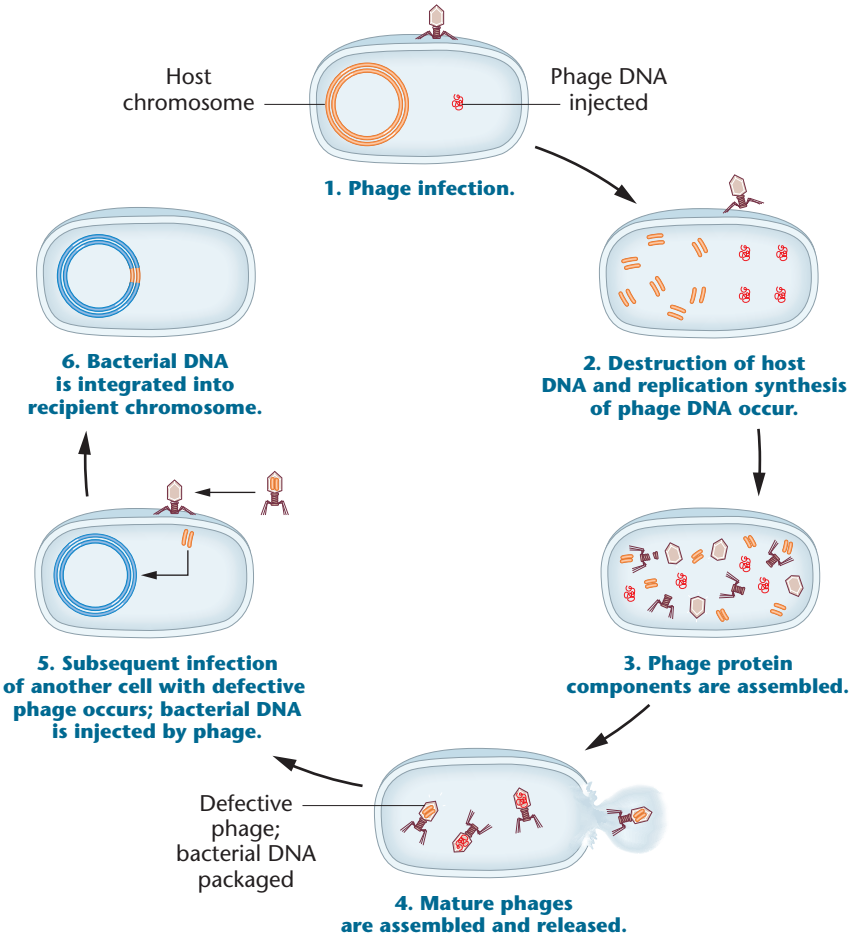


FIGURE 8–15 The process of transduction, where bacteriophages mediate bacterial recombination.

CASE STUDY | To treat or not to treat

A 4-month-old infant had been running a moderate fever for 36 hours, and a nervous mother made a call to her pediatrician. Examination and testing revealed no outward signs of infection or cause of the fever. The anxious mother asked the pediatrician about antibiotics, but the pediatrician recommended watching the infant carefully for two days before making a decision. He explained that decades of rampant use of antibiotics in medicine and agriculture had caused a worldwide surge in bacteria that are now resistant to such drugs. He also said that the reproductive behavior of bacteria allows them to exchange antibiotic resistance traits with a wide range of other disease-causing bacteria, and that many strains are now resistant to multiple antibiotics. The physician's information raises several interesting questions.

1. Was the physician correct in saying that bacteria can share resistance?
2. Where do bacteria carry antibiotic resistance genes, and how are they exchanged?
3. If the infant were given an antibiotic as a precaution, how might it contribute to the production of resistant bacteria?
4. Aside from hospitals, where else would infants and children come in contact with antibiotic-resistant strains of bacteria? Does the presence of such bacteria in the body always mean an infection?

INSIGHTS AND SOLUTIONS

1. Time mapping is performed in a cross involving the genes *his, leu, mal,* and *xyl.* The recipient cells are auxotrophic for all four genes. After 25 minutes, mating is interrupted, with the results in recipient cells shown below. Diagram the positions of these genes relative to the origin (*O*) of the F factor and to one another.

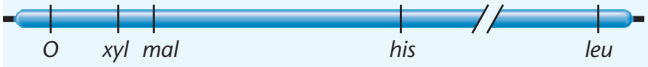
   (a) 90% are *xyl*$^+$

   (b) 80% are *mal*$^+$

   (c) 20% are *his*$^+$

   (d) None are *leu*$^+$

   Solution: The *xyl* gene is transferred most frequently, so it is closest to *O* (very close). The *mal* gene is next and reasonably close to *xyl,* followed by the more distant *his* gene. The *leu* gene is far beyond these three, since no recovered recombinants include it. The diagram shows these relative locations along a piece of the circular chromosome.



2. In four Hfr strains of bacteria, all derived from an original F$^+$ culture grown over several months, a group of hypothetical genes is studied and shown to transfer in the orders shown in the following table. (a) Assuming *b* is the first gene along the chromosome, determine the sequence of all genes shown. (b) One strain creates an apparent dilemma. Which one is it? Explain why the dilemma is only apparent, not real.

Hfr Strain	Order of Transfer
1	*e r i u m b*
2	*u m b a c t*
3	*c t e r i u*
4	*r e t c a b*

Solution:

(a) The sequence is found by overlapping the genes in each strain.

Strain 2 *u m* ***b a c*** *t*
Strain 3 *c* ***t e r*** *i u*
Strain 1 *e r* ***i u m*** *b*

Starting with *b* in strain 2, the gene sequence is *bacterium.*

(b) Strain 4 creates a dilemma, which is resolved when we realize that the F factor is integrated in the opposite orientation. Thus, the genes enter in the opposite sequence, starting with gene *r.*

retcab →

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we have focused on genetic systems present in bacteria and the viruses that use bacteria as hosts (bacteriophages). In particular, we discussed mechanisms by which bacteria and their phages undergo genetic recombination, the basis of chromosome mapping. Based on your knowledge of these topics, answer several fundamental questions:
   (a) How do we know that bacteria undergo genetic recombination, allowing the transfer of genes from one organism to another?
   (b) How do we know that conjugation leading to genetic recombination between bacteria involves cell contact, which precedes the transfer of genes from one bacterium to another?
   (c) How do we know that during transduction bacterial cell-to-cell contact is not essential?
   (d) How do we know that intergenic exchange occurs in bacteriophages?

CONCEPT QUESTION

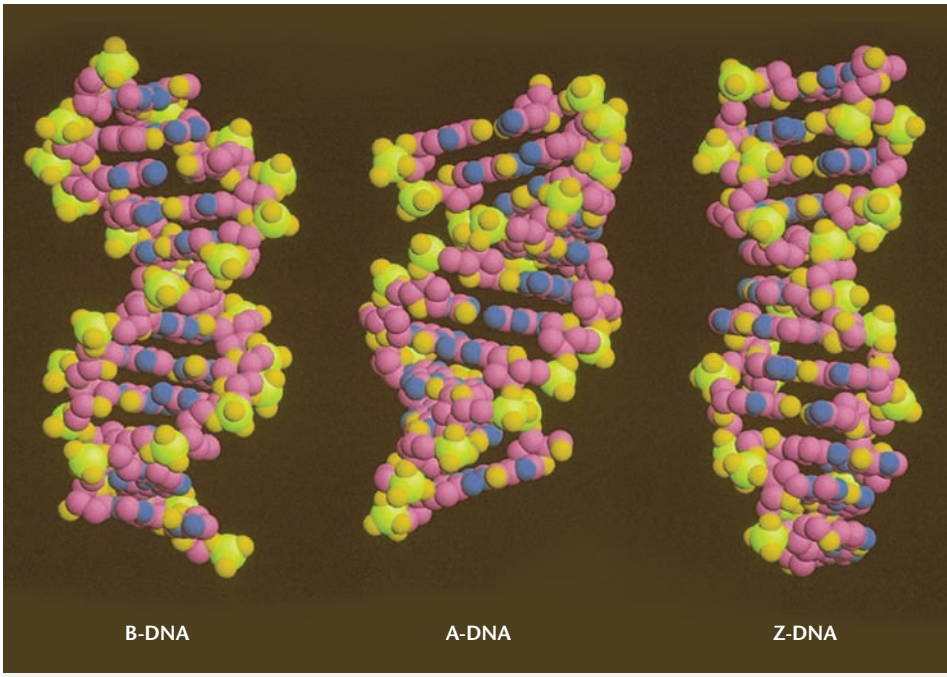
2. Review the Chapter Concepts list on p. 159. Many of these center around the findings that genetic recombination occurs in bacteria and in bacteriophages. Write a short summary that contrasts how recombination occurs in bacteria and bacteriophages. ■
3. Distinguish among the three modes of recombination in bacteria.
4. With respect to F^+ and F^- bacterial matings,
   (a) How was it established that physical contact was necessary?
   (b) How was it established that chromosome transfer was unidirectional?
   (c) What is the genetic basis of a bacterium being F^+?
5. List all of the differences between $F^+ \times F^-$ and Hfr $\times$ F^- bacterial crosses and between F^+, F^-, Hfr, and F′ bacteria.
6. Explain how the interrupted mating technique used in the Hfr $\times$ F^- crosses served as the basis for the first genetic mapping of the *E. coli* chromosome.
7. Describe how different strains of *E. coli* can reveal different linkage arrangements of genes in Hfr crosses.
8. Describe the origin of F′ bacteria and merozygotes.
9. Describe the mechanism of transformation.
10. The bacteriophage genome consists primarily of genes encoding proteins that make up the head, collar and tail, and tail fibers. When these genes are transcribed following phage infection, how are these proteins synthesized, since the phage genome lacks genes essential to ribosome structure?
11. Describe the temporal sequence of the bacteriophage life cycle.
12. In the plaque assay, what is the precise origin of a single plaque?
13. Explain the connection between plaque formation and lytic bacteriophage.
14. A plaque assay is performed beginning with 1.0 mL of a solution containing bacteriophages. This solution is serially diluted three times by taking 0.1 mL and adding it to 9.9 mL of liquid medium. The final dilution is plated and yields 17 plaques. What is the initial density of bacteriophages in the original 1.0 mL?
15. Describe the difference between the lytic cycle and lysogeny when bacteriophage infection occurs.
16. Distinguish between temperate and virulent bacteriophages.
17. Explain the observations that led Zinder and Lederberg to conclude that the prototrophs recovered in their transduction experiments were not the result of Hfr-mediated conjugation.
18. Describe the execution of and rationale behind linkage and mapping studies of bacterial genes during transduction experiments.
19. Assume that one counted 67 plaques on a bacterial plate where 0.1 ml of a 10-5 dilution of phage was added to bacterial culture. What was the initial concentration of the undiluted phage?
20. A phage-infected bacterial culture was subjected to a series of dilutions, and a plaque assay was performed in each case, with the following results. What conclusion can be drawn in the case of each dilution?

	Dilution Factor	Assay Results
(a)	10^4	All bacteria lysed
(b)	10^6	14 plaques
(c)	10^8	0 plaques

9 DNA Structure and Analysis

CHAPTER CONCEPTS

- With the exception of some viruses, DNA serves as the genetic material in all living organisms on Earth.
- According to the Watson–Crick model, DNA exists in the form of the right-handed double helix.
- The strands of the double helix are antiparallel and held together by hydrogen bonding between complementary nitrogenous bases.
- The structure of DNA provides the basis for storing and expressing genetic information.
- RNA has many similarities to DNA but exists mostly as a single-stranded molecule.
- In some viruses, RNA serves as the genetic material.
- Many techniques have been developed that facilitate the analysis of nucleic acids, most based on detection of the complementarity of nitrogenous bases.



Computer-generated space-filling models of alternative forms of DNA.

Up to this point in the text, we have described chromosomes as structures containing genes that control phenotypic traits that are transmitted through gametes to future offspring. Logically, genes must contain some sort of information that, when passed to a new generation, influences the form and characteristics of each individual. We refer to that information as the **genetic material.** Logic also suggests that this same information in some way directs the many complex processes that lead to an organism's adult form.

Until 1944, it was not clear what chemical component of the chromosome makes up genes and constitutes the genetic material. Because chromosomes were known to have both a nucleic acid and a protein component, both were candidates. In 1944, however, direct experimental evidence emerged showing that the nucleic acid DNA serves as the informational basis for heredity.

Once the importance of DNA in genetic processes was realized, work intensified with the hope of discerning not only the structural basis of this molecule but also the relationship of its structure to its function. Between 1944 and 1953, many scientists sought information that might answer the most significant and intriguing question in the history of biology: How does DNA serve as the genetic basis for the living process? Researchers believed the answer depended strongly on the chemical structure of the DNA molecule, given the complex but orderly functions ascribed to it.

These efforts were rewarded in 1953 when James Watson and Francis Crick set forth their hypothesis for the double-helical nature of DNA. The

assumption that the molecule's functions would be clarified more easily once its general structure was determined proved to be correct. In this chapter, we initially review the evidence that DNA is the genetic material and then discuss the elucidation of its structure.

9.1 The Genetic Material Must Exhibit Four Characteristics

For a molecule to serve as the genetic material, it must possess four major characteristics: **replication, storage of information, expression of information,** and **variation by mutation.** Replication of the genetic material is one facet of the cell cycle, a fundamental property of all living organisms. Once the genetic material of cells replicates and is doubled in amount, it must then be partitioned equally into daughter cells. During the formation of gametes, the genetic material is also replicated but is partitioned so that each cell gets only one-half of the original amount of genetic material—the process of meiosis. Although the products of mitosis and meiosis differ, both of these processes are part of the more general phenomenon of cellular reproduction.

Storage of information requires the molecule to act as a repository of genetic information that may or may not be expressed by the cell in which it resides. It is clear that while most cells contain a complete copy of the organism's genome, at any point in time they express only a part of this genetic potential. For example, in bacteria many genes "turn on" in response to specific environmental conditions and "turn off" when conditions change. In vertebrates, skin cells may display active melanin genes but never activate their hemoglobin genes; in contrast, digestive cells activate many genes specific to their function but do not activate their melanin genes.

Expression of the stored genetic information is the basis of the process of **information flow** within the cell (**Figure 9–1**). The initial event is the **transcription** of DNA, in which three main types of RNA molecules are synthesized: messenger RNA (mRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA). Of these, mRNAs are translated into proteins. Each mRNA is the product of a specific gene and directs the synthesis of a different protein. In **translation,** the chemical information in mRNA directs the construction of a chain of amino acids, called a polypeptide, which then folds into a protein. Collectively, these processes form the **central dogma of molecular genetics:** "DNA makes RNA, which makes proteins."

The genetic material is also the source of variation among organisms through the process of mutation. If a mutation—a change in the chemical composition of DNA—occurs, the alteration may be reflected during transcription and translation, affecting the specific protein. If such a mutation is present in gametes, it may be passed to future generations and, with time, become distributed throughout the population. Genetic variation, which also includes alterations of chromosome number and rearrangements within and between chromosomes, provides the raw material for the process of evolution.

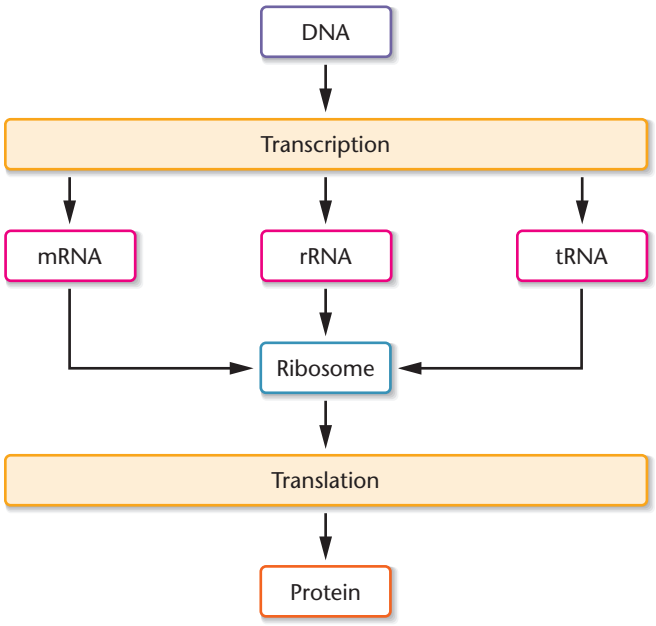


FIGURE 9–1 Simplified view of information flow (the central dogma) involving DNA, RNA, and proteins within cells.

9.2 Until 1944, Observations Favored Protein as the Genetic Material

The idea that genetic material is physically transmitted from parent to offspring has been accepted for as long as the concept of inheritance has existed. Beginning in the late nineteenth century, research into the structure of biomolecules progressed considerably, setting the stage for describing the genetic material in chemical terms. Although both proteins and nucleic acid were major candidates for the role of the genetic material, until the 1940s many geneticists favored proteins. This is not surprising because a diversity of proteins was known to be abundant in cells, and much more was known about protein chemistry.

DNA was first studied in 1868 by a Swiss chemist, Friedrich Miescher. He isolated cell nuclei and derived an acid substance containing DNA that he called **nuclein.** As investigations progressed, however, DNA, which was shown to be present in chromosomes, seemed to lack the chemical diversity necessary to store extensive genetic

information. This conclusion was based largely on Phoebus A. Levene's observations in 1910 that DNA contained approximately equal amounts of four similar molecules called *nucleotides.* Levene postulated incorrectly that identical groups of these four components were repeated over and over, which was the basis of his **tetranucleotide hypothesis** for DNA structure. Attention was thus directed away from DNA, favoring proteins. However, in the 1940s, Erwin Chargaff showed that Levene's proposal was incorrect when he demonstrated that most organisms do not contain precisely equal proportions of the four nucleotides. We shall see later that the structure of DNA accounts for Chargaff's observations.

ESSENTIAL POINT

Although both proteins and nucleic acids were initially considered as possible candidates, proteins were initially favored to serve as the genetic material. ■

9.3 Evidence Favoring DNA as the Genetic Material Was First Obtained during the Study of Bacteria and Bacteriophages

Oswald Avery, Colin MacLeod, and Maclyn McCarty's 1944 publication on the chemical nature of a "transforming principle" in bacteria was the initial event that led to the acceptance of DNA as the genetic material. Their work, along with subsequent findings of other research teams, constituted the first direct experimental proof that DNA, and not protein, is the biomolecule responsible for heredity. It marked the beginning of the era of molecular genetics, a period of discovery in biology that made biotechnology feasible and has moved us closer to understanding the basis of life. The impact of the initial findings on future research and thinking paralleled that of the publication of Darwin's theory of evolution and the subsequent rediscovery of Mendel's postulates of transmission genetics. Together, these events constitute the three great revolutions in biology.

Transformation Studies

The research that provided the foundation for Avery, MacLeod, and McCarty's work was initiated in 1927 by Frederick Griffith, a medical officer in the British Ministry of Health. He experimented with several different strains of the bacterium *Diplococcus pneumoniae.** Some were **virulent strains,** which cause pneumonia in certain vertebrates (notably humans and mice), while others were **avirulent strains,** which do not cause illness.

The difference in virulence depends on the existence of a polysaccharide capsule; virulent strains have this capsule, whereas avirulent strains do not. The nonencapsulated bacteria are readily engulfed and destroyed by phagocytic cells in the animal's circulatory system. Virulent bacteria, which possess the polysaccharide coat, are not easily engulfed; they multiply and cause pneumonia.

The presence or absence of the capsule causes a visible difference between colonies of virulent and avirulent strains. Encapsulated bacteria form **smooth colonies** (*S*) with a shiny surface when grown on an agar culture plate; nonencapsulated strains produce **rough colonies** (*R*). Thus, virulent and avirulent strains are easily distinguished by standard microbiological culture techniques.

Each strain of *Diplococcus* may be one of dozens of different types called **serotypes.** The specificity of the serotype is due to the detailed chemical structure of the polysaccharide constituent of the thick, slimy capsule. Serotypes are identified by immunological techniques and are usually designated by Roman numerals. Griffith used the avirulent type II*R* and the virulent type III*S* in his critical experiments. Table 9.1 summarizes the characteristics of these strains.

Griffith knew from the work of others that only living virulent cells produced pneumonia in mice. If heat-killed virulent bacteria were injected into mice, no pneumonia resulted, just as living avirulent bacteria failed to produce the disease. Griffith's critical experiment involved injecting mice with living II*R* (avirulent) cells combined with heat-killed III*S* (virulent) cells. Since neither cell type caused death in mice when injected alone, Griffith expected that the double injection would not kill the mice. But, after five days, all of the mice that had received both types of cells were dead. Paradoxically, analysis of their blood revealed large numbers of living type III*S* bacteria.

As far as could be determined, these III*S* bacteria were identical to the III*S* strain from which the heat-killed cell preparation had been made. Control mice, injected only with living avirulent II*R* bacteria, did not develop pneumonia and remained healthy. This ruled out the possibility that the avirulent II*R* cells simply changed (or mutated) to virulent III*S* cells in the absence of the heat-killed III*S* bacteria. Instead, some type of interaction had taken place between living II*R* and heat-killed III*S* cells.

TABLE 9.1 Strains of *Diplococcus pneumoniae* Used by Frederick Griffith in His Original Transformation Experiments

Serotype	Colony Morphology	Capsule	Virulence
II*R*	Rough	Absent	Avirulent
III*S*	Smooth	Present	Virulent

*This organism is now named *Streptococcus pneumoniae.*

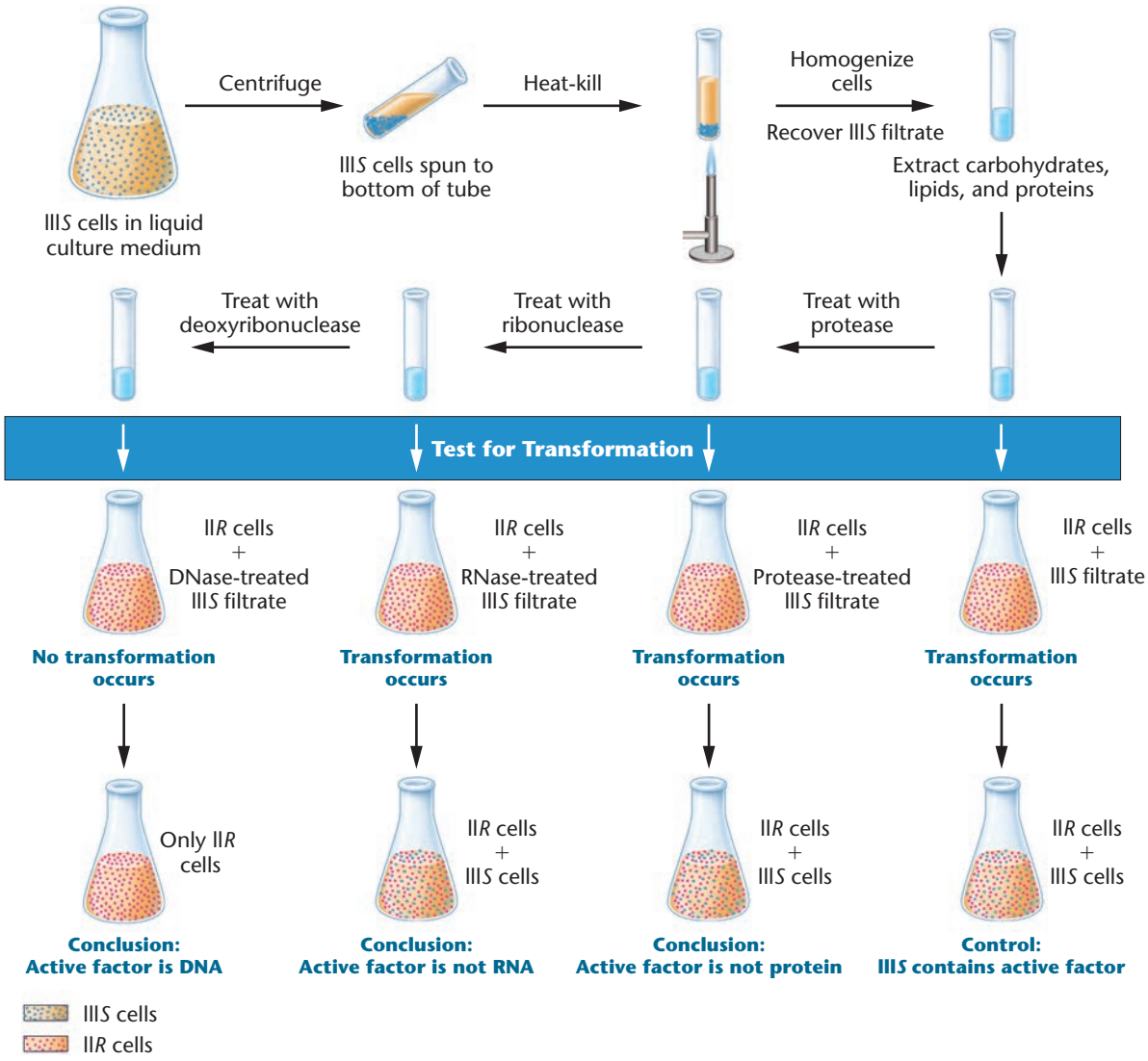


FIGURE 9–2 Summary of Avery, MacLeod, and McCarty's experiment demonstrating that DNA is the transforming principle.

Griffith concluded that the heat-killed III*S* bacteria somehow converted live avirulent II*R* cells into virulent III*S* cells. Calling the phenomenon **transformation,** he suggested that the **transforming principle** might be some part of the polysaccharide capsule or a compound required for capsule synthesis, although the capsule alone did not cause pneumonia. To use Griffith's term, the transforming principle from the dead III*S* cells served as a "pabulum" for the II*R* cells.

Griffith's work led others to explore the phenomenon of transformation. By 1931, Henry Dawson and his coworkers showed that transformation could occur *in vitro* (in a test tube containing only bacterial cells). That is, injection into mice was not necessary for transformation to occur. By 1933, Lionel J. Alloway had refined the *in vitro* experiments using extracts from *S* cells added to living *R* cells. The soluble filtrate from the heat-killed III*S* cells was as effective in inducing transformation as were the intact cells. Alloway and others did not view transformation as a genetic event, but rather as a physiological modification of some sort. Nevertheless, the experimental evidence that a chemical substance was responsible for transformation was quite convincing.

Then, in 1944, after ten years of work, Avery, MacLeod, and McCarty published their results in what is now regarded as a classic paper in the field of molecular genetics. They reported that they had obtained the transforming principle in a highly purified state and that they strongly believed that it was DNA.

The details of their work are illustrated in **Figure 9–2**. The researchers began their isolation procedure with large quantities (50–75 L) of liquid cultures of type III*S* virulent cells. The cells were centrifuged, collected, and heat-killed. Following various chemical treatments, a soluble filtrate was derived from these cells, which retained the ability to induce transformation of type II*R* avirulent

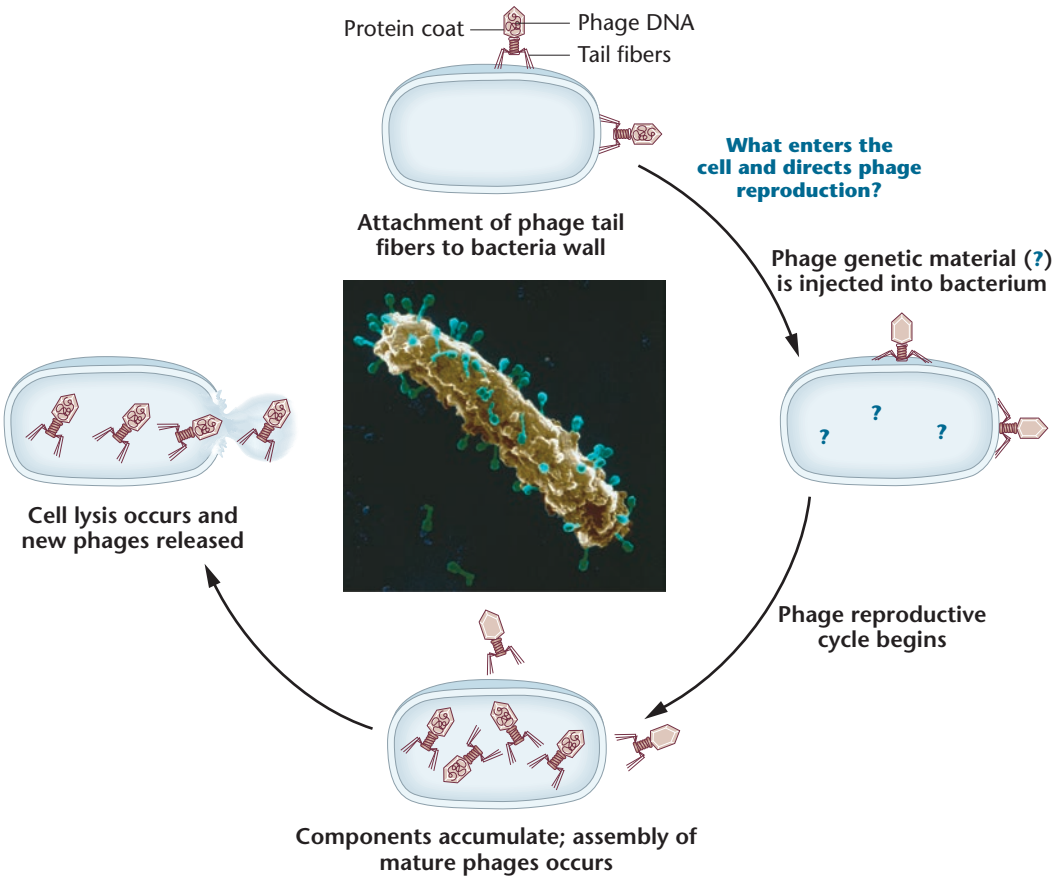


FIGURE 9–3 Life cycle of a T-even bacteriophage. The electron micrograph shows an *E. coli* cell during infection by numerous T2 phages (shown in blue).

cells. The soluble filtrate was treated with a protein-digesting enzyme, called a protease, and an RNA-digesting enzyme, called **ribonuclease.** Such treatment destroyed the activity of any remaining protein and RNA. Nevertheless, transforming activity still remained. They concluded that neither protein nor RNA was responsible for transformation. The final confirmation came with experiments using crude samples of the DNA-digesting enzyme **deoxyribonuclease,** isolated from dog and rabbit sera. Digestion with this enzyme destroyed transforming activity present in the filtrate; thus, Avery and his coworkers were certain that the active transforming principle in these experiments was DNA.

The great amount of work, the confirmation and reconfirmation of the conclusions, and the logic of the experimental design involved in the research of these three scientists are truly impressive. Their conclusion in the 1944 publication, however, was stated very simply: "The evidence presented supports the belief that a nucleic acid of the desoxyribose* type is the fundamental unit of the transforming principle of *Pneumococcus* type III."

The researchers also immediately recognized the genetic and biochemical implications of their work. They suggested that the transforming principle interacts with the II*R* cell and gives rise to a coordinated series of enzymatic reactions that culminates in the synthesis of the type III*S* capsular polysaccharide. They emphasized that, once transformation occurs, the capsular polysaccharide is produced in successive generations. Transformation is therefore heritable, and the process affects the genetic material.

Transformation, originally introduced in Chapter 8, has now been shown to occur in *Hemophilus influenzae, Bacillus subtilis, Shigella paradysenteriae,* and *Escherichia coli,* among many other microorganisms. Transformation of numerous genetic traits other than colony morphology has been demonstrated, including those that resist antibiotics. These observations further strengthened the belief that transformation by DNA is primarily a genetic event rather than simply a physiological change.

The Hershey–Chase Experiment

The second major piece of evidence supporting DNA as the genetic material was provided during the study of the bacterium *E. coli* and one of its infecting viruses, **bacteriophage T2.** Often referred to simply as a **phage,** the virus consists of a protein coat surrounding a core of DNA. Electron micrographs reveal that the phage's external structure is composed of a hexagonal head plus a tail. Figure 9–3 shows the life cycle

* Desoxyribose is now spelled deoxyribose.

of a T-even bacteriophage such as T2, as it was known in 1952. Recall that the phage adsorbs to the bacterial cell and that some component of the phage enters the bacterial cell. Following infection, the viral information "commandeers" the cellular machinery of the host and undergoes viral reproduction. In a reasonably short time, many new phages are constructed and the bacterial cell is lysed, releasing the progeny viruses.

In 1952, Alfred Hershey and Martha Chase published the results of experiments designed to clarify the events leading to phage reproduction. Several of the experiments clearly established the independent functions of phage protein and nucleic acid in the reproduction process of the bacterial cell. Hershey and Chase knew from this existing data that:

1. T2 phages consist of approximately 50 percent protein and 50 percent DNA.
2. Infection is initiated by adsorption of the phage by its tail fibers to the bacterial cell.
3. The production of new viruses occurs within the bacterial cell.

It appeared that some molecular component of the phage, DNA and/or protein, entered the bacterial cell and directed viral reproduction. Which was it?

Hershey and Chase used radioisotopes to follow the molecular components of phages during infection. Both ^{32}P and ^{35}S, radioactive forms of phosphorus and sulfur, respectively, were used. DNA contains phosphorus but not sulfur, so ^{32}P effectively labels DNA. Because proteins contain sulfur, but not phosphorus, ^{35}S labels protein. *This is a key point in the experiment.* If *E. coli* cells are first grown in the presence of *either* ^{32}P *or* ^{35}S and then infected with T2 viruses, the progeny phage will have *either* a labeled DNA core *or* a labeled protein coat, respectively. These radioactive phages can be isolated and used to infect unlabeled bacteria (**Figure 9–4**).

When labeled phage and unlabeled bacteria were mixed, an adsorption complex was formed as the phages attached their tail fibers to the bacterial wall. These complexes were isolated and subjected to a high shear force by placing them in a blender. This force stripped off the attached phages, which were then analyzed separately. By tracing the radioisotopes, Hershey and Chase were able to demonstrate that most of the ^{32}P-labeled DNA had transferred into the bacterial cell following adsorption; on the other hand, almost all of the ^{35}S-labeled protein remained outside the bacterial cell and was recovered in the phage "ghosts" (empty phage coats) after the blender treatment. Following separation, the bacterial cells, which now contained viral DNA, were eventually lysed as new phages were produced. These progeny contained ^{32}P but not ^{35}S.

Hershey and Chase interpreted these results as indicating that the protein of the phage coat remains outside the host cell and is not involved in the production of new phages. On the other hand, and most important, phage DNA enters the host cell and directs phage reproduction. Hershey and Chase had demonstrated that the genetic material in phage T2 is DNA, not protein.

These experiments, along with those of Avery and his colleagues, provided convincing evidence that DNA is the molecule responsible for heredity. This conclusion has since served as the cornerstone of the field of molecular genetics.

NOW SOLVE THIS

9–1 Would an experiment similar to that performed by Hershey and Chase work if the basic design were applied to the phenomenon of transformation? Explain why or why not.

HINT: *This problem involves an understanding of the protocol of the Hershey–Chase experiment as applied to the investigation of transformation. The key to its solution is to remember that in transformation, exogenous DNA enters the soon-to-be transformed cell and that no cell-to-cell contact is involved in the process.*

Transfection Experiments

During the eight years following publication of the Hershey–Chase experiment, additional research with bacterial viruses provided even more solid proof that DNA is the genetic material. In 1957, several reports demonstrated that if *E. coli* is treated with the enzyme lysozyme, the outer wall of the cell can be removed without destroying the bacterium. Enzymatically treated cells are naked, so to speak, and contain only the cell membrane as the outer boundary of the cell; these structures are called **protoplasts** (or **spheroplasts**). John Spizizen and Dean Fraser independently reported that by using protoplasts, they were able to initiate phage multiplication with disrupted T2 particles. That is, provided protoplasts were used, a virus did not have to be intact for infection to occur.

Similar but more refined experiments were reported in 1960 using only DNA purified from bacteriophages. This process of infection by only the viral nucleic acid, called **transfection,** proves conclusively that phage DNA alone contains all the necessary information for producing

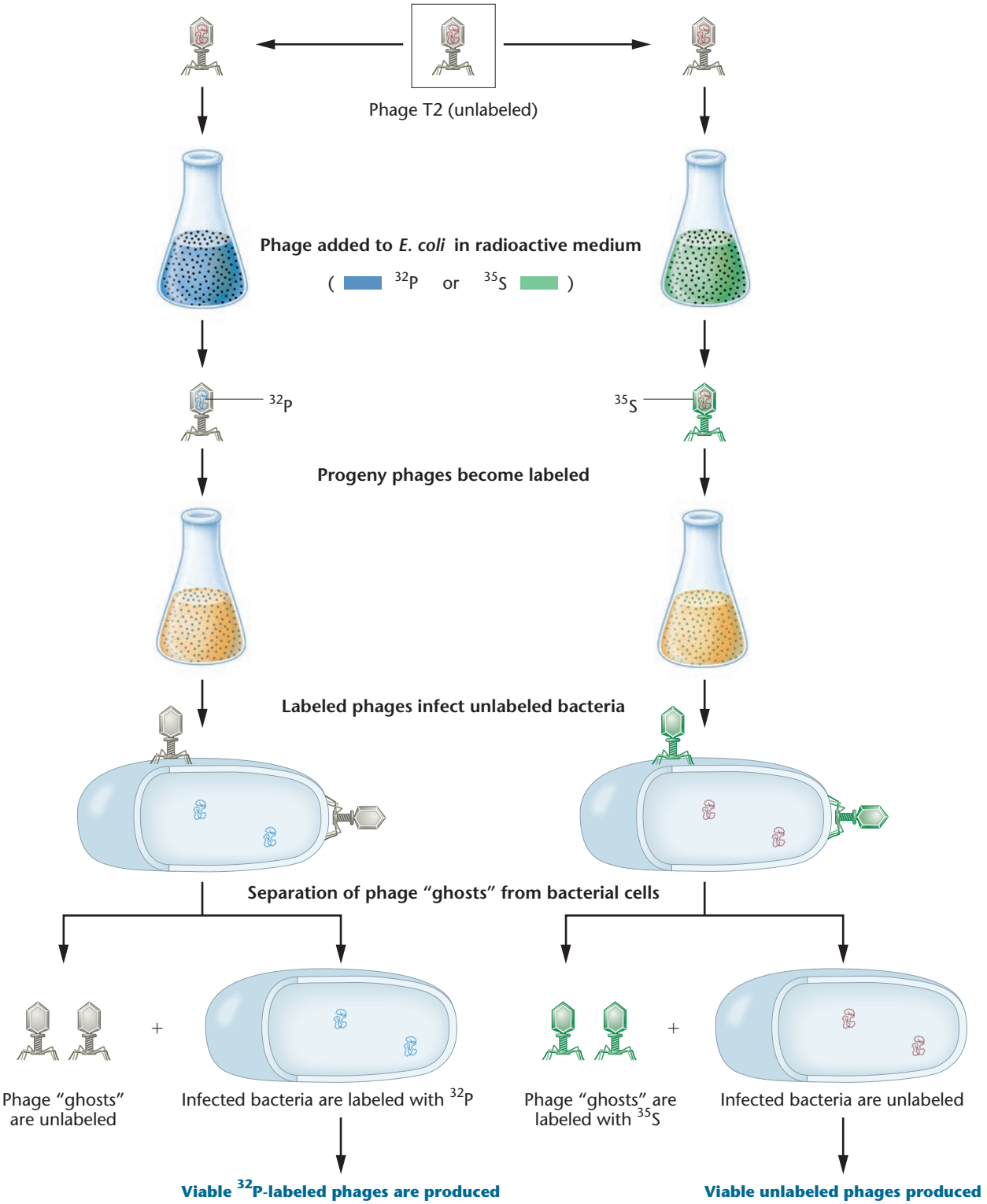


FIGURE 9–4 Summary of the Hershey–Chase experiment demonstrating that DNA, not protein, is responsible for directing the reproduction of phage T2 during the infection of *E. coli*.

mature viruses. Thus, the evidence that DNA serves as the genetic material in all organisms was further strengthened, even though all direct evidence had been obtained from bacterial and viral studies.

ESSENTIAL POINT

By 1952, transformation studies and experiments using bacteria infected with bacteriophages strongly suggested that DNA is the genetic material in bacteria and most viruses. ■

9.4 Indirect and Direct Evidence Supports the Concept that DNA Is the Genetic Material in Eukaryotes

In 1950, eukaryotic organisms were not amenable to the types of experiments that used bacteria and viruses to demonstrate that DNA is the genetic material. Nevertheless, it was generally assumed that the genetic material would be a universal substance and also serve this role in eukaryotes. Initially, support for this assumption relied on several circumstantial observations that, taken together, indicated that DNA is also the genetic material in eukaryotes. Subsequently, direct evidence established unequivocally the central role of DNA in genetic processes.

Indirect Evidence: Distribution of DNA

The genetic material should be found where it functions—in the nucleus as part of chromosomes. Both DNA and protein fit this criterion. However, protein is also abundant in the cytoplasm, whereas DNA is not. Both mitochondria and chloroplasts are known to perform genetic functions, and DNA is also present in these organelles. Thus, DNA is found only where primary genetic function is known to occur. Protein, however, is found everywhere in the cell. These observations are consistent with the interpretation favoring DNA over protein as the genetic material.

Because it had been established earlier that chromosomes within the nucleus contain the genetic material, a correlation was expected between the ploidy (*n*, 2*n*, etc.) of cells and the quantity of the molecule that functions as the genetic material. Meaningful comparisons can be made between gametes (sperm and eggs) and somatic or body cells. The somatic cells are recognized as being diploid (2*n*) and containing twice the number of chromosomes as gametes, which are haploid (*n*).

Table 9.2 compares the amount of DNA found in haploid sperm and the diploid nucleated precursors of red blood cells from a variety of organisms. The amount of DNA and the number of sets of chromosomes are closely correlated. No consistent correlation can be observed between gametes and diploid cells for proteins, thus again favoring DNA over proteins as the genetic material of eukaryotes.

TABLE 9.2 DNA Content of Haploid Versus Diploid Cells of Various Species (in picograms)*

Organism	*n*	2*n*
Human	3.25	7.30
Chicken	1.26	2.49
Trout	2.67	5.79
Carp	1.65	3.49
Shad	0.91	1.97

*Sperm (*n*) and nucleated precursors to red blood cells (2*n*) were used to contrast ploidy levels.

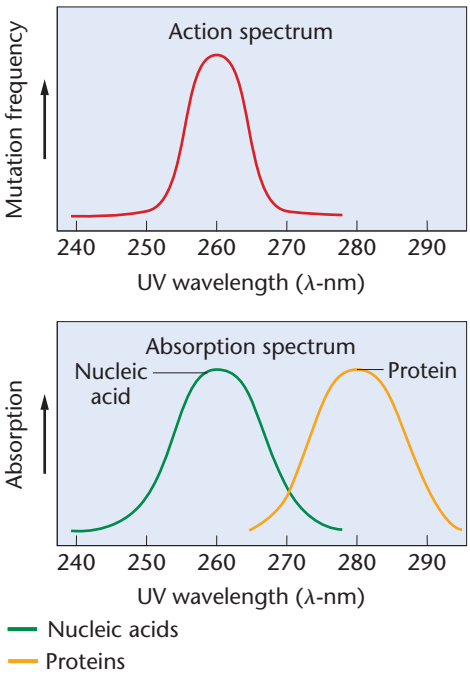


FIGURE 9–5 Comparison of the action spectrum, which determines the most effective mutagenic UV wavelength, and the absorption spectrum, which shows the range of wavelengths where nucleic acids and proteins absorb UV light.

Indirect Evidence: Mutagenesis

Ultraviolet (UV) light is one of many agents capable of inducing mutations in the genetic material. Simple organisms such as yeast and other fungi can be irradiated with various wavelengths of UV light, and the effectiveness of each wavelength can then be measured by the number of mutations it induces. When the data are plotted, an **action spectrum** of UV light as a mutagenic agent is obtained. This action spectrum can then be compared with the **absorption spectrum** of any molecule suspected to be genetic material (Figure 9–5). *The molecule serving as the genetic material is expected to absorb at the wavelengths shown to be mutagenic.*

UV light is most mutagenic at the wavelength (λ) of about 260 nanometers (nm), and both DNA and RNA strongly absorb UV light at 260 nm. On the other hand, protein absorbs most strongly around 280 nm, yet no significant mutagenic effects are observed at this wavelength. This indirect evidence also supports the idea that a nucleic acid is the genetic material and tends to exclude protein.

Direct Evidence: Recombinant DNA Studies

Although the circumstantial evidence described earlier does not constitute direct proof that DNA is the genetic material

in eukaryotes, these observations spurred researchers to forge ahead, basing their work on this hypothesis. Today, there is no doubt of the validity of this conclusion. DNA *is* the genetic material in eukaryotes. The strongest evidence is provided by molecular analysis utilizing **recombinant DNA technology.** In this procedure, segments of eukaryotic DNA corresponding to specific genes are isolated and spliced into bacterial DNA. This complex can then be inserted into a bacterial cell, and its genetic expression is monitored. If a eukaryotic gene is introduced, the presence of the corresponding eukaryotic protein product demonstrates directly that this DNA is functional in the bacterial cell. This has been shown to be the case in countless instances. For example, the products of the human genes specifying insulin and interferon are produced by bacteria after they have incorporated the human genes that encode these proteins. As the bacterium divides, the eukaryotic DNA replicates along with the host DNA and is distributed to the daughter cells, which also express the human genes and synthesize the corresponding proteins.

The availability of vast amounts of DNA coding for specific genes, derived from recombinant DNA research, has led to other direct evidence that DNA serves as the genetic material. Work done in the laboratory of Beatrice Mintz demonstrated that DNA encoding the human β-globin protein, when microinjected into a fertilized mouse egg, is later found to be present and expressed in adult mouse tissue, and it is transmitted to and expressed in that mouse's progeny. These mice are examples of **transgenic animals,** which are now commonplace in genetic research. They clearly demonstrate that DNA meets the requirement of expression of genetic information in eukaryotes.

ESSENTIAL POINT

Although at first only indirect observations supported the hypothesis that DNA controls inheritance in eukaryotes, subsequent studies involving recombinant DNA techniques and transgenic mice provided direct experimental evidence that the eukaryotic genetic material is DNA. ■

9.5 RNA Serves as the Genetic Material in Some Viruses

Some viruses contain an RNA core rather than a DNA core. In these viruses, it appears that RNA serves as the genetic material—an exception to the general rule that DNA performs this function. In 1956, it was demonstrated that when purified RNA from **tobacco mosaic virus (TMV)** was spread on tobacco leaves, the characteristic lesions caused by viral infection subsequently appeared on the leaves. Thus, it was concluded that RNA is the genetic material of this virus.

In 1965 and 1966, Norman Pace and Sol Spiegelman demonstrated further that RNA from the phage Qβ can be isolated and replicated *in vitro.* Replication depends on an enzyme, **RNA replicase,** which is isolated from host *E. coli* cells following normal infection. When the RNA replicated *in vitro* is added to *E. coli* protoplasts, infection and viral multiplication (transfection) occur. Thus, RNA synthesized in a test tube serves as the genetic material in these phages by directing the production of all the components necessary for viral reproduction.

One other group of RNA-containing viruses bears mention. These are the **retroviruses,** which replicate in an unusual way. Their RNA serves as a template for the synthesis of the complementary DNA molecule. The process, **reverse transcription,** occurs under the direction of an RNA-dependent DNA polymerase enzyme called **reverse transcriptase.** This DNA intermediate can be incorporated into the genome of the host cell, and when the host DNA is transcribed, copies of the original retroviral RNA chromosomes are produced. Retroviruses include the human immunodeficiency virus (HIV), which causes AIDS, as well as RNA tumor viruses.

ESSENTIAL POINT

RNA serves as the genetic material in some bacteriophages as well as some plant and animal viruses. ■

9.6 The Structure of DNA Holds the Key to Understanding Its Function

Having established that DNA is the genetic material in all living organisms (except certain viruses), we turn now to the structure of this nucleic acid. In 1953, James Watson and Francis Crick proposed that the structure of DNA is in the form of a double helix. Their proposal was published in a short paper in the journal *Nature*. In a sense, this publication was the finish of a highly competitive scientific race to obtain what some consider to be the most significant finding in the history of biology. This race, as recounted in Watson's book *The Double Helix,* demonstrates the human interaction, genius, frailty, and intensity involved in the scientific effort that eventually led to elucidation of the DNA structure.

The data available to Watson and Crick, crucial to the development of their proposal, came primarily from two sources: (1) base composition analysis of hydrolyzed samples of DNA and (2) X-ray diffraction studies of DNA. Watson and Crick's analytical success can be attributed to model building that conformed to the existing data. If the correct solution to the structure of DNA is viewed as a puzzle, Watson and Crick, working in the Cavendish Laboratory in Cambridge, England, were the first to fit the pieces together successfully.

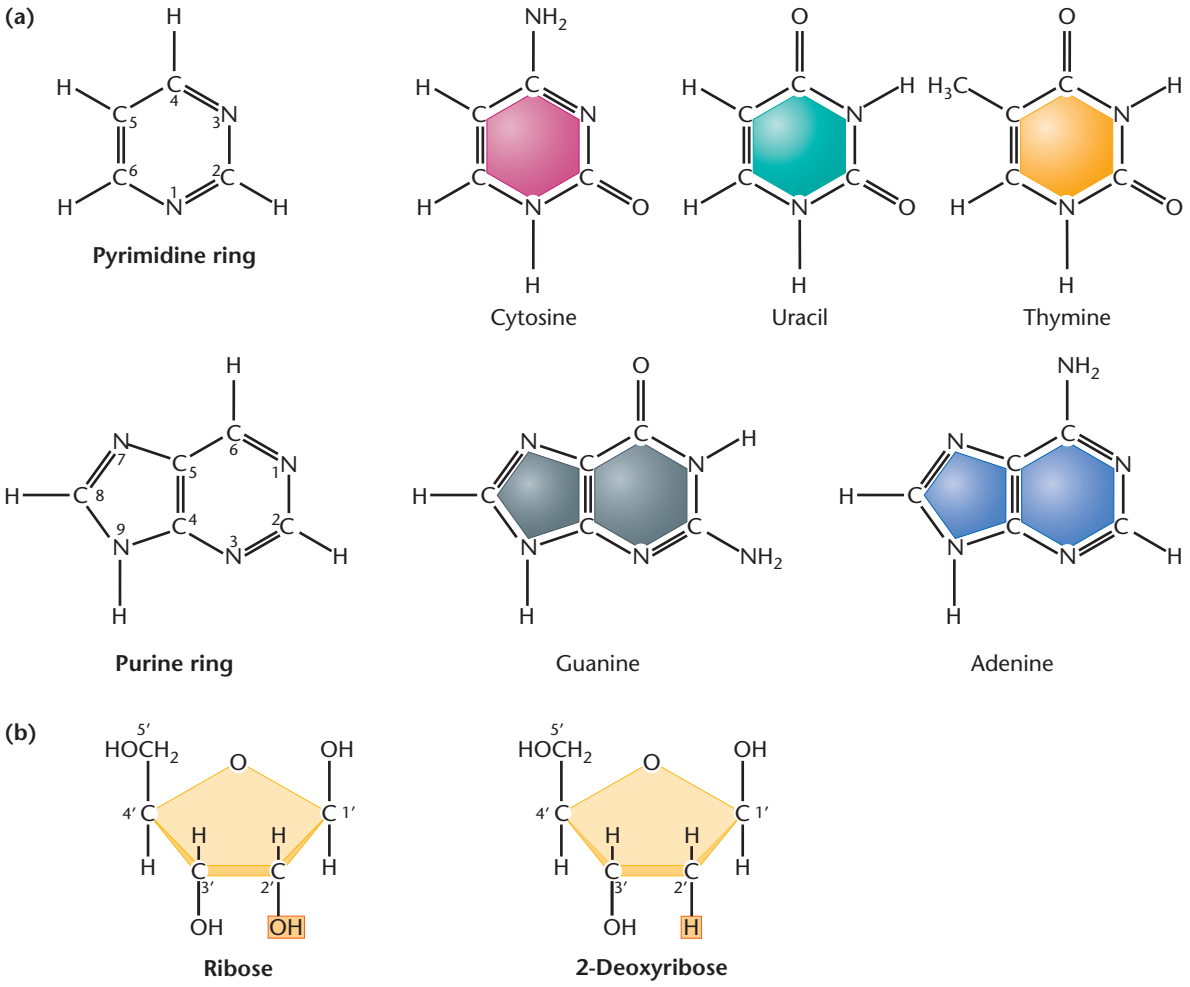


FIGURE 9–6 (a) Chemical structures of the pyrimidines and purines that serve as the nitrogenous bases in RNA and DNA. (b) Chemical ring structures of ribose and 2-deoxyribose, which serve as the pentose sugars in RNA and DNA, respectively.

Nucleic Acid Chemistry

Before turning to this work, a brief introduction to nucleic acid chemistry is in order. This chemical information was well known to Watson and Crick during their investigation and served as the basis of their model building.

DNA is a nucleic acid, and nucleotides are the building blocks of all nucleic acid molecules. Sometimes called mononucleotides, these structural units have three essential components: a **nitrogenous base,** a **pentose sugar** (a five-carbon sugar), and a **phosphate group.** There are two kinds of nitrogenous bases: the nine-member double-ring **purines** and the six-member single-ring **pyrimidines.**

Two types of purines and three types of pyrimidines are found in nucleic acids. The two purines are **adenine** and **guanine,** abbreviated A and G, respectively. The three pyrimidines are **cytosine, thymine,** and **uracil** (respectively, C, T, and U). The chemical structures of the five bases are shown in **Figure 9–6(a)**. Both DNA and RNA contain A, G, and C, but only DNA contains the base T and only RNA contains the base U. Each nitrogen or carbon atom of the ring structures of purines and pyrimidines is designated by a number. Note that corresponding atoms in the purine and pyrimidine rings are numbered differently.

The pentose sugars found in nucleic acids give them their names. **Ribonucleic acids (RNA)** contain **ribose,** while **deoxyribonucleic acids (DNA)** contain **deoxyribose.** **Figure 9–6(b)** shows the chemical structures for these two pentose sugars. Each carbon atom is distinguished by a number with a prime sign (′). As you can see in Figure 9–6(b), compared with ribose, deoxyribose has a hydrogen atom rather than a hydroxyl group at the (C-2′) position. The absence of a hydroxyl group at the (C-2′) position thus distinguishes DNA from RNA. In the absence of the (C-2′) hydroxyl group, the sugar is more specifically named **2-deoxyribose.**

If a molecule is composed of a purine or pyrimidine base and a ribose or deoxyribose sugar, the chemical unit is called a **nucleoside.** If a phosphate group is added to the nucleoside, the molecule is now called a **nucleotide.** Nucleosides and nucleotides are named according to the specific nitrogenous base (A, G, C, T, or U) that is part of the

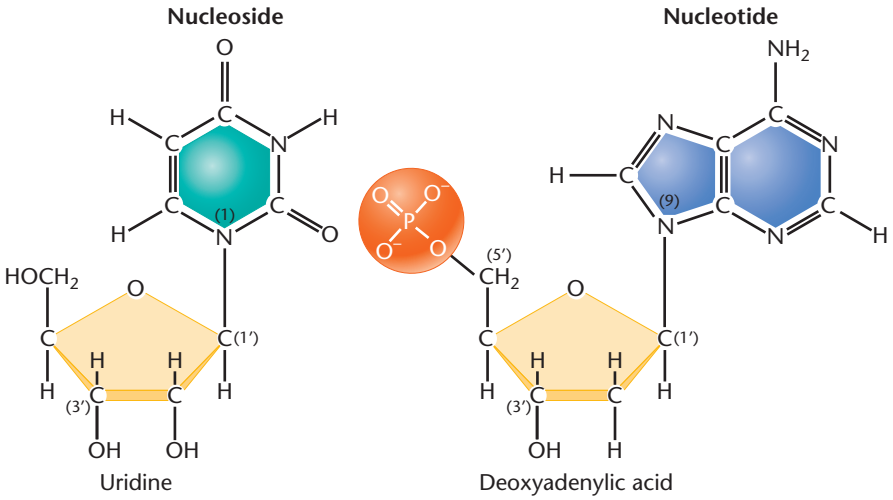


FIGURE 9–7 Structures and names of the nucleosides and nucleotides of RNA and DNA.

Ribonucleosides	Ribonucleotides
Adenosine	Adenylic acid
Cytidine	Cytidylic acid
Guanosine	Guanylic acid
Uridine	Uridylic acid
Deoxyribonucleosides	**Deoxyribonucleotides**
Deoxyadenosine	Deoxyadenylic acid
Deoxycytidine	Deoxycytidylic acid
Deoxyguanosine	Deoxyguanylic acid
Deoxythymidine	Deoxythymidylic acid

molecule. The structure of a nucleotide and the nomenclature used in naming DNA nucleotides and nucleosides are shown in **Figure 9–7**.

The bonding between components of a nucleotide is highly specific. The (C-1′) atom of the sugar is involved in the chemical linkage to the nitrogenous base. If the base is a purine, the N-9 atom is covalently bonded to the sugar; if the base is a pyrimidine, the N-1 atom bonds to the sugar. In deoxyribonucleotides, the phosphate group may be bonded to the (C-2′), (C-3′), or (C-5′) atom of the sugar. The (C-5′)-phosphate configuration is shown in Figure 9–7. It is by far the most prevalent one in biological systems and the one found in DNA and RNA.

Nucleotides are also described by the term **nucleoside monophosphate (NMP).** The addition of one or two phosphate groups results in **nucleoside diphosphates (NDP)** and **triphosphates (NTP),** respectively, as shown in **Figure 9–8**. The triphosphate form is significant because it

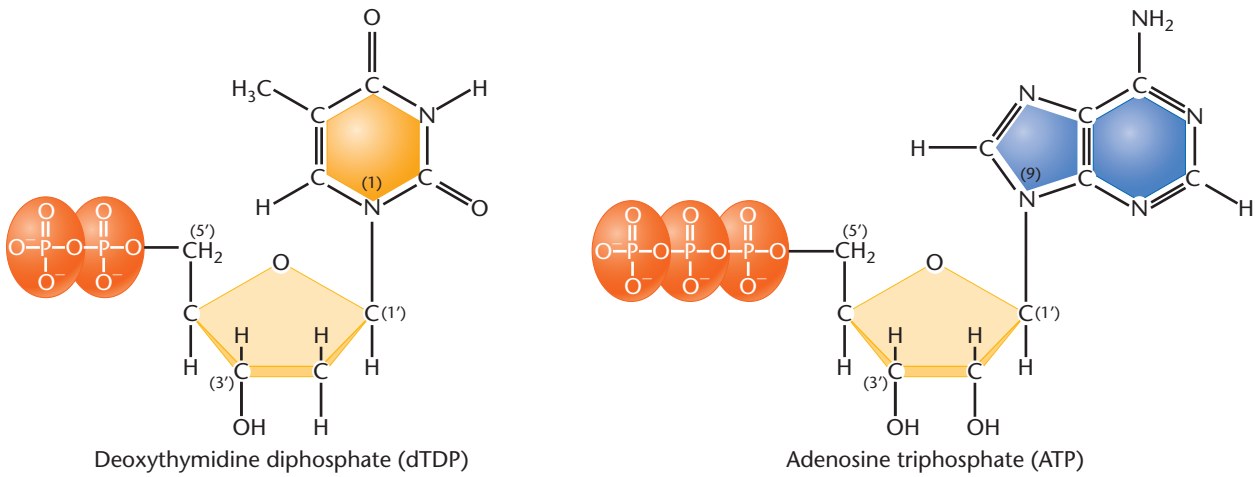


FIGURE 9–8 Basic structures of nucleoside diphosphates and triphosphates. Deoxythymidine diphosphate and adenosine triphosphate are diagrammed here.

is the precursor molecule during nucleic acid synthesis within the cell. In addition, **adenosine triphosphate (ATP)** and **guanosine triphosphate (GTP)** are important in cell bioenergetics because of the large amount of energy involved in adding or removing the terminal phosphate group. The hydrolysis of ATP or GTP to ADP or GDP and inorganic phosphate (P_i) is accompanied by the release of a large amount of energy in the cell. When these chemical conversions are coupled to other reactions, the energy produced is used to drive them. As a result, ATP and GTP are involved in many cellular activities.

The linkage between two mononucleotides involves a phosphate group linked to two sugars. A **phosphodiester bond** is formed as phosphoric acid is joined to two alcohols (the hydroxyl groups on the two sugars) by an ester linkage on both sides. Figure 9–9 shows the resultant phosphodiester bond in DNA. Each structure has a (C-3′) end and a (C-5′) end. Two joined nucleotides form a dinucleotide; three nucleotides, a trinucleotide; and so forth. Short chains consisting of up to 20 nucleotides or so are called **oligonucleotides;** longer chains are **polynucleotides.**

Long polynucleotide chains account for the large molecular weight of DNA and explain its most important property—storage of vast quantities of genetic information.

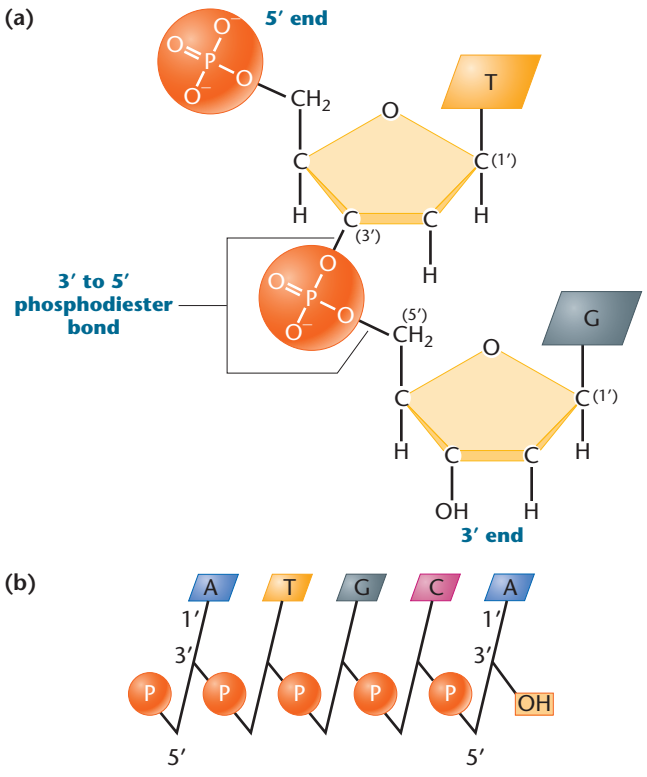


FIGURE 9–9 (a) Linkage of two nucleotides by the formation of a C-3′ to C-5′ (3′–5′) phosphodiester bond, producing a dinucleotide. (b) Shorthand notation for a polynucleotide chain.

If each nucleotide position in this long chain can be occupied by any one of four nucleotides, extraordinary variation is possible. For example, a polynucleotide only 1000 nucleotides in length can be arranged 4^{1000} different ways, each one different from all other possible sequences. This potential variation in molecular structure is essential if DNA is to store the vast amounts of chemical information necessary to direct cellular activities.

Base-Composition Studies

Between 1949 and 1953, Erwin Chargaff and his colleagues used chromatographic methods to separate the four bases in DNA samples from various organisms. Quantitative methods were then used to determine the amounts of the four nitrogenous bases from each source. On the basis of these data, the following conclusions may be drawn:

1. The amount of adenine residues is proportional to the amount of thymine residues in DNA. Also, the amount of guanine residues is proportional to the amount of cytosine residues.
2. Based on this proportionality, the sum of the purines (A + G) equals the sum of the pyrimidines (C + T).
3. The percentage of (G + C) does not necessarily equal the percentage of (A + T). Instead, this ratio varies greatly between different organisms.

These conclusions indicate a definite pattern of base composition of DNA molecules. The data were critical to Watson and Crick's successful model of DNA. They also directly refuted Levene's tetranucleotide hypothesis, which stated that all four bases are present in equal amounts.

X-Ray Diffraction Analysis

When fibers of a DNA molecule are subjected to X-ray bombardment, the X rays scatter according to the molecule's atomic structure. The pattern of scatter can be captured as spots on photographic film and analyzed, particularly for the overall shape of and regularities within the molecule. This process, **X-ray diffraction analysis,** was applied successfully to the study of protein structure by Linus Pauling and other chemists. The technique had been attempted on DNA as early as 1938 by William Astbury. By 1947, he had detected a periodicity of 3.4 angstroms (Å)* within the structure of the molecule, which suggested to him that the bases were stacked like coins on top of one another.

* Today, measurement in nanometers (nm) is favored (1 nm = 10 Å).

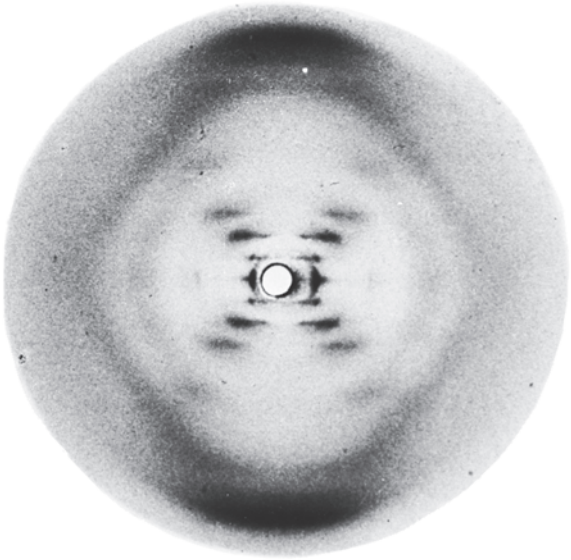


FIGURE 9–10 X-ray diffraction photograph of purified DNA fibers. The strong arcs on the periphery show closely spaced aspects of the molecule, providing an estimate of the periodicity of nitrogenous bases, which are 3.4 Å apart. The inner cross pattern of spots shows the grosser aspect of the molecule, indicating its helical nature.

Between 1950 and 1953, Rosalind Franklin, working in the laboratory of Maurice Wilkins, obtained improved X-ray data from more purified samples of DNA (**Figure 9–10**). Her work confirmed the 3.4 Å periodicity seen by Astbury and suggested that the structure of DNA was some sort of helix. However, she did not propose a definitive model. Pauling had analyzed the work of Astbury and others and proposed incorrectly that DNA is a triple helix.

The Watson–Crick Model

Watson and Crick published their analysis of DNA structure in 1953. By building models under the constraints of the information just discussed, they proposed the double-helical form of DNA shown in **Figure 9–11(a)**. This model has the following major features:

1. Two long polynucleotide chains are coiled around a central axis, forming a right-handed double helix.
2. The two chains are antiparallel; that is, their (C-5′) to (C-3′) orientations run in opposite directions.
3. The bases of both chains are flat structures, lying perpendicular to the axis; they are "stacked" on one another, 3.4 Å (0.34 nm) apart, and located on the inside of the structure.
4. The nitrogenous bases of opposite chains are *paired* as the result of hydrogen bonds; in DNA, only A═T and G≡C pairs occur.
5. Each complete turn of the helix is 34 Å (3.4 nm) long; thus, 10 bases exist per turn in each chain.
6. In any segment of the molecule, alternating larger **major grooves** and smaller **minor grooves** are apparent along the axis.
7. The double helix measures 20 Å (2.0 nm) in diameter.

The nature of *base pairing* (point 4 above) is the most genetically significant feature of the model. Before we

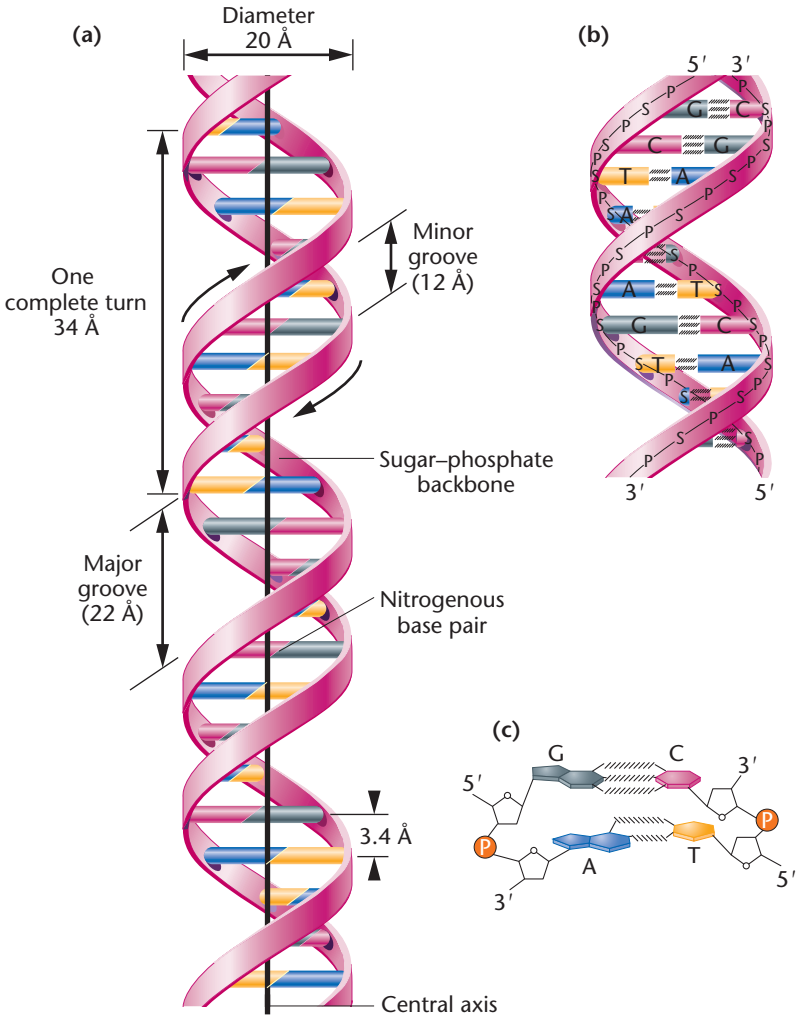


FIGURE 9–11 (a) The DNA double helix as proposed by Watson and Crick. The ribbonlike strands constitute the sugar–phosphate backbones, and the horizontal rungs constitute the nitrogenous base pairs, of which there are 10 per complete turn. The major and minor grooves are shown. The solid vertical bar represents the central axis. (b) A detailed view labeled with the bases, sugars, phosphates, and hydrogen bonds of the helix. (c) A demonstration of the antiparallel nature of the helix and the horizontal stacking of the bases.

discuss it in detail, several other important features warrant emphasis. First, the antiparallel nature of the two chains is a key part of the double helix model. While one chain runs in the 5′-to-3′ orientation (what seems right side up to us), the other chain goes in the 3′-to-5′ orientation (and thus appears upside down). This is illustrated in **Figure 9–11(b)** and **(c)**. Given the constraints of the bond angles of the various nucleotide components, the double helix could not be constructed easily if both chains ran parallel to one another.

The key to the model proposed by Watson and Crick is the specificity of base pairing. Chargaff's data suggested that the amounts of A equaled T and that the amounts of G equaled C. Watson and Crick realized that if A pairs with T and C pairs with G, this would account for these proportions and that such pairing could occur as a result of hydrogen bonding between base pairs [Figure 9–11(c)], providing the chemical stability necessary to hold the two chains together. Arranged in this way, both major and minor grooves become apparent along the axis. Further, a purine (A or G) opposite a pyrimidine (T or C) on each "rung of the spiral staircase" of the proposed double helix accounts for the 20 Å (2 nm) diameter suggested by X-ray diffraction studies.

The specific A═T and G≡C base pairing is the basis for **complementarity.** This term describes the chemical affinity provided by hydrogen bonding between the bases. As we shall see, complementarity is very important in DNA replication and gene expression.

It is appropriate to inquire into the nature of a hydrogen bond and to ask whether it is strong enough to stabilize the helix. A **hydrogen bond** is a very weak electrostatic attraction between a covalently bonded hydrogen atom and an atom with an unshared electron pair. The hydrogen atom assumes a partial positive charge, while the unshared electron pair—characteristic of covalently bonded oxygen and nitrogen atoms—assumes a partial negative charge. These opposite charges are responsible for the weak chemical attractions. As oriented in the double helix, adenine forms two hydrogen bonds with thymine, and guanine forms three hydrogen bonds with cytosine. Although two or three individual hydrogen bonds are energetically very weak, 2000 to 3000 bonds in tandem (typical of two long polynucleotide chains) provide great stability to the helix.

Another stabilizing factor is the arrangement of sugars and bases along the axis. In the Watson–Crick model, the *hydrophobic* ("water-fearing") nitrogenous bases are stacked almost horizontally on the interior of the axis and are thus shielded from water. The *hydrophilic* ("water-loving") sugar–phosphate backbone is on the outside of the axis, where both components can interact with water. These molecular arrangements provide significant chemical stabilization to the helix.

A more recent and accurate analysis of the form of DNA that served as the basis for the Watson–Crick model has revealed a minor structural difference between the substance and the model. A precise measurement of the number of base pairs per turn has demonstrated a value of 10.4, rather than the 10.0 predicted by Watson and Crick. In the classic model, each base pair is rotated 36° around the helical axis relative to the adjacent base pair, but the new finding requires a rotation of 34.6°. This results in slightly more than 10 base pairs per 360° turn.

The Watson–Crick model had an immediate effect on the emerging discipline of molecular biology. Even in their initial 1953 article, the authors noted, "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." Two months later, Watson and Crick pursued this idea in a second article in *Nature,* suggesting a specific mechanism of replication of DNA—the **semiconservative mode of replication.** The second article alluded to two new concepts: (1) the storage of genetic information in the sequence of the bases, and (2) the mutations or genetic changes that would result from an alteration of the bases. These ideas have received vast amounts of experimental support since 1953 and are now universally accepted.

Watson and Crick's synthesis of ideas was highly significant with regard to subsequent studies of genetics and biology. The nature of the gene and its role in genetic mechanisms could now be viewed and studied in biochemical terms. Recognition of their work, along with that of Wilkins, led to their receipt of the Nobel Prize in Physiology or Medicine in 1962. Unfortunately, Rosalind Franklin had died in 1958 at the age of 37, making her contributions ineligible for consideration since the award is not given posthumously. The Nobel Prize was to be one of many such awards bestowed for work in the field of molecular genetics.

ESSENTIAL POINT

As proposed by Watson and Crick, DNA exists in the form of a right-handed double helix composed of two long antiparallel polynucleotide chains held together by hydrogen bonds formed between complementary, nitrogenous base pairs. ■

NOW SOLVE THIS

9–2 In sea urchin DNA, which is double-stranded, 17.5 percent of the bases were shown to be cytosine (C). What percentages of the other three bases are expected to be present in this DNA?

■ **HINT:** *This problem asks you to extrapolate from one measurement involving a unique DNA molecule to three other values characterizing the molecule. The key to its solution is to understand the base-pairing rules in the Watson–Crick model of DNA.*

EVOLVING CONCEPT OF THE GENE

Based on the model of DNA put forward by Watson and Crick in 1953, the gene was viewed for the first time in molecular terms as a sequence of nucleotides in a DNA helix that encodes genetic information. ■

9.7 Alternative Forms of DNA Exist

Under different conditions of isolation, several conformational forms of DNA have been recognized. At the time Watson and Crick performed their analysis, two forms—**A-DNA** and **B-DNA**—were known. Watson and Crick's analysis was based on X-ray studies of B-DNA performed by Franklin, which is present under aqueous, low-salt conditions and is believed to be the biologically significant conformation.

While DNA studies around 1950 relied on the use of X-ray diffraction, more recent investigations have been performed using **single-crystal X-ray analysis.** The earlier studies achieved limited resolution of about 5 Å, but single crystals diffract X rays at about 1 Å, near atomic resolution. As a result, every atom is "visible" and much greater structural detail is available during analysis.

With this modern technique, A-DNA, which is prevalent under high-salt or dehydration conditions, has now been scrutinized. In comparison to B-DNA A-DNA is slightly more compact, with 11 bp in each complete turn of the helix, which is 23 Å (2.3 nm) in diameter. It is also a right-handed helix, but the orientation of the bases is somewhat different—they are tilted and displaced laterally in relation to the axis of the helix. As a result, the appearance of the major and minor grooves is modified. It seems doubtful that A-DNA occurs *in vivo* (under physiological conditions).

Other forms of DNA (e.g., C-, D-, E-, and most recently, P-DNA) are now known, but it is **Z-DNA** that has drawn the most attention. Discovered by Andrew Wang, Alexander Rich, and their colleagues in 1979 when they examined a small synthetic DNA fragment containing only G≡C base pairs, Z-DNA takes on the rather remarkable configuration of a *left-handed double helix*. Like A- and B-DNA, Z-DNA consists of two antiparallel chains held together by Watson–Crick base pairs. Beyond these characteristics, Z-DNA is quite different. The left-handed helix is 18 Å (1.8 nm) in diameter, contains 12 bp per turn, and assumes a zigzag conformation (hence its name). The major groove present in B-DNA is nearly eliminated. Z-DNA is compared with the A and B forms in the chapter opening photograph on p. 176.

Speculation abounds over the possibility that regions of Z-DNA exist in the chromosomes of living organisms. The unique helical arrangement could provide an important recognition point for the interaction with other molecules. However, it is still not clear whether Z-DNA occurs *in vivo*.

Still other forms of DNA have been studied, including P-DNA, named after Linus Pauling. It is produced by artificial "stretching" of DNA, creating a longer, narrower version with the phosphate groups on the interior.

9.8 The Structure of RNA Is Chemically Similar to DNA, but Single-Stranded

The structure of RNA molecules resembles DNA, with several important exceptions. Although RNA also has nucleotides linked with polynucleotide chains, the sugar ribose replaces deoxyribose, and the nitrogenous base uracil replaces thymine. Another important difference is that most RNA is single-stranded, although there are two important exceptions. First, RNA molecules sometimes fold back on themselves to form double-stranded regions of complementary base pairs. Second, some animal viruses that have RNA as their genetic material contain double-stranded helices.

As established earlier (see Figure 9–1), three major classes of cellular RNA molecules function during the expression of genetic information: **ribosomal RNA (rRNA), messenger RNA (mRNA),** and **transfer RNA (tRNA).** These molecules all originate as complementary copies of deoxyribonucleotide sequences of DNA. Because uracil replaces thymine in RNA, uracil is complementary to adenine during transcription and RNA base pairing.

Different RNAs are distinguished by their sedimentation behavior in a centrifugal field. Sedimentation behavior depends on a molecule's density, mass, and shape, and its measure is called the **Svedberg coefficient** (*S*). Although higher *S* values almost always designate molecules of greater molecular weight, the correlation is not direct; that is, a two-fold increase in molecular weight does not lead to a two-fold increase in *S*.

Ribosomal RNAs are generally the largest of these molecules and usually constitute about 80 percent of all RNA in the cell. Ribosomal RNAs are important structural components of **ribosomes,** which function as nonspecific workbenches where proteins are synthesized during translation. The various forms of rRNA found in prokaryotes and eukaryotes differ distinctly in size.

Messenger RNA molecules carry genetic information from the DNA of the gene to the ribosome. The mRNA molecules vary considerably in size, which reflects the

variation in the size of the protein encoded by the mRNA as well as the gene serving as the template for transcription of mRNA.

Transfer RNA, the smallest class of these RNA molecules, carries amino acids to the ribosome during translation. Since more than one tRNA molecule interacts simultaneously with the ribosome, the molecule's smaller size facilitates these interactions.

Other unique RNAs exist that perform various genetic roles, especially in eukaryotes. For example, **telomerase RNA** is involved in DNA replication at the ends of chromosomes (the telomeres). **Small nuclear RNA (snRNA)** participates in processing mRNAs, and **antisense RNA, microRNA (miRNA),** and **short interfering RNA (siRNA)** are involved in gene regulation.

ESSENTIAL POINT

The second category of nucleic acids important in genetic function is RNA, which is similar to DNA with the exceptions that it is usually single-stranded, the sugar ribose replaces the deoxyribose, and the pyrimidine uracil replaces thymine. ■

NOW SOLVE THIS

9–3 German measles results from an infection of the rubella virus, which can cause a multitude of health problems in newborns. What conclusions can you reach from a nucleic acid analysis of the virus that reveals an A + G/U + C ratio of 1.13?

■ **HINT:** *This problem asks you to analyze information about the chemical composition of a nucleic acid serving as the genetic material of a virus. The key to its solution is to apply your knowledge of nucleic acid chemistry, in particular your understanding of base pairing.*

9.9 Many Analytical Techniques Have Been Useful during the Investigation of DNA and RNA

Since 1953, the role of DNA as the genetic material and the role of RNA in transcription and translation have been clarified through detailed analysis of nucleic acids. Several important methods of analysis are based on the unique nature of the hydrogen bond that is so integral to the structure of nucleic acids. For example, if DNA is subjected to heat, the double helix is denatured and unwinds. During unwinding, the viscosity of DNA decreases and UV absorption increases (called the **hyperchromic shift**). A melting profile, in which OD_{260} is plotted against temperature, is shown for two DNA molecules in **Figure 9–12**.

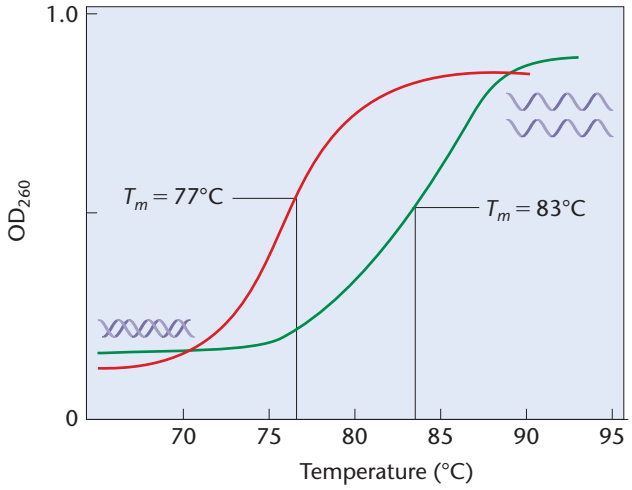


FIGURE 9–12 A melting profile shows the increase in UV absorption versus temperature (the hyperchromic effect) for two DNA molecules with different G≡C contents. The molecule with a melting point (T_m) of 83°C has a greater G≡C content than the molecule with a T_m of 77°C.

The midpoint of each curve is called the **melting temperature** T_m where 50 percent of the strands have unwound. The molecule with a higher T_m has a higher percentage of G≡C base pairs than A═T base pairs since G≡C pairs share three hydrogen bonds compared to the two bonds between A═T pairs.

The denaturation/renaturation of nucleic acids is the basis for one of the most useful techniques in molecular genetics—**molecular hybridization.** Provided that a reasonable degree of base complementarity exists between any two nucleic acid strands, denaturation can be reversed whereby molecular hybridization is possible. Duplexes can be re-formed between DNA strands, even from different organisms, and between DNA and RNA strands. For example, an RNA molecule will hybridize with the segment of DNA from which it was transcribed. As a result, nucleic acid **probes** are often used to identify complementary sequences.

The technique can even be performed using the DNA present in chromosomal preparations as the "target" for hybrid formation. This process is called ***in situ* molecular hybridization.** Mitotic cells are first fixed to slides and then subjected to hybridization conditions. Single-stranded DNA or RNA is added (a probe), and hybridization is monitored. The nucleic acid that is added may be either radioactive or contain a fluorescent label to allow its detection. In the former case, autoradiography is used.

Figure 9–13 illustrates the use of a fluorescent label. A short fragment of DNA that is complementary to DNA in the chromosomes' centromere regions has been hybridized. Fluorescence occurs only in the centromere regions

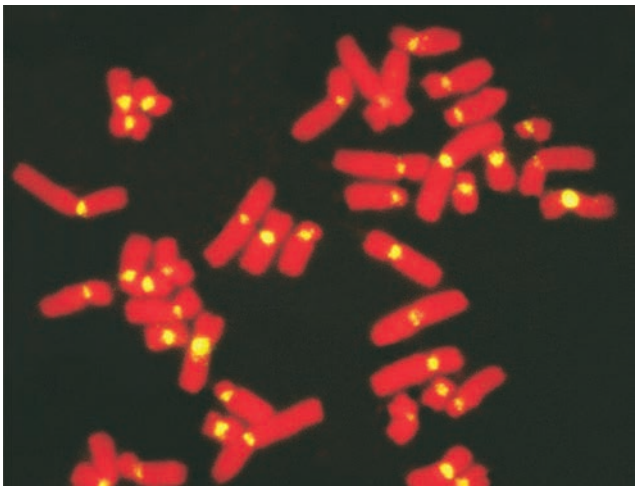


FIGURE 9–13 Fluorescent *in situ* hybridization (FISH) of human metaphase chromosomes. The probe, specific to centromeric DNA, produces a yellow fluorescence signal indicating hybridization. The red fluorescence is produced by propidium iodide counterstaining of chromosomal DNA.

and thus identifies each one along its chromosome. Because fluorescence is used, the technique is known by the acronym **FISH (fluorescent *in situ* hybridization).** The use of this technique to identify chromosomal locations housing specific genetic information has been a valuable addition to geneticists' repertoire of experimental techniques.

Electrophoresis

Another technique essential to the analysis of nucleic acids is **electrophoresis.** This technique may be adapted to separate different-sized fragments of DNA and RNA chains and is invaluable in current research investigations in molecular genetics.

Electrophoresis separates the molecules in a mixture by causing them to migrate under the influence of an electric field. A sample is placed on a porous substance, such as a semisolid gel, which is then placed in a solution that conducts electricity. Mixtures of molecules with a similar charge–mass ratio but of different sizes will migrate at different rates through the gel based on their size. For example, two polynucleotide chains of different lengths, such as 10 versus 20 nucleotides, are both negatively charged (based on the phosphate groups of the nucleotides) and will both move to the positively charged pole (the anode), but at different rates. Using a medium such as a **polyacrylamide gel** or an **agarose gel,** which can be prepared with various pore sizes, the *shorter chains migrate at a faster rate through the gel than larger chains* (**Figure 9–14**). Once electrophoresis is complete, bands representing the variously sized molecules are identified either by autoradiography (if a component of the molecule is radioactive) or by use of a fluorescent dye that binds to nucleic acids. The resolving power is so great that polynucleotides that vary by just one nucleotide in length may be separated.

Electrophoretic separation of nucleic acids is at the heart of a variety of other commonly used research techniques. Of particular note are the various "blotting" techniques (e.g., Southern blots and Northern blots), as well as DNA sequencing methods, which we will discuss in detail later in Chapter 17.

ESSENTIAL POINT

Various methods of analysis of nucleic acids, particularly molecular hybridization and electrophoresis, have led to studies essential to our understanding of genetic mechanisms. ■

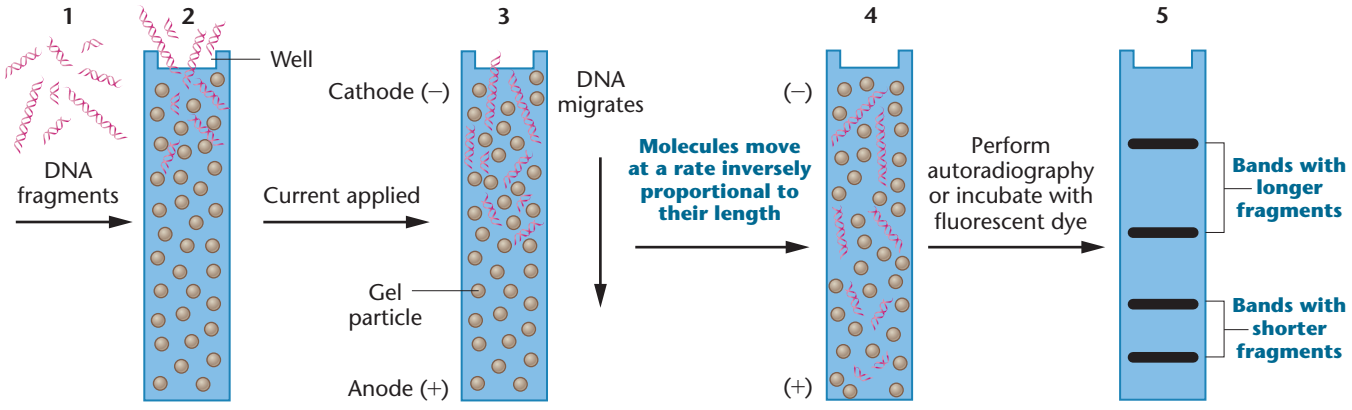
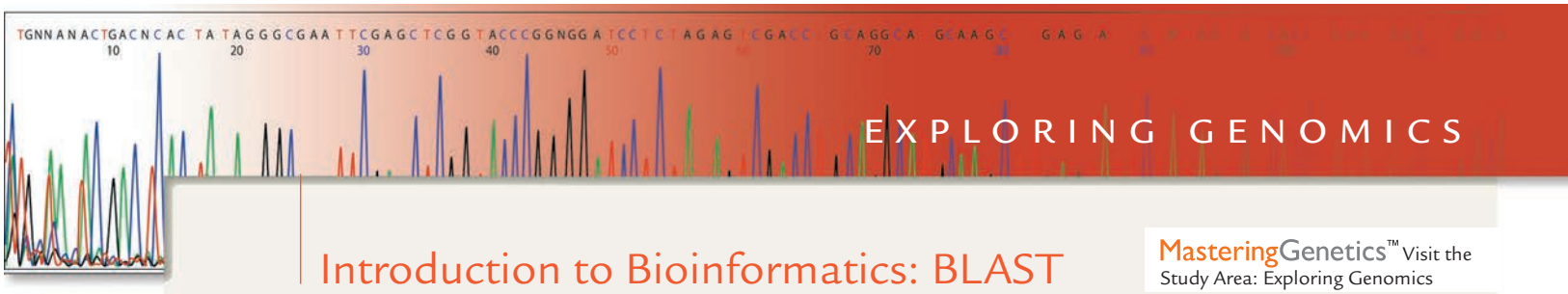


FIGURE 9–14 Electrophoretic separation of a mixture of DNA fragments that vary in length. The photograph at the bottom right shows an autoradiogram derived from an agarose gel that reveals DNA bands.



Introduction to Bioinformatics: BLAST

MasteringGenetics™ Visit the Study Area: Exploring Genomics

In this chapter, we focused on the structural details of DNA. In Chapter 17, you will learn how scientists can clone and sequence DNA. The explosion of DNA and protein sequence data that has occurred in the last 15 years has launched the field of *bioinformatics,* an interdisciplinary science that applies mathematics and computing technology to develop hardware and software for storing, sharing, comparing, and analyzing nucleic acid and protein sequence data.

A large number of sequence databases that make use of bioinformatics have been developed. An example is **GenBank** (http://www.ncbi.nlm.nih.gov/genbank/), which is the National Institutes of Health sequence database. This global resource, with access to databases in Europe and Japan, currently contains more than 152 billion base pairs of sequence data!

In the Exploring Genomics exercises for Chapter 7, you were introduced to the National Center for Biotechnology Information (NCBI) Genes and Disease site. Now we will use an NCBI application called **BLAST, Basic Local Alignment Search Tool**. BLAST is an invaluable program for searching through GenBank and other databases to find DNA- and protein-sequence similarities.

Exercise I – Introduction to BLAST

1. Access BLAST from the NCBI Web site at http://blast.ncbi.nlm.nih.gov/Blast.cgi.
2. Click on "nucleotide blast." This feature allows you to search DNA databases to look for a similarity between a sequence you enter and other sequences in the database. You will do a nucleotide search with the following sequence:

   CCAGAGTCCAGCTGCTGCTCATA
   CTACTGATACTGCTGGG
3. Imagine that this sequence is a short part of a gene you cloned in your laboratory. You want to know if this gene or others with similar sequences have been discovered. Enter this sequence into the "Enter Query Sequence" text box at the top of the page. Near the bottom of the page, under the "Program Selection" category, choose "blastn"; then click on the "BLAST" button at the bottom of the page to run the search. It may take several minutes for results to be available because BLAST is using powerful algorithms to scroll through billions of bases of sequence data! A new page will appear with the results of your search.
4. On the search results page, below the Graphic Summary you will see a category called Descriptions and a table showing significant matches to the sequence you searched with (called the query sequence). BLAST determines significant matches based on statistical measures that consider the length of the query sequence, the number of matches with sequences in the database, and other factors. Significant *alignments,* regions of significant similarity in the query and subject sequences, typically have E values less than 1.0.
5. The top part of the table lists matches to transcripts (mRNA sequences), and the lower part lists matches to genomic DNA sequences, in order of highest to lowest number of matches.
6. Alignments are indicated by horizontal lines. BLAST adjusts for gaps in the sequences, that is, for areas that may not align precisely because of missing bases in otherwise similar sequences. Scroll below the table to see the aligned sequences from this search, and then answer the following questions:

   a. What were the top three matches to your query sequence?

   b. For each alignment, BLAST also indicates the percent *identity* and the number of gaps in the match between the query and subject sequences. What was the percent identity for the top three matches? What percentage of each aligned sequence showed gaps indicating sequence differences?

   c. Click on the links for the first matched sequence (far-right column). These will take you to a wealth of information, including the size of the sequence; the species it was derived from; a PubMed-linked chronology of research publications pertaining to this sequence; the complete sequence; and if the sequence encodes a polypeptide, the predicted amino acid sequence coded by the gene. Skim through the information presented for this gene. What is the gene's function?
7. A BLAST search can also be done by entering the *accession number* for a sequence, which is a unique identifying number assigned to a sequence before it can be put into a database. For example, search with the accession number NM_007305. What did you find?

CASE STUDY | Zigs and zags of the smallpox virus

Smallpox, a once highly lethal contagious disease, has been eradicated worldwide. However, research continues with stored samples of variola, the smallpox virus, because it is a potential weapon in bioterrorism. Human cells protect themselves from the variola virus (and other viruses) by activating genes that encode protective proteins. It has recently been discovered that in response to variola, human cells create small transitory stretches of Z-DNA at sites that regulate these genes. The smallpox virus can bypass this cellular defense mechanism by specifically targeting the segments of Z-DNA and inhibiting the synthesis of the protective proteins. This discovery raises some interesting questions:

1. What is unique about Z-DNA that might make it a specific target during viral infection?
2. How might the virus target host-cell Z-DNA formation to block the synthesis of antiviral proteins?
3. To study the interaction between viral proteins and Z-DNA, how could Z-DNA-forming DNA be synthesized in the lab?
4. How could this research lead to the development of drugs to combat infection by variola and related viruses?

INSIGHTS AND SOLUTIONS

This chapter recounts some of the initial experimental analyses that launched the era of molecular genetics. Quite fittingly, then, our "Insights and Solutions" section shifts its emphasis from problem solving to experimental rationale and analytical thinking.

1. Based strictly on the transformation analysis of Avery, MacLeod, and McCarty, what objection might be made to the conclusion that DNA is the genetic material? What other conclusion might be considered?

   Solution: Based solely on their results, we could conclude that DNA is essential for transformation. However, DNA might have been a substance that caused capsular formation by converting nonencapsulated cells *directly* to cells with a capsule. That is, DNA may simply have played a catalytic role in capsular synthesis, leading to cells that display smooth, type III colonies.

2. What observations argue against this objection?

   Solution: First, transformed cells pass the trait on to their progeny cells, thus supporting the conclusion that DNA is responsible for heredity, not for the direct production of polysaccharide coats. Second, subsequent transformation studies over the next five years showed that other traits, such as antibiotic resistance, could be transformed. Therefore, the transforming factor has a broad general effect, not one specific to polysaccharide synthesis.

3. If RNA were the universal genetic material, how would this have affected the Avery experiment and the Hershey–Chase experiment?

   Solution: In the Avery experiment, ribonuclease (RNase), rather than deoxyribonuclease (DNase), would have eliminated transformation. Had this occurred, Avery and his colleagues would have concluded that RNA was the transforming factor. Hershey and Chase would have obtained identical results, since ^{32}P would also label RNA but not protein.

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we have focused on DNA, the molecule that stores genetic information in all living things. In particular, we discussed its structure and delved into how we analyze this molecule. Based on your knowledge of these topics, answer several fundamental questions:
   (a) How were we able to determine that DNA, and not some other molecule, serves as the genetic material in bacteria, bacteriophages, and eukaryotes?
   (b) How do we know that the structure of DNA is in the form of a right-handed double-helical molecule?
   (c) How do we know that in DNA G pairs with C and that A pairs with T as complementary strands are formed?
   (d) How do we know that repetitive DNA sequences exist in eukaryotes?

CONCEPTS QUESTION

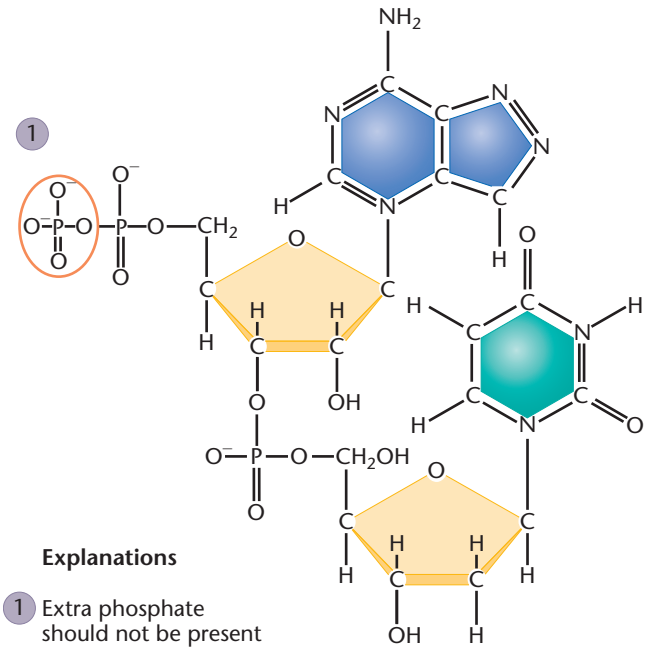
2. Review the Chapter Concepts list on p. 176. Most center on DNA and RNA and their role of serving as the genetic material. Write a short essay that contrasts these molecules, including a comparison of advantages conferred by their structure that each of them has over the other in serving in this role. ■
3. Discuss the reasons why proteins were generally favored over DNA as the genetic material before 1940. What was the role of the tetranucleotide hypothesis in this controversy?
4. Contrast the various contributions made to our understanding of transformation by Griffith, Alloway, and Avery.
5. When Avery and his colleagues had obtained what was concluded to be the transforming factor from the IIIS virulent cells, they treated the fraction with proteases, ribonuclease, and deoxyribonuclease, followed by the assay for retention or loss of transforming ability. What were the purpose and results of these experiments? What conclusions were drawn?
6. Why were ^{32}P and ^{35}S chosen in the Hershey–Chase experiment? Discuss the rationale and conclusions of this experiment.
7. Present an overview of two classical experiments that demonstrated that DNA is the genetic material. Can RNA be the genetic material? Explain.
8. What observations are consistent with the conclusion that DNA serves as the genetic material in eukaryotes? List and discuss them.
9. What are the exceptions to the general rule that DNA is the genetic material in all organisms? What evidence supports these exceptions?
10. Draw the chemical structure of the three components of a nucleotide, and then link them together. What atoms are removed from the structures when the linkages are formed?

11. What are the structural differences between (a) purines and pyrimidines, and (b) ribose and deoxyribose sugars?
12. Cytosine may also be named 2-oxy-4-amino pyrimidine. How would you name the other four nitrogenous bases, using this alternative system? (CH_3 is methyl.)
13. Draw the chemical structure of a dinucleotide composed of A and G. Opposite this structure, draw the dinucleotide composed of T and C in an antiparallel (or upside-down) fashion. Form the possible hydrogen bonds.
14. Describe the various characteristics of the Watson–Crick double helix model for DNA.
15. What evidence did Watson and Crick have at their disposal in 1953? What was their approach in arriving at the structure of DNA?
16. What might Watson and Crick have concluded, had Chargaff's data from a single source indicated the following base composition?

	A	T	G	C
%	29	19	21	31

Why would this conclusion be contradictory to Wilkins and Franklin's data?

17. If the GC content of a DNA molecule is 60%, what are the molar percentages of the four bases (G, C, T, A)?
18. If an RNA strand generates its complementary strand, thus producing a double helix, will a molecule of this be structurally identical to that of DNA? Explain.
19. What are the three major types of RNA molecules? How is each related to the concept of information flow?
20. What component of the nucleotide is responsible for the absorption of ultraviolet light? How is this technique important in the analysis of nucleic acids?
21. What is the physical state of DNA after being denatured by heat?
22. What is the hyperchromic effect? How is it measured? What does T_m imply?
23. A certain DNA molecule has 40% A + T, while another has 30% G + C. Which of these will have a higher T_m and why?
24. What is the chemical basis of molecular hybridization?
25. What is meant by the semiconservative nature of DNA replication?
26. A genetics student was asked to draw the chemical structure of an adenine- and thymine-containing dinucleotide derived from DNA. His answer is shown below. The student made more than six major errors. One of them is circled, numbered 1, and explained. Find five others. Circle them, number them 2–6, and briefly explain each by following the example given.



Explanations

1 Extra phosphate should not be present

27. A primitive eukaryote was discovered that displayed a unique nucleic acid as its genetic material. Analysis revealed the following observations:
(a) X-ray diffraction studies display a general pattern similar to DNA, but with somewhat different dimensions and more irregularity.
(b) A major hyperchromic shift is evident upon heating and monitoring UV absorption at 260 nm.
(c) Base-composition analysis reveals four bases in the following proportions:

Adenine = 8% Hypoxanthine = 18%
Guanine = 37% Xanthine = 37%

(d) About 75 percent of the sugars are deoxyribose, whereas 25 percent are ribose.
Attempt to solve the structure of this molecule by postulating a model that is consistent with the foregoing observations.
28. While demethylation can convert thymine to uracil, deamination can convert cytosine to uracil. Suppose these two mutations occur in a cell. What would be the impact on the DNA structure?
29. Consider the structure of double-stranded DNA. When DNA is placed into distilled water, it denatures; however, by adding NaCl, the DNA renatures. Why?
30. *Newsdate: March 1, 2030.* A unique creature has been discovered during exploration of outer space. Recently, its genetic material has been isolated and analyzed, and has been found to be similar in some ways to DNA in chemical makeup. It contains in abundance the 4-carbon sugar erythrose and a molar equivalent of phosphate groups. In addition, it contains six nitrogenous bases: adenine (A), guanine (G), thymine (T), cytosine (C), hypoxanthine (H), and xanthine (X). These bases exist in the following relative proportion:

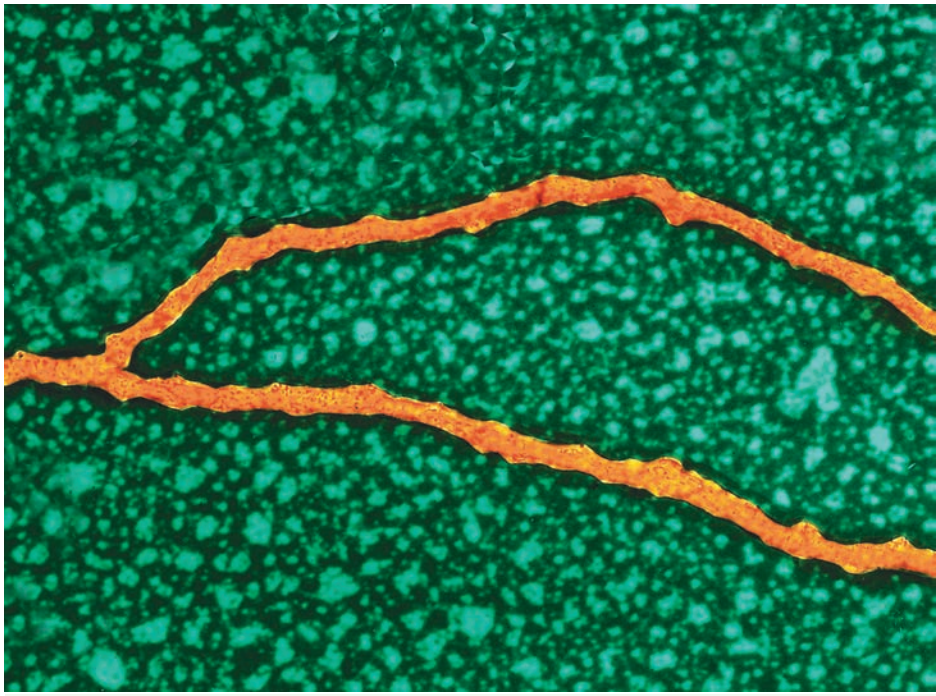
$$A = T = H \quad \text{and} \quad C = G = X$$

X-ray diffraction studies have established a regularity in the molecule and a constant diameter of about 30 Å. Together, these data have suggested a model for the structure of this molecule. (a) Propose a general model of this molecule, and briefly describe it. (b) What base-pairing properties must exist for H and for X in the model? (c) Given the constant diameter of 30 Å, do you think *either* (i) both H and X are purines or both pyrimidines, *or* (ii) one is a purine and one is a pyrimidine?
31. You are provided with DNA samples from two newly discovered bacterial viruses. Based on the various analytical techniques discussed in this chapter, construct a research protocol that would be useful in characterizing and contrasting the DNA of both viruses. Indicate the type of information you hope to obtain for each technique included in the protocol.
32. During electrophoresis, DNA molecules can easily be separated according to size because all DNA molecules have the same charge–mass ratio and the same shape (long rod). Would you expect RNA molecules to behave in the same manner as DNA during electrophoresis? Why or why not?
33. Assume that you are interested in separating short (25–40 nucleotides) DNA molecules from a pool of longer molecules in the 900–1000 nucleotide range. You have two recipes for making your polyacrylamide gels; one recipe uses 12 percent acrylamide and would be considered a "hard gel," while the other uses 4 percent acrylamide and would be considered a loose gel. Which recipe would you consider using and why?

10 DNA Replication

CHAPTER CONCEPTS

- Genetic continuity between parental and progeny cells is maintained by semiconservative replication of DNA, as predicted by the Watson–Crick model.
- Semiconservative replication uses each strand of the parent double helix as a template, and each newly replicated double helix includes one "old" and one "new" strand of DNA.
- DNA synthesis is a complex but orderly process, occurring under the direction of a myriad of enzymes and other proteins.
- DNA synthesis involves the polymerization of nucleotides into polynucleotide chains.
- DNA synthesis is similar in prokaryotes and eukaryotes, but more complex in eukaryotes.
- In eukaryotes, DNA synthesis at the ends of chromosomes (telomeres) poses a special problem, overcome by a unique RNA-containing enzyme, telomerase.



Transmission electron micrograph of human DNA from a HeLa cell, illustrating a replication fork characteristic of active DNA synthesis.

Following Watson and Crick's proposal for the structure of DNA, scientists focused their attention on how this molecule is replicated. Replication is an essential function of the genetic material and must be executed precisely if genetic continuity between cells is to be maintained following cell division. It is an enormous, complex task. Consider for a moment that more than 3×10^9 (3 billion) base pairs exist within the human genome. To duplicate faithfully the DNA of just one of these chromosomes requires a mechanism of extreme precision. Even an error rate of only 10^{-6} (one in a million) will still create 3000 errors (obviously an excessive number) during each replication cycle of the genome. Although it is not error free, and much of evolution would not have occurred if it were, an extremely accurate system of DNA replication has evolved in all organisms.

As Watson and Crick noted in the concluding paragraph of their 1953 paper, their proposed model of the double helix provided the initial insight into how replication occurs. Called semiconservative replication, this mode of DNA duplication was soon to receive strong support from numerous studies of viruses, prokaryotes, and eukaryotes. Once the general *mode* of replication was clarified, research to determine the precise details of the *synthesis* of DNA intensified. What has since been discovered is that numerous enzymes and other proteins are needed to copy a DNA helix. Because of the complexity of the chemical events during synthesis, this subject remains an extremely active area of research.

In this chapter, we will discuss the general mode of replication, as well as the specific details of DNA synthesis. The research leading to such knowledge is another link in our understanding of life processes at the molecular level.

10.1 DNA Is Reproduced by Semiconservative Replication

Watson and Crick recognized that, because of the arrangement and nature of the nitrogenous bases, each strand of a DNA double helix could serve as a template for the synthesis of its complement (**Figure 10–1**). They proposed that, if the helix were unwound, each nucleotide along the two parent strands would have an affinity for its complementary nucleotide. As we learned in Chapter 9, the complementarity is due to the potential hydrogen bonds that can be formed. If thymidylic acid (T) were present, it would "attract" adenylic acid (A); if guanylic acid (G) were present, it would attract cytidylic acid (C); likewise, A would attract T, and C would attract G. If these nucleotides were then covalently linked into polynucleotide chains along both templates, the result would be the production of two identical double strands of DNA. Each replicated DNA molecule would consist of one "old" and one "new" strand, hence the reason for the name **semiconservative replication.**

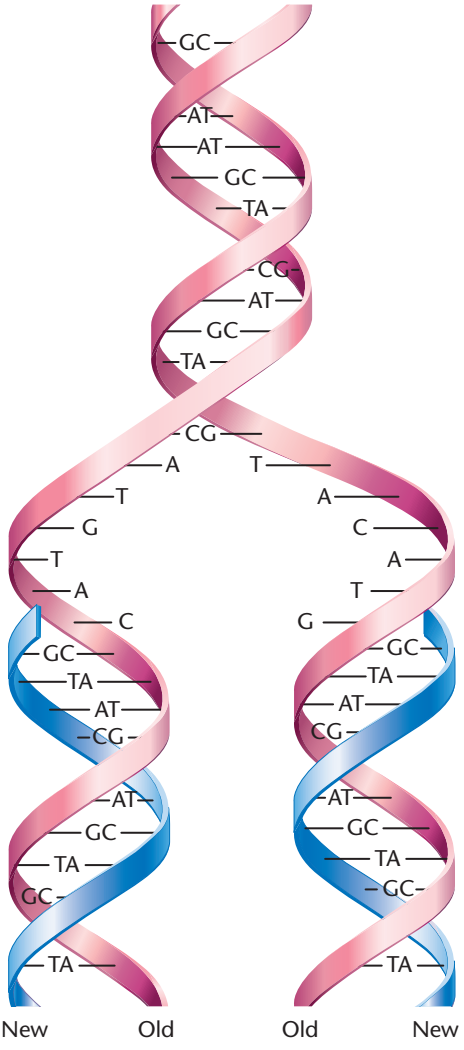


FIGURE 10–1 Generalized model of semiconservative replication of DNA. New synthesis is shown in blue.

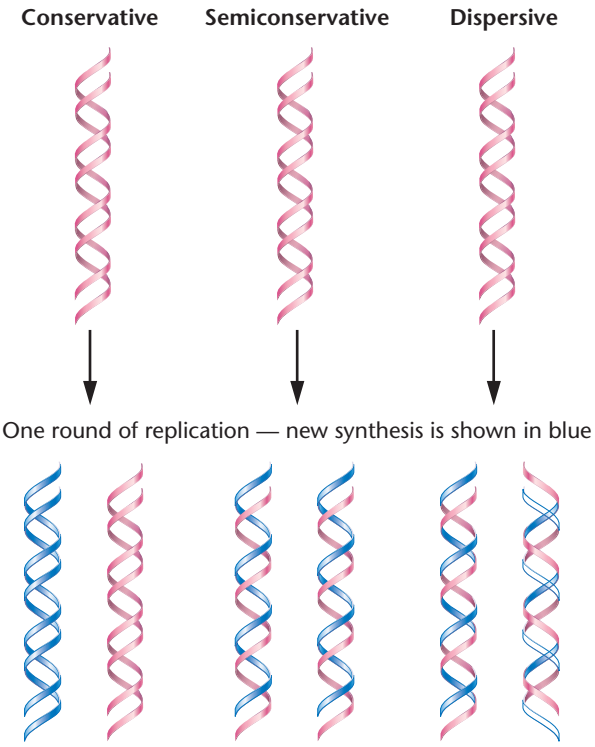


FIGURE 10–2 Results of one round of replication of DNA for each of the three possible modes by which replication could be accomplished.

Two other theoretical modes of replication are possible that also rely on the parental strands as a template (**Figure 10–2**). In **conservative replication,** complementary polynucleotide chains are synthesized as described earlier. Following synthesis, however, the two newly created strands then come together and the parental strands reassociate. The original helix is thus "conserved."

In the second alternative mode, called **dispersive replication,** the parental strands are dispersed into two new double helices following replication. Hence, each strand consists of both old and new DNA. This mode would involve cleavage of the parental strands during replication. It is the most complex of the three possibilities and is therefore considered to be least likely to occur. It could not, however, be

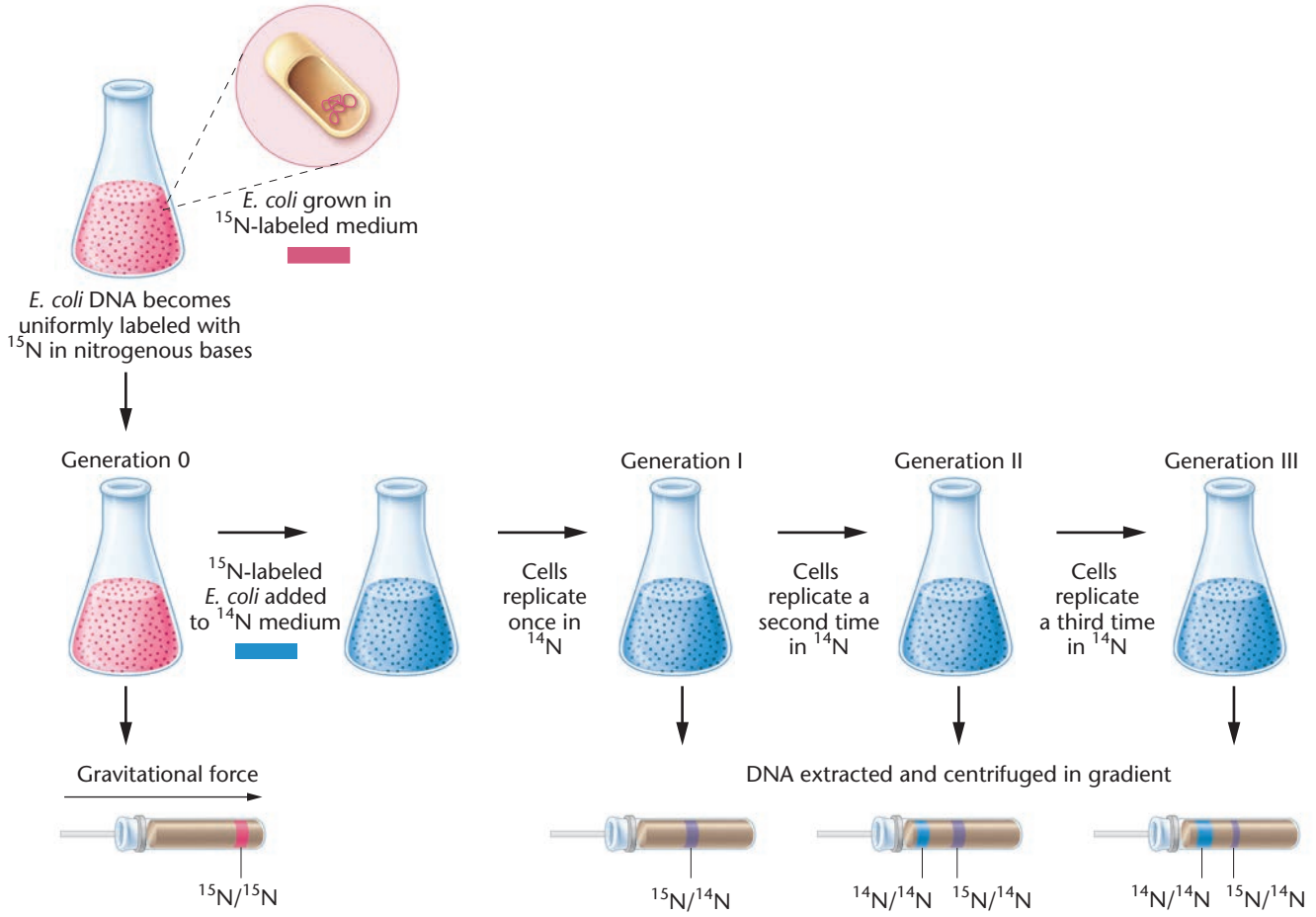


FIGURE 10–3 The Meselson–Stahl experiment.

ruled out as an experimental model. Figure 10–2 shows the theoretical results of a single round of replication by each of the three different modes.

The Meselson–Stahl Experiment

In 1958, Matthew Meselson and Franklin Stahl published the results of an experiment providing strong evidence that semiconservative replication is the mode used by bacterial cells to produce new DNA molecules. They grew *E. coli* cells for many generations in a medium that had $^{15}NH_4Cl$ (ammonium chloride) as the only nitrogen source. A "heavy" isotope of nitrogen, ^{15}N contains one more neutron than the naturally occurring ^{14}N isotope; thus, molecules containing ^{15}N are more dense than those containing ^{14}N. Unlike radioactive isotopes, ^{15}N is stable. After many generations in this medium, almost all nitrogen-containing molecules in the *E. coli* cells, including the nitrogenous bases of DNA, contained the heavier isotope.

Critical to the success of this experiment, DNA containing ^{15}N can be distinguished from DNA containing ^{14}N. The experimental procedure involves the use of a technique referred to as **sedimentation equilibrium centrifugation** (also called buoyant density gradient centrifugation). Samples are forced by centrifugation through a density gradient of a heavy metal salt, such as cesium chloride. Molecules of DNA will reach equilibrium when their density equals the density of the gradient medium. In this case, ^{15}N-DNA will reach this point at a position closer to the bottom of the tube than will ^{14}N-DNA.

In this experiment (**Figure 10–3**), uniformly labeled ^{15}N cells were transferred to a medium containing only $^{14}NH_4Cl$. Thus, all "new" synthesis of DNA during replication contained only the "lighter" isotope of nitrogen. The time of transfer to the new medium was taken as time zero $t = 0$. The *E. coli* cells were allowed to replicate over several generations, with cell samples removed after each replication cycle. DNA was isolated from each sample and subjected to sedimentation equilibrium centrifugation.

After one generation, the isolated DNA was present in only a single band of intermediate density—the expected result for semiconservative replication in which each replicated molecule was composed of one new ^{14}N-strand and one

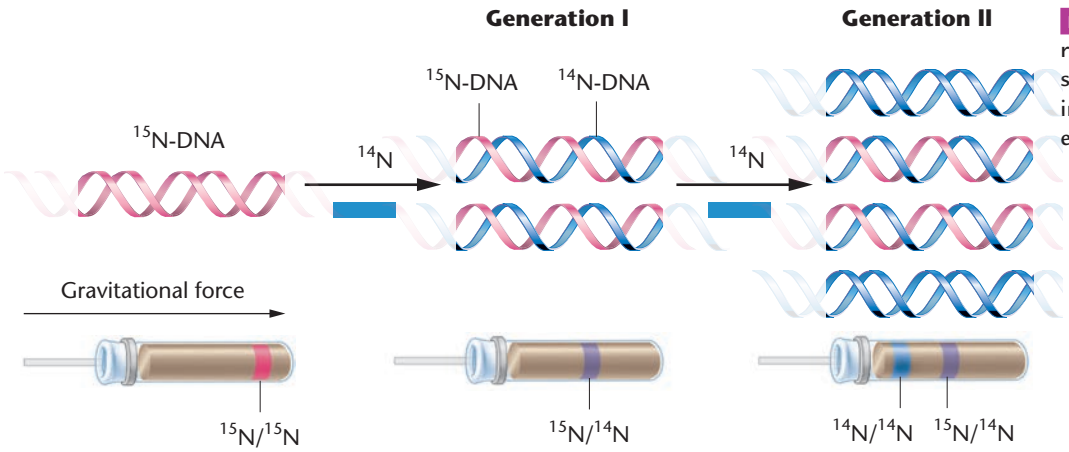


FIGURE 10–4 The expected results of two generations of semiconservative replication in the Meselson–Stahl experiment.

old ^{15}N-strand (**Figure 10–4**). This result was not consistent with the prediction of conservative replication, in which two distinct bands would occur; thus this mode may be ruled out.

After two cell divisions, DNA samples showed two density bands—one intermediate band and one lighter band corresponding to the ^{14}N position in the gradient. Similar results occurred after a third generation, except that the proportion of the lighter band increased. This was again consistent with the interpretation that replication is semiconservative.

You may have realized that a molecule exhibiting intermediate density is also consistent with dispersive replication. However, Meselson and Stahl also ruled out this mode of replication on the basis of two observations. First, after the first generation of replication in an ^{14}N-containing medium, they isolated the hybrid molecule and heat denatured it.

Recall from Chapter 9 that heating will separate a duplex into single strands. When the densities of the single strands of the hybrid were determined, they exhibited *either* an ^{15}N profile *or* an ^{14}N profile, but *not* an intermediate density. This observation is consistent with the semiconservative mode but inconsistent with the dispersive mode.

Furthermore, if replication were dispersive, *all* generations after $t = 0$ would demonstrate DNA of an intermediate density. In each generation after the first, the ratio of ^{15}N/^{14}N would decrease, and the hybrid band would become lighter and lighter, eventually approaching the ^{14}N band. This result was not observed. The Meselson–Stahl experiment provided conclusive support for semiconservative replication in bacteria and tended to rule out both the conservative and dispersive modes.

ESSENTIAL POINT

In 1958, Meselson and Stahl resolved the question of which of three potential modes of replication is utilized by *E. coli* during the duplication of DNA in favor of semiconservative replication, showing that newly synthesized DNA consists of one old strand and one new strand. ■

NOW SOLVE THIS

10–1 In the Meselson–Stahl experiment, which of the three modes of replication could be ruled out after one round of replication? After two rounds?

■ **HINT:** *This problem involves an understanding of the nature of the experiment as well as the difference between the three possible modes of replication. The key to its solution is to determine which mode will not create "hybrid" helices after one round of replication.*

Semiconservative Replication in Eukaryotes

In 1957, the year before the work of Meselson and Stahl was published, J. Herbert Taylor, Philip Woods, and Walter Hughes presented evidence that semiconservative replication also occurs in eukaryotic organisms. They experimented with root tips of the broad bean *Vicia faba,* which are an excellent source of dividing cells. These researchers were able to monitor the process of replication by labeling DNA with ^{3}H-thymidine, a radioactive precursor of DNA, and performing autoradiography.

Autoradiography is a common technique that, when applied cytologically, pinpoints the location of a radioisotope in a cell. In this procedure, a photographic emulsion is placed over a histological preparation containing cellular material (root tips, in this experiment), and the preparation is stored in the dark. The slide is then developed, much as photographic film is processed. Because the radioisotope emits energy, upon development the emulsion turns black at the approximate point of emission. The end result is the presence of dark spots or "grains" on the surface of the section, identifying the location of newly synthesized DNA within the cell.

Taylor and his colleagues grew root tips for approximately one generation in the presence of the radioisotope and then placed them in unlabeled medium in which cell division continued. At the conclusion of each generation,

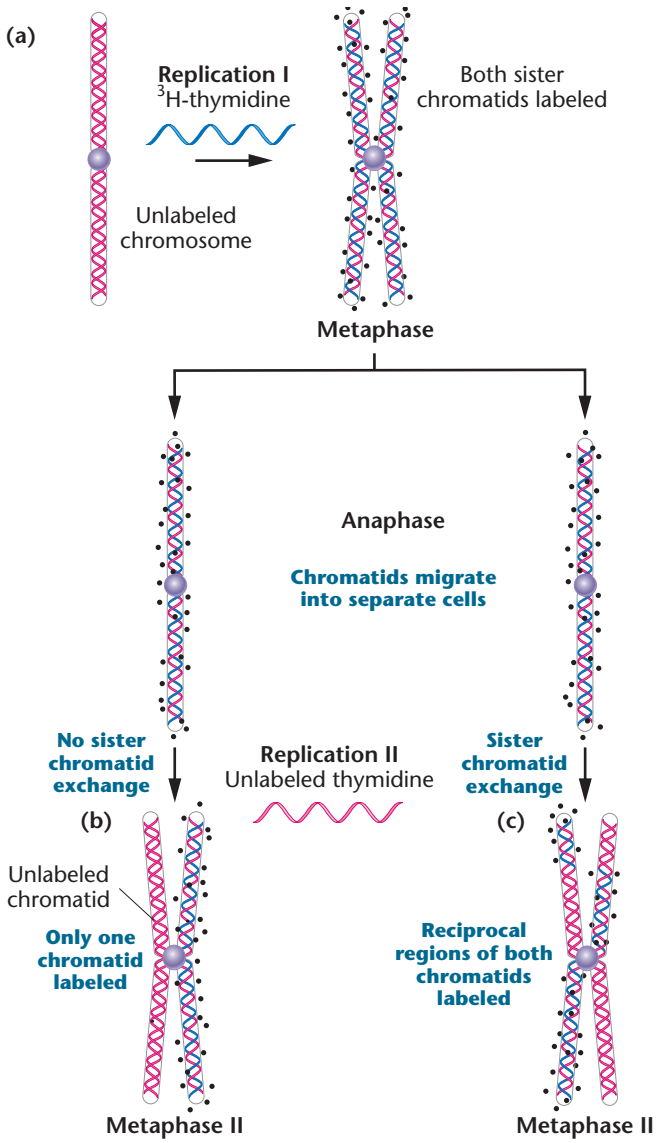


FIGURE 10–5 The Taylor–Woods–Hughes experiment, demonstrating the semiconservative mode of replication of DNA in the root tips of *Vicia faba*. (a) An unlabeled chromosome proceeds through the cell cycle in the presence of ^{3}H-thymidine. As it enters mitosis, both sister chromatids of the chromosome are labeled, as shown, by autoradiography. After a second round of replication (b), this time in the absence of ^{3}H-thymidine, only one chromatid of each chromosome is expected to be surrounded by grains. Except where a reciprocal exchange has occurred between sister chromatids (c), the expectation was upheld.

they arrested the cultures at metaphase by adding colchicine (a chemical derived from the crocus plant that poisons the spindle fibers) and then examined the chromosomes by autoradiography. They found radioactive thymidine only in association with chromatids that contained newly synthesized DNA. **Figure 10–5** illustrates the replication of a single chromosome over two division cycles, including the distribution of grains.

These results are compatible with the semiconservative mode of replication. After the first replication cycle in the presence of the isotope, both sister chromatids show radioactivity, indicating that each chromatid contains one new radioactive DNA strand and one old unlabeled strand. After the second replication cycle, *which takes place in unlabeled medium,* only one of the two sister chromatids of each chromosome should be radioactive because half of the parent strands are unlabeled. With only the minor exceptions of *sister chromatid exchanges* (discussed in Chapter 7), this result was observed.

Together, the Meselson–Stahl experiment and the experiment by Taylor, Woods, and Hughes soon led to the general acceptance of the semiconservative mode of replication. Later studies with other organisms reached the same conclusion and also strongly supported Watson and Crick's proposal for the double helix model of DNA.

ESSENTIAL POINT

Taylor, Woods, and Hughes demonstrated semiconservative replication in eukaryotes using the root tips of the broad bean as the source of dividing cells. ■

Origins, Forks, and Units of Replication

To enhance our understanding of semiconservative replication, let's briefly consider a number of relevant issues. The first concerns the **origin of replication.** Where along the chromosome is DNA replication initiated? Is there only a single origin, or does DNA synthesis begin at more than one point? Is any given point of origin random, or is it located at a specific region along the chromosome? Second, once replication begins, does it proceed in a single direction or in both directions away from the origin? In other words, is replication **unidirectional** or **bidirectional**?

To address these issues, we need to introduce two terms. First, at each point along the chromosome where replication is occurring, the strands of the helix are unwound, creating what is called a **replication fork** (see the Chapter Opening photograph on p. 196). Such a fork will initially appear at the point of origin of synthesis and then move along the DNA duplex as replication proceeds. If replication is bidirectional, two such forks will be present, migrating in opposite directions away from the origin. The second term refers to the length of DNA that is replicated following one initiation event at a single origin. This is a unit referred to as the **replicon.**

The evidence is clear regarding the origin and direction of replication. John Cairns tracked replication in *E. coli,* using radioactive precursors of DNA synthesis and autoradiography. He was able to demonstrate that in *E. coli* there is only a single region, called *oriC,* where replication is initiated. The presence of only a single origin is characteristic of bacteria, which have only one circular chromosome. Since DNA synthesis in bacteriophages and bacteria originates at

a single point, the entire chromosome constitutes one replicon. In *E. coli,* the replicon consists of the entire genome of 4.6 Mb (4.6 million base pairs).

Figure 10–6 illustrates Cairns's interpretation of DNA replication in *E. coli.* This interpretation and the accompanying micrograph do not answer the question of unidirectional versus bidirectional synthesis. However, other results, derived from studies of bacteriophage lambda, have demonstrated that replication is bidirectional, moving away from *oriC* in both directions. Figure 10–6 therefore interprets Cairns's work with that understanding. Bidirectional replication creates two replication forks that migrate farther and farther apart as replication proceeds. These forks eventually merge, as semiconservative replication of the entire chromosome is completed, at a termination region, called *ter.*

Later in this chapter, we will see that in eukaryotes, each chromosome contains multiple points of origin.

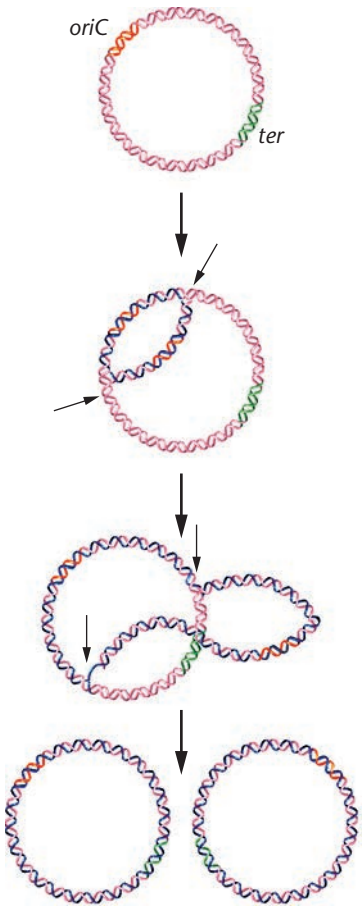


FIGURE 10–6 Bidirectional replication of the *E. coli* chromosome. The thin black arrows identify the advancing replication forks. The micrograph is of a bacterial chromosome in the process of replication, comparable to the figure next to it.

10.2 DNA Synthesis in Bacteria Involves Five Polymerases, as Well as Other Enzymes

To say that replication is semiconservative and bidirectional describes the overall *pattern* of DNA duplication and the association of finished strands with one another once synthesis is completed. However, it says little about the more complex issue of how the actual *synthesis* of long complementary polynucleotide chains occurs on a DNA template. Like most questions in molecular biology, this one was first studied using microorganisms. Research on DNA synthesis began about the same time as the Meselson–Stahl work, and the topic is still an active area of investigation. What is most apparent in this research is the tremendous complexity of the biological synthesis of DNA.

DNA Polymerase I

Studies of the enzymology of DNA replication were first reported by Arthur Kornberg and colleagues in 1957. They isolated an enzyme from *E. coli* that was able to direct DNA synthesis in a cell-free (*in vitro*) system. The enzyme is called **DNA polymerase I,** because it was the first of several similar enzymes to be isolated.

Kornberg determined that there were two major requirements for *in vitro* DNA synthesis under the direction of DNA polymerase I: (1) all four deoxyribonucleoside triphosphates (dNTPs) and (2) template DNA. If any one of the four deoxyribonucleoside triphosphates was omitted from the reaction, no measurable synthesis occurred. If derivatives of these precursor molecules other than the nucleoside triphosphate were used (nucleotides or nucleoside diphosphates), synthesis also did not occur. If no template DNA was added, synthesis of DNA occurred but was reduced greatly.

Most of the synthesis directed by Kornberg's enzyme appeared to be exactly the type required for semiconservative replication. The reaction is summarized in **Figure 10–7**, which depicts the addition of a single nucleotide. The enzyme has since been shown to consist of a single polypeptide containing 928 amino acids.

The way in which each nucleotide is added to the growing chain is a function of the specificity of DNA polymerase I. As shown in **Figure 10–8**, the precursor dNTP contains the three phosphate groups attached to the 5′-carbon of deoxyribose. As the two terminal phosphates are cleaved during synthesis, the remaining phosphate attached to the 5′-carbon is covalently linked to the 3′-OH group of the deoxyribose to which it is added. Thus, **chain elongation** occurs in the **5′ to 3′ direction** by the addition of one nucleotide at a time to the growing 3′ end. Each step provides a newly exposed 3′-OH group that can participate in the next addition of a nucleotide as DNA synthesis proceeds.

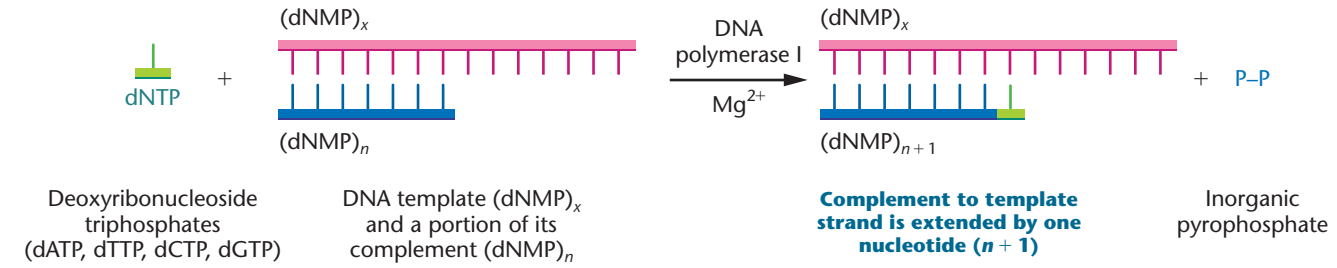


FIGURE 10–7 The chemical reaction catalyzed by DNA polymerase I. During each step, a single nucleotide is added to the growing complement of the DNA template using a nucleoside triphosphate as the substrate. The release of inorganic pyrophosphate drives the reaction energetically.

Having isolated DNA polymerase I and demonstrated its catalytic activity, Kornberg next sought to demonstrate the accuracy, or fidelity, with which the enzyme replicated the DNA template. Because technology for ascertaining the nucleotide sequences of the template and newly synthesized strand was not yet available in 1957, he initially had to rely on several indirect methods.

One of Kornberg's approaches was to compare the nitrogenous base compositions of the DNA template with those of the recovered DNA product. Using several sources of DNA (phage T2, *E. coli,* and calf thymus), he discovered that, within experimental error, the base composition of each product agreed with the template DNA used. This suggested that the templates were replicated faithfully.

ESSENTIAL POINT

Arthur Kornberg isolated the enzyme DNA polymerase I from *E. coli* and showed that it is capable of directing *in vitro* DNA synthesis, provided that a template and precursor nucleoside triphosphates were supplied. ■

DNA Polymerase II, III, IV, and V

While DNA polymerase I clearly directs the synthesis of DNA, a serious reservation about the enzyme's true biological role was raised in 1969. Paula DeLucia and John Cairns discovered a mutant strain of *E. coli* that was deficient in polymerase I activity. The mutation was designated *polA1*. In the absence of the functional enzyme, this mutant strain of *E. coli* still duplicated its DNA and successfully reproduced. However, the cells were deficient in their ability to repair DNA. For example, the mutant strain is highly sensitive to ultraviolet light (UV) and radiation, both of which damage DNA and are mutagenic. Nonmutant bacteria are able to repair a great deal of UV-induced damage.

These observations led to two conclusions:

1. At least one other enzyme that is responsible for replicating DNA *in vivo* is present in *E. coli* cells.
2. DNA polymerase I serves a secondary function *in vivo,* now believed to be critical to the maintenance of fidelity of DNA synthesis.

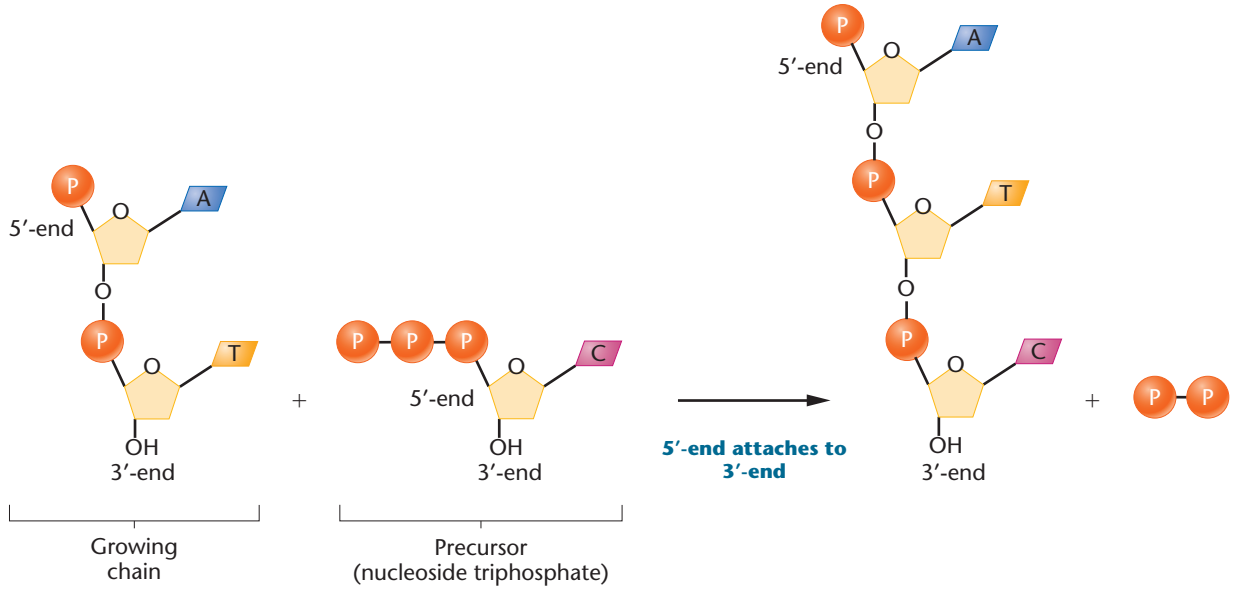


FIGURE 10–8 Demonstration of 5′ to 3′ synthesis of DNA.

TABLE 10.1 Properties of Bacterial DNA Polymerases I, II, and III

Properties	I	II	III
Initiation of chain synthesis	–	–	–
5′–3′ polymerization	+	+	+
3′–5′ exonuclease activity	+	+	+
5′–3′ exonuclease activity	+	–	–
Molecules of polymerase/cell	400	?	15

To date, four other unique DNA polymerases have been isolated from cells lacking polymerase I activity and from normal cells that contain polymerase I. Table 10.1 contrasts several characteristics of DNA polymerase I with **DNA polymerase II** and **III.** Although none of the three can *initiate* DNA synthesis on a template, all three can *elongate* an existing DNA strand, called a **primer.**

All the DNA polymerase enzymes are large proteins exhibiting a molecular weight in excess of 100,000 Daltons (Da). All three possess 3′ to 5′ exonuclease activity, which means that they have the potential to polymerize in one direction and then pause, reverse their direction, and excise nucleotides just added. As we will discuss later in the chapter, this activity provides a capacity to proofread newly synthesized DNA and to remove and replace incorrect nucleotides.

DNA polymerase I also demonstrates 5′ to 3′ exonuclease activity. This activity allows the enzyme to excise nucleotides, starting at the end at which synthesis begins and proceeding in the same direction of synthesis. Two final observations probably explain why Kornberg isolated polymerase I and not polymerase III: polymerase I is present in greater amounts than is polymerase III, and it is also much more stable.

What then are the roles of the polymerases *in vivo*? Polymerase III is the enzyme responsible for the 5′ to 3′ polymerization essential to *in vivo* replication. Its 3′ to 5′ exonuclease activity also provides a proofreading function that is activated when it inserts an incorrect nucleotide. When this occurs, synthesis stalls and the polymerase "reverses course," excising the incorrect nucleotide. Then, it proceeds back in the 5′ to 3′ direction, synthesizing the complement of the template strand. Polymerase I is believed to be responsible for removing the primer, as well as for the synthesis that fills gaps produced after this removal. Its exonuclease activities also allow for its participation in DNA repair. Polymerase II, as well as **polymerase IV and V,** are involved in various aspects of repair of DNA that has been damaged by external forces, such as ultraviolet light. Polymerase II is encoded by a gene whose transcription is activated by disruption of DNA synthesis at the replication fork.

ESSENTIAL POINT

The discovery of the *polA1* mutant strain of *E. coli,* capable of DNA replication despite its lack of polymerase I activity, cast doubt on the enzyme's hypothesized *in vivo* replicative function. Polymerase III has been identified as the enzyme responsible for DNA replication *in vivo.* ■

The DNA Polymerase III Holoenzyme

We conclude this section by emphasizing the complexity of the DNA polymerase III enzyme, henceforth referred to as DNA Pol III. The active form of DNA Pol III, referred to as the **holoenzyme,** is made up of unique polypeptide subunits, ten of which have been identified (Table 10.2). The largest subunit, α, along with subunits ε and θ, form a complex called the **core enzyme,** which imparts the catalytic function to the holoenzyme. In *E. coli*, each holoenzyme contains two, and possibly three, core enzyme complexes. As part of each core, the α subunit is responsible for DNA synthesis along the template strands, whereas the ε subunit possesses 3′ to 5′ exonuclease capability, essential to proofreading. The need for more than one core enzyme will soon become apparent. A second group of five subunits (γ, δ, δ', χ, and υ) are complexed to form what is called the **sliding clamp loader,** which pairs with the core enzyme and facilitates the function of a critical component of the holoenzyme, called the **sliding DNA clamp.** The enzymatic function of the sliding clamp loader is dependent on energy generated by the hydrolysis of ATP. The sliding DNA clamp links to the core enzyme and is made up of multiple copies of the β subunit, taking on the shape of a donut, whereby it can open and shut, to encircle the unreplicated DNA helix. By doing so, and being linked to the core enzyme, the clamp leads the way during synthesis, maintaining the binding of the core enzyme to the template during polymerization of nucleotides. Thus, the length of DNA that is replicated by the core enzyme before it detaches from the template, a property referred

TABLE 10.2 Subunits of the DNA Polymerase III Holoenzyme

Subunit	Function	Groupings
α	5′–3′ polymerization	Core enzyme: Elongates polynucleotide chain and proofreads
ε	3′–5′ exonuclease	
θ	Core assembly	
γ, δ, δ', χ, υ	Loads enzyme on template (serves as clamp loader)	γ complex
β	Sliding clamp structure (processivity factor)	
τ	Dimerizes core complex	

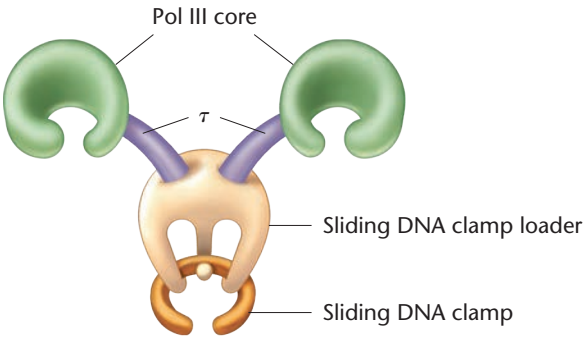


FIGURE 10–9 The components making up the DNA Pol III holoenzyme, as described in the text. While there may be three core enzyme complexes present in the holoenzyme, for simplicity, we illustrate only two.

to as **processivity,** is vastly increased. There is one sliding clamp per core enzyme. Finally, one τ subunit interacts with each core enzyme, linking it to the sliding clamp loader.

The DNA Pol III holoenzyme is diagrammatically illustrated in **Figure 10–9**. You should compare the diagram to the description of each component above. Note that we have shown the holoenzyme to contain two core enzyme complexes, although as stated above, a third one may be present. The components of the DNA Pol III holoenzyme will be referred to in the discussion that follows.

10.3 Many Complex Issues Must Be Resolved during DNA Replication

We have thus far established that in bacteria and viruses replication is semiconservative and bidirectional along a single replicon. We also know that synthesis is catalyzed by DNA polymerase III and occurs in the 5′ to 3′ direction. Bidirectional synthesis creates two replication forks that move in opposite directions away from the origin of synthesis. As we can see from the following list, many issues remain to be resolved in order to provide a comprehensive understanding of DNA replication:

1. The helix must undergo localized unwinding, and the resulting "open" configuration must be stabilized so that synthesis may proceed along both strands.
2. As unwinding and subsequent DNA synthesis proceed, increased coiling creates tension further down the helix, which must be reduced.
3. A primer of some sort must be synthesized so that polymerization can commence under the direction of DNA polymerase III. Surprisingly, RNA, not DNA, serves as the primer.
4. Once the RNA primers have been synthesized, DNA polymerase III begins to synthesize the DNA complement of both strands of the parent molecule. Because the two strands are antiparallel to one another, continuous synthesis in the direction that the replication fork moves is possible along only one of the two strands. On the other strand, synthesis must be discontinuous and thus involves a somewhat different process.
5. The RNA primers must be removed prior to completion of replication. The gaps that are temporarily created must be filled with DNA complementary to the template at each location.
6. The newly synthesized DNA strand that fills each temporary gap must be joined to the adjacent strand of DNA.
7. While DNA polymerases accurately insert complementary bases during replication, they are not perfect, and, occasionally, incorrect nucleotides are added to the growing strand. A proofreading mechanism that also corrects errors is an integral process during DNA synthesis.

As we consider these points, examine Figures 10–10, 10–11, and 10–12 to see how each issue is resolved. Figure 10–13 summarizes the model of DNA synthesis.

Unwinding the DNA Helix

As discussed earlier, there is a single point of origin along the circular chromosome of most bacteria and viruses at which DNA synthesis is initiated. This region of the *E. coli* chromosome has been particularly well studied. Called ***oriC,*** it consists of 245 nucleotide pairs characterized by repeating sequences of 9 and 13 bases (called **9mers** and **13mers**). One particular protein, called **DnaA** (because it is encoded by the gene called *dnaA*), is responsible for the initial step in unwinding the helix. A number of subunits of the DnaA protein bind to each of several 9mers. This step facilitates the subsequent binding of **DnaB** and **DnaC** proteins that further open and destabilize the helix. Proteins such as these, which require the energy supplied by the hydrolysis of ATP in order to break hydrogen bonds and denature the double helix, are called **helicases.** Other proteins, called **single-stranded binding proteins (SSBPs),** stabilize this open conformation.

As unwinding proceeds, a coiling tension is created ahead of the replication fork, often producing **supercoiling.** In circular molecules, supercoiling may take the form of added twists and turns of the DNA, much like the coiling you can create in a rubber band by stretching it out and then twisting one end. Such supercoiling can be relaxed by **DNA gyrase,** a member of a larger group of enzymes referred to as **DNA topoisomerases.** The gyrase makes

either single- or double-stranded "cuts" and also catalyzes localized movements that have the effect of "undoing" the twists and knots created during supercoiling. The strands are then resealed. These various reactions are driven by the energy released during ATP hydrolysis.

Together, the DNA, the polymerase complex, and associated enzymes make up an array of molecules that participate in DNA synthesis and are part of what we have previously called the *replisome*.

ESSENTIAL POINT

During the initiation of DNA synthesis, the double helix unwinds, forming a replication fork at which synthesis begins. Proteins stabilize the unwound helix and assist in relaxing the coiling tension created ahead of the replication fork. ■

Initiation of DNA Synthesis Using an RNA Primer

Once a small portion of the helix is unwound, what else is needed to initiate synthesis? As we have seen, DNA polymerase III requires a primer with a free 3′-hydroxyl group in order to elongate a polynucleotide chain. Since none is available in a circular chromosome, this absence prompted researchers to investigate how the first nucleotide could be added. It is now clear that RNA serves as the primer that initiates DNA synthesis.

A short segment of RNA (about 10 to 12 ribonucleotides long), complementary to DNA, is first synthesized on the DNA template. Synthesis of the RNA is directed by a form of RNA polymerase called **primase,** which does not require a free 3′ end to initiate synthesis. It is to this short segment of RNA that DNA polymerase III begins to add deoxyribonucleotides, initiating DNA synthesis. A conceptual diagram of initiation on a DNA template is shown in Figure 10–10. Later, the RNA primer is clipped out and replaced with DNA. This is thought to occur under the direction of DNA polymerase I. Recognized in viruses, bacteria, and several eukaryotic organisms, RNA priming is a universal phenomenon during the initiation of DNA synthesis.

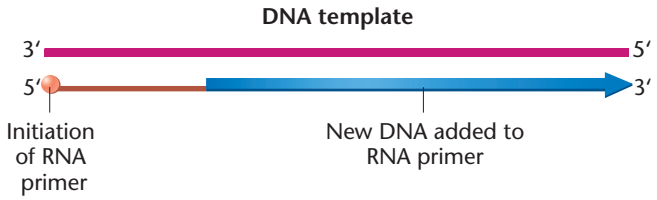


FIGURE 10–10 The initiation of DNA synthesis. A complementary RNA primer is first synthesized, to which DNA is added. All synthesis is in the 5′ to 3′ direction. Eventually, the RNA primer is replaced with DNA under the direction of DNA polymerase I.

ESSENTIAL POINT

DNA synthesis is initiated at specific sites along each template strand by the enzyme primase, resulting in short segments of RNA that provide suitable 3′ ends upon which DNA polymerase III can begin polymerization. ■

Continuous and Discontinuous DNA Synthesis

We must now revisit the fact that the two strands of a double helix are **antiparallel** to each other—that is, one runs in the 5′–3′ direction, while the other has the opposite 3′–5′ polarity. Because DNA polymerase III synthesizes DNA in only the 5′–3′ direction, synthesis along an advancing replication fork occurs in one direction on one strand and in the opposite direction on the other.

As a result, as the strands unwind and the replication fork progresses down the helix (Figure 10–11), only

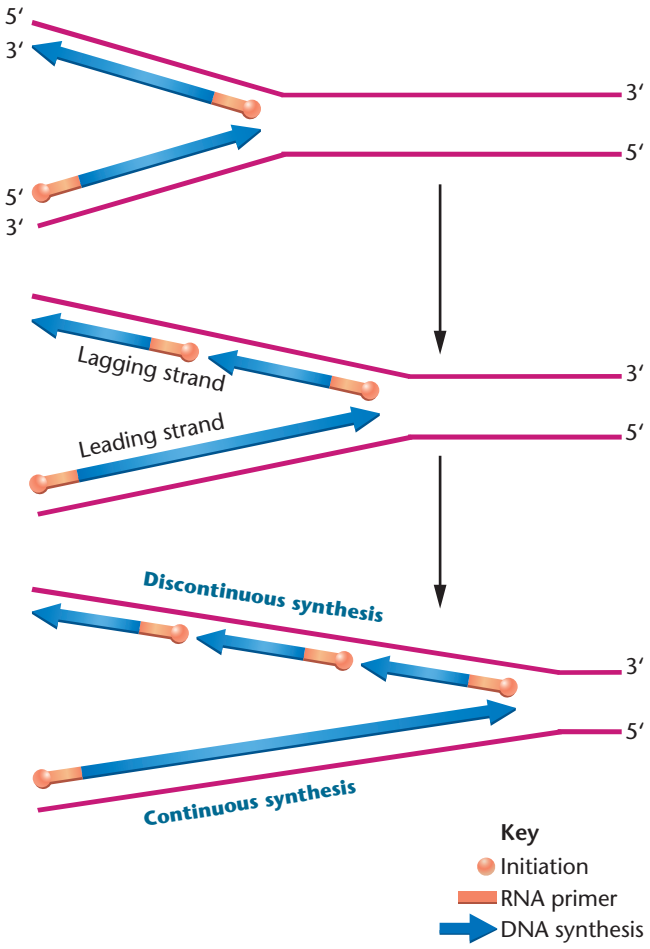


FIGURE 10–11 Opposite polarity of synthesis along the two strands of DNA is necessary because they run antiparallel to one another, and because DNA polymerase III synthesizes in only one direction (5′ to 3′). On the lagging strand, synthesis must be discontinuous, resulting in the production of Okazaki fragments. On the leading strand, synthesis is continuous. RNA primers are used to initiate synthesis on both strands.

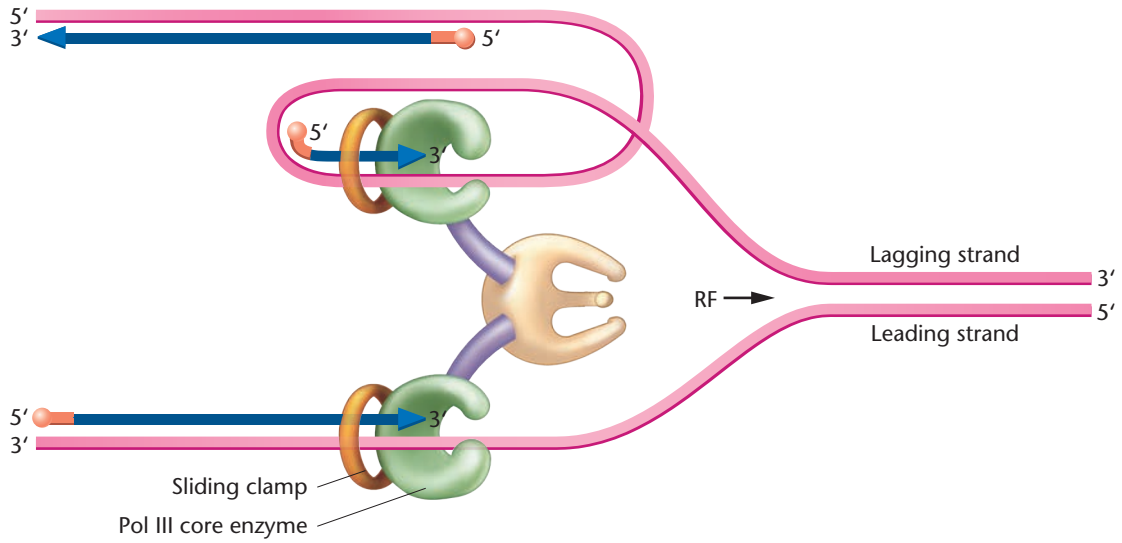


FIGURE 10–12 Illustration of how concurrent DNA synthesis may be achieved on both the leading and lagging strands at a single replication fork (RF). The lagging template strand is "looped" in order to invert the physical direction of synthesis, but not the biochemical direction. The enzyme functions as a dimer, with each core enzyme achieving synthesis on one or the other strand.

one strand can serve as a template for **continuous DNA synthesis.** This newly synthesized DNA is called the **leading strand.** As the fork progresses, many points of initiation are necessary on the opposite DNA template, resulting in **discontinuous DNA synthesis*** of the **lagging strand.**

Evidence supporting the occurrence of discontinuous DNA synthesis was first provided by Reiji and Tuneko Okazaki. They discovered that when bacteriophage DNA is replicated in *E. coli,* some of the newly formed DNA that is hydrogen bonded to the template strand is present as small fragments containing 1000 to 2000 nucleotides. RNA primers are part of each such fragment. These pieces, now called **Okazaki fragments,** are converted into longer and longer DNA strands of higher molecular weight as synthesis proceeds.

Discontinuous synthesis of DNA requires enzymes that both remove the RNA primers and unite the Okazaki fragments into the lagging strand. As we have noted, DNA polymerase I removes the primers and replaces the missing nucleotides. Joining the fragments is the work of **DNA ligase,** which is capable of catalyzing the formation of the phosphodiester bond that seals the nick between the discontinuously synthesized strands. The evidence that DNA ligase performs this function during DNA synthesis is strengthened by the observation of a ligase-deficient mutant strain (*lig*) of *E. coli,* in which a large number of unjoined Okazaki fragments accumulate.

*Because DNA synthesis is continuous on one strand and discontinuous on the other, the term **semidiscontinuous synthesis** is sometimes used to describe the overall process.

Concurrent Synthesis Occurs on the Leading and Lagging Strands

Given the model just discussed, we might ask how the holoenzyme of DNA Pol III synthesizes DNA on both the leading and lagging strands. Can both strands be replicated simultaneously at the same replication fork, or are the events distinct, involving two separate copies of the enzyme? Evidence suggests that both strands are replicated simultaneously, with each strand acted upon by one of the two core enzymes that are part of the DNA Pol III holoenzyme. As **Figure 10–12** illustrates, if the lagging strand is spooled out, forming a loop, nucleotide polymerization can occur simultaneously on both template strands under the direction of the holoenzyme. After the synthesis of 1000 to 2000 nucleotides, the monomer of the enzyme on the lagging strand will encounter a completed Okazaki fragment, at which point it releases the lagging strand. A new loop of the lagging strand is spooled out, and the process is repeated. Looping inverts the orientation of the template but not the direction of actual synthesis on the lagging strand, which is always in the 5′ to 3′ direction. As mentioned above, it is believed that there is a third core enzyme associated with the DNA Pol III holoenzyme, and that it functions in the synthesis of Okazaki fragments. For simplicity, we will include only two core enzymes in this and subsequent figures.

Another important feature of the holoenzyme that facilitates synthesis at the replication fork is the donut-shaped sliding DNA clamp that surrounds the unreplicated double helix and is linked to the advancing core enzyme. This clamp prevents the core enzyme from dissociating from the template as polymerization proceeds. By doing so,

the clamp is responsible for vastly increasing the processivity of the core enzyme—that is, the number of nucleotides that may be continually added prior to dissociation from the template. This function is critical to the rapid *in vivo* rate of DNA synthesis during replication.

ESSENTIAL POINT

Concurrent DNA synthesis occurs continuously on the leading strand and discontinuously on the opposite lagging strand, resulting in short Okazaki fragments that are later joined by DNA ligase. ■

NOW SOLVE THIS

10–2 An alien organism was investigated. When DNA replication was studied, a unique feature was apparent: No Okazaki fragments were observed. Create a model of DNA that is consistent with this observation.

■ **HINT:** *This problem involves an understanding of the process of DNA synthesis in prokaryotes, as depicted in Figure 10–12. The key to its solution is to consider why Okazaki fragments are observed during DNA synthesis and how their formation relates to DNA structure, as described in the Watson–Crick model.*

Proofreading and Error Correction Occur during DNA Replication

The immediate purpose of DNA replication is the synthesis of a new strand that is precisely complementary to the template strand at each nucleotide position. Although the action of DNA polymerases is very accurate, synthesis is not perfect and a noncomplementary nucleotide is occasionally inserted erroneously. To compensate for such inaccuracies, the DNA polymerases all possess 3′ to 5′ exonuclease activity. This property imparts the potential for them to detect and excise a mismatched nucleotide (in the 3′ to 5′ direction). Once the mismatched nucleotide is removed, 5′ to 3′ synthesis can again proceed. This process, called **proofreading,** increases the fidelity of synthesis by a factor of about 100. In the case of the holoenzyme form of DNA polymerase III, the epsilon (ε) subunit is directly involved in the proofreading step. In strains of *E. coli* with a mutation that has rendered the ε subunit nonfunctional, the error rate (the mutation rate) during DNA synthesis is increased substantially.

10.4 A Coherent Model Summarizes DNA Replication

We can now combine the various aspects of DNA replication occurring at a single replication fork into a coherent model, as shown in **Figure 10–13**. At the advancing fork, a helicase is unwinding the double helix. Once unwound, single-stranded binding proteins associate with the strands, preventing the re-formation of the helix. In advance of the replication fork, DNA gyrase functions to diminish the tension created as the helix supercoils. Each half of the dimeric polymerase is a core enzyme bound to one of the template strands by a β-subunit sliding clamp. Continuous synthesis occurs on the leading strand, while the lagging strand must loop out and around the polymerase in order for simultaneous (concurrent) synthesis to occur on both strands. Not shown in the figure, but essential to replication on the lagging strand, is the action of DNA polymerase I and DNA ligase, which together replace the RNA primers with DNA and join the Okazaki fragments, respectively.

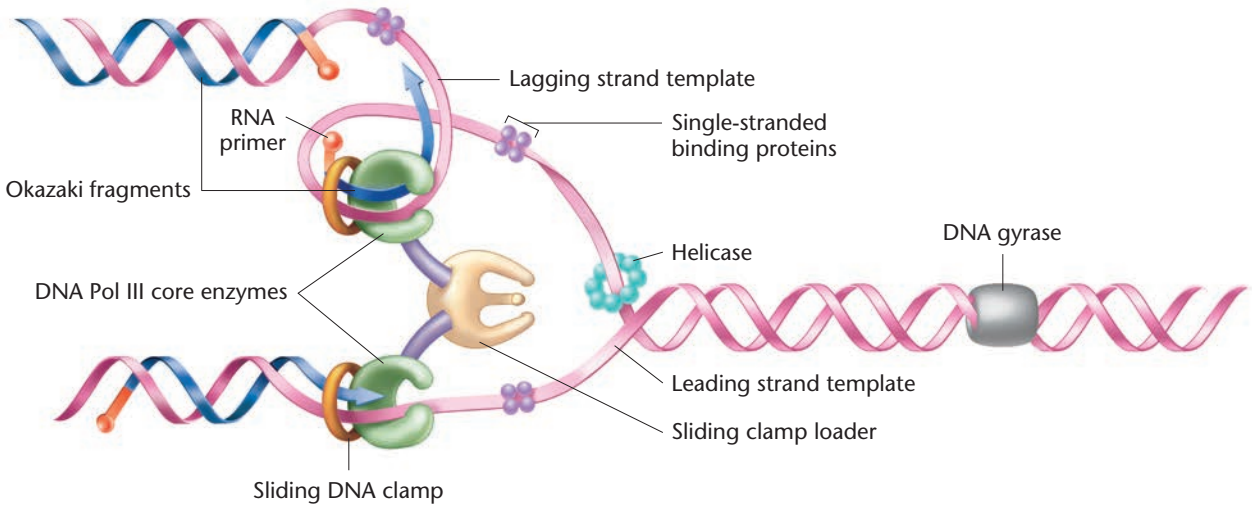


FIGURE 10–13 Summary of DNA synthesis at a single replication fork. Various enzymes and proteins essential to the process are shown.

Because the investigation of DNA synthesis is still an extremely active area of research, this model will no doubt be extended in the future. In the meantime, it provides a summary of DNA synthesis against which genetic phenomena can be interpreted.

10.5 Replication Is Controlled by a Variety of Genes

Much of what we know about DNA replication in viruses and bacteria is based on genetic analysis of the process. For example, we have already discussed studies involving the *polA1* mutation, which revealed that DNA polymerase I is not the major enzyme responsible for replication. Many other mutations interrupt or seriously impair some aspect of replication, such as the ligase-deficient and the proofreading-deficient mutations mentioned previously. Because such mutations are lethal, genetic analysis frequently uses **conditional mutations,** which are expressed under one condition but not under a different condition. For example, a **temperature-sensitive mutation** may not be expressed at a particular *permissive* temperature. When mutant cells are grown at a *restrictive* temperature, the mutant phenotype is expressed and can be studied. By examining the effect of the loss of function associated with the mutation, the investigation of such temperature-sensitive mutants can provide insight into the product and the associated function of the normal, nonmutated gene.

As shown in Table 10.3, a variety of genes in *E. coli* specify the subunits of the DNA polymerases and encode products involved in specification of the origin of synthesis, helix unwinding and stabilization, initiation and priming, relaxation of supercoiling, repair, and ligation.

TABLE 10.3 Some of the Various *E. coli* Genes and Their Products or Role in Replication

Gene	Product or Role
polA	DNA polymerase I
polB	DNA polymerase II
dnaE, N, Q, X, Z	DNA polymerase III subunits
dnaG	Primase
dnaA, I, P	Initiation
dnaB, C	Helicase at *oriC*
gyrA, B	Gyrase subunits
lig	DNA ligase
rep	DNA helicase
ssb	Single-stranded binding proteins
rpoB	RNA polymerase subunit

The discovery of such a large group of genes attests to the complexity of the process of replication, even in the relatively simple prokaryote. Given the enormous quantity of DNA that must be unerringly replicated in a very brief time, this level of complexity is not unexpected. As we will see in the next section, the process is even more involved and therefore more difficult to investigate in eukaryotes.

10.6 Eukaryotic DNA Replication Is Similar to Replication in Prokaryotes, but Is More Complex

Eukaryotic DNA replication shares many features with replication in bacteria. In both systems, double-stranded DNA is unwound at replication origins, replication forks are formed, and bidirectional DNA synthesis creates leading and lagging strands from single-stranded DNA templates under the direction of DNA polymerase. Eukaryotic polymerases have the same fundamental requirements for DNA synthesis as do bacterial polymerases: four deoxyribonucleoside triphosphates, a template, and a primer. However, eukaryotic DNA replication is more complex, due to several features of eukaryotic DNA. For example, eukaryotic cells contain much more DNA, this DNA is complexed with nucleosomes, and eukaryotic chromosomes are linear rather than circular. In this section, we will describe some of the ways in which eukaryotes deal with this added complexity.

Initiation at Multiple Replication Origins

The most obvious difference between eukaryotic and prokaryotic DNA replication is that eukaryotic replication must deal with greater amounts of DNA. For example, yeast cells contain three times as much DNA, and *Drosophila* cells contain 40 times as much as *E. coli* cells. In addition, eukaryotic DNA polymerases synthesize DNA at a rate 25 times slower (about 2000 nucleotides per minute) than that in prokaryotes. Under these conditions, replication from a single origin on a typical eukaryotic chromosome would take days to complete. However, replication of entire eukaryotic genomes is usually accomplished in a matter of minutes to hours.

To facilitate the rapid synthesis of large quantities of DNA, eukaryotic chromosomes contain multiple replication origins. Yeast genomes contain between 250 and 400 origins, and mammalian genomes have as many as 25,000. Multiple origins are visible under the electron microscope as “replication bubbles” that form as the DNA helix opens up, each bubble providing two potential replication forks

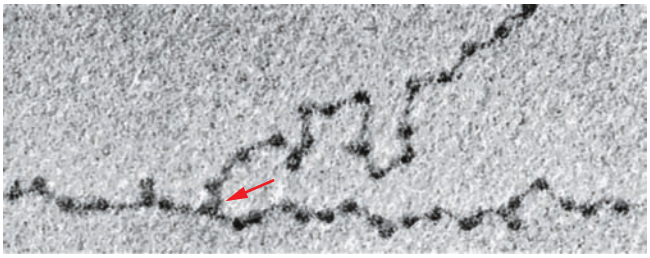


FIGURE 10–14 An electron micrograph of a eukaryotic replicating fork demonstrating the presence of histone-protein-containing nucleosomes on both branches.

(Figure 10–14). Origins in yeast, called **autonomously replicating sequences (ARSs),** consist of approximately 120 base pairs containing a **consensus sequence** (meaning a sequence that is the same, or nearly the same, in all yeast ARSs) of 11 base pairs. Origins in mammalian cells appear to be unrelated to specific sequence motifs and may be defined more by chromatin structure over a 6–55 kb region.

Multiple Eukaryotic DNA Polymerases

To accommodate the increased number of replicons, eukaryotic cells contain many more DNA polymerase molecules than do bacterial cells. For example, a single *E. coli* cell contains about 15 molecules of DNA polymerase III, but a mammalian cell contains tens of thousands of DNA polymerase molecules.

Eukaryotes also utilize a larger number of different DNA polymerase types than do prokaryotes. The human genome contains genes that encode at least 14 different DNA polymerases, only three of which are involved in the majority of nuclear genome DNA replication.

Pol α, δ, and ε are the major forms of the enzyme involved in initiation and elongation during eukaryotic nuclear DNA synthesis, so we will concentrate our discussion on these. Two of the four subunits of the **Pol α enzyme** synthesize RNA primers on both the leading and lagging strands. After the RNA primer reaches a length of about 10 ribonucleotides, another subunit adds 10–20 complementary deoxyribonucleotides. Pol α is said to possess low **processivity,** a term that refers to the strength of the association between the enzyme and its substrate, and thus the length of DNA that is synthesized before the enzyme dissociates from the template. Once the primer is in place, an event known as **polymerase switching** occurs, whereby Pol α dissociates from the template and is replaced by Pol δ or ε. These enzymes extend the primers on opposite strands of DNA, possess much greater processivity, and exhibit 3′ to 5′ exonuclease activity, thus having the potential to proofread. Pol ε synthesizes DNA on the leading strand, and Pol δ synthesizes the lagging strand. Both Pol δ and ε participate in other DNA synthesizing events in the cell, including several types of DNA repair and recombination. All three DNA polymerases are essential for viability.

As in prokaryotic DNA replication, the final stages in eukaryotic DNA replication involve replacing the RNA primers with DNA and ligating the Okazaki fragments on the lagging strand. In eukaryotes, the Okazaki fragments are about ten times smaller (100 to 150 nucleotides) than in prokaryotes.

Included in the remainder of DNA-replicating enzymes is Pol γ, which is found exclusively in mitochondria, synthesizing the DNA present in that organelle. Other DNA polymerases are involved in DNA repair and replication through regions of the DNA template that contain damage or distortions.

Replication through Chromatin

One of the major differences between prokaryotic and eukaryotic DNA is that eukaryotic DNA is complexed with DNA-binding proteins, existing in the cell as *chromatin*. As we will discuss later in the text (see Chapter 11), chromatin consists of regularly repeating units called nucleosomes, each of which consists of about 200 base pairs of DNA wrapped around eight histone protein molecules. Before polymerases can begin synthesis, nucleosomes and other DNA-binding proteins must be stripped away or otherwise modified to allow the passage of replication proteins. As DNA synthesis proceeds, the histones and nonhistone proteins must rapidly reassociate with the newly formed duplexes, reestablishing the characteristic nucleosome pattern. Electron microscopy studies, such as the one shown in Figure 10–14, show that nucleosomes form immediately after new DNA is synthesized at replication forks.

In order to re-create nucleosomal chromatin on replicated DNA, the synthesis of new histone proteins is tightly coupled to DNA synthesis during the S phase of the cell cycle. Research data suggest that nucleosomes are disrupted just ahead of the replication fork and that the preexisting histone proteins can assemble with newly synthesized histone proteins into new nucleosomes. The new nucleosomes are assembled behind the replication fork, onto the two daughter strands of DNA. The assembly of new nucleosomes is carried out by **chromatin assembly factors (CAFs)** that move along with the replication fork.

ESSENTIAL POINT

DNA replication in eukaryotes is more complex than replication in prokaryotes, using multiple replication origins, multiple forms of DNA polymerases, and factors that disrupt and assemble nucleosomal chromatin. ■

10.7 The Ends of Linear Chromosomes Are Problematic during Replication

A final difference between prokaryotic and eukaryotic DNA synthesis stems from the structural differences in their chromosomes. Unlike the closed, circular DNA of bacteria and most bacteriophages, eukaryotic chromosomes are linear. During replication, two special problems arise at the "ends" of these linear double-stranded DNA molecules.

The first problem is that the double-stranded "ends" of DNA molecules at the termini of linear chromosomes potentially resemble the **double-stranded breaks (DSBs)** that can occur when a chromosome becomes fragmented internally as a result of DNA damage. In such cases, double-stranded loose ends can fuse, resulting in chromosomal translocations. If the ends do not fuse, they are vulnerable to degradation by nucleases. The second problem occurs during DNA replication, because DNA polymerases cannot synthesize new DNA at the tips of single-stranded 5′ ends.

To deal with these two problems, linear eukaryotic chromosomes end in distinctive sequences called **telomeres** that help preserve the integrity and stability of the chromosomes. Telomeres create inert chromosome ends, protecting intact eukaryotic chromosomes from improper fusion or degradation. They also solve the 5′ end replication problem, as we will describe next.

Telomere Structure

First discovered by Elizabeth Blackburn and Joe Gall in their study of micronuclei—the smaller of two nuclei in the ciliated protozoan *Tetrahymena*—the DNA at the protozoan's chromosome ends consists of the short tandem repeating sequence TTGGGG. This sequence is present many times on one of the two DNA strands making up each telomere. This strand is referred to as the G-rich strand, in contrast to its complementary strand, the so-called C-rich strand, which displays the repeated sequence AACCCC. In a similar way, all vertebrates contain the sequence TTAGGG at the ends of G-rich strands, repeated several thousand times in somatic cells. Since each linear chromosome ends with two helical DNA strands running antiparallel to one another, one strand has a 3′-ending and the other has a 5′-ending. It is the 3′-strand that is the G-rich one. This has special significance during telomere replication.

But first, let's describe how this tandemly repeated DNA confers inertness to the chromosome ends. One model is based on the discovery that the 3′-ending G-rich strand extends as an overhang, lacking a complement, and thus forms a single-stranded tail at the terminus of each telomere. In *Tetrahymena,* this tail is only 12 to 16 nucleotides long. However, in vertebrates, it may be several hundred nucleotides long. The final conformation of these tails has been correlated with chromosome inertness. Though not considered complementary in the same way as A-T and G-C base pairs are, G-containing nucleotides are nevertheless capable of base pairing with one another when several are aligned opposite another G-rich sequence. Thus, the G-rich single-stranded tails are capable of looping back on themselves, forming multiple G-G hydrogen bonds to create what are referred to as **G-quartets.** The resulting loops at the chromosome ends (called **t-loops**) are much like those created when you tie your shoelaces into a bow. It is believed that these structures, in combination with specific proteins that bind to them, essentially close off the ends of chromosomes and make them resistant to nuclease digestion and DNA fusions.

Replication at the Telomere

Now let's consider the problem that semiconservative replication poses the 5′-ends of double-stranded DNA molecules. As we have learned previously in this chapter, DNA replication initiates from short RNA primers, synthesized on both leading and lagging strands (**Figure 10–15**). Primers are

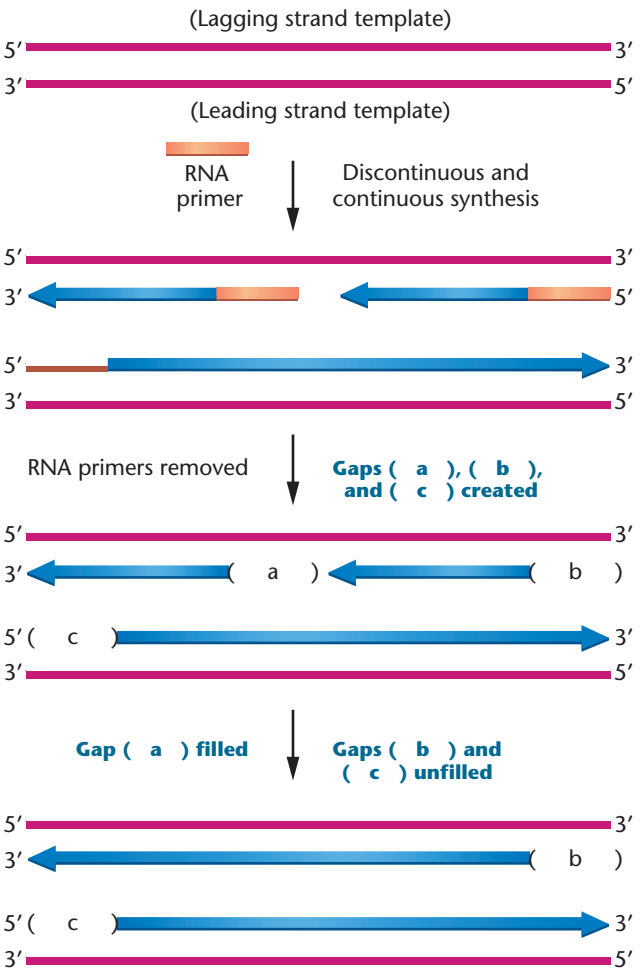


FIGURE 10–15 Diagram illustrating the difficulty encountered during the replication of the ends of linear chromosomes. Gaps (b) and (c) are left following synthesis both the leading and lagging strands.

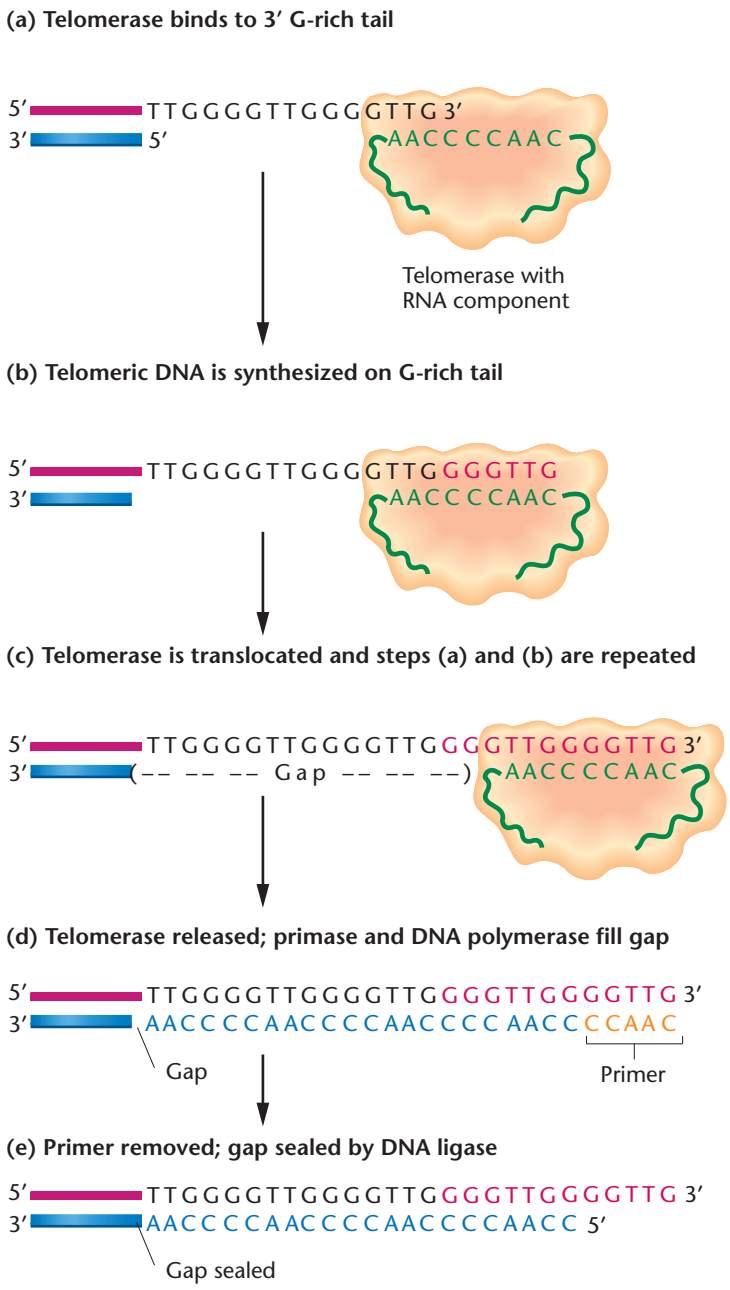


FIGURE 10–16 The predicted solution to the problem posed in Figure 10–15. The enzyme telomerase (with its RNA component shown in green) directs synthesis of repeated TTGGGG sequences, resulting in the formation of an extended 3′-overhang. This facilitates DNA synthesis on the opposite strand, filling in the gap that would otherwise have been created on the ends of linear chromosomes during each replication cycle.

necessary because DNA polymerase requires a free 3′-OH on which to initiate synthesis. After replication is completed, these RNA primers are removed. The resulting gaps within the new daughter strands are filled by DNA polymerase and sealed by ligase. These internal gaps have free 3′-OH groups available at the ends of the Okazaki fragments for DNA polymerase to initiate synthesis. The problem arises at the gaps left at the 5′ ends of the newly synthesized DNA [gaps (b) and (c) in Figure 10–15]. These gaps cannot be filled by DNA polymerase because no free 3′-OH groups are available for the initiation of synthesis.

Thus, in the situation depicted in Figure 10–15, gaps remain on newly synthesized DNA strands at each successive round of synthesis, shortening the double-stranded ends of the chromosome by the length of the RNA primer. With each round of replication, the shortening becomes more severe in each daughter cell, *eventually extending beyond the telomere and potentially deleting gene-coding regions.*

The solution to this so-called *end-replication problem* is provided by a unique eukaryotic enzyme called **telomerase.** Telomerase was first discovered by Elizabeth Blackburn and her graduate student, Carol Greider, in studies of *Tetrahymena*. As noted earlier, telomeric DNA in eukaryotes consists of many short, repeated nucleotide sequences, with the G-rich strand overhanging in the form of a single-stranded tail. In *Tetrahymena* the tail contains several repeats of the sequence 5′-TTGGGG-3′. As we will see, telomerase is capable of adding several more repeats of this six-nucleotide sequence to the 3′-end of the G-rich strand (using 5′–3′ synthesis). Detailed investigation by Blackburn and Greider of how the *Tetrahymena* telomerase enzyme accomplishes this synthesis yielded an extraordinary finding. The enzyme is highly unusual in that it is a *ribonucleoprotein,* containing within its molecular structure a short piece of RNA that is essential to its catalytic activity. The telomerase RNA component (TERC) serves as both a “guide” to proper attachment of the enzyme to the telomere and a “template” for synthesis of its DNA complement. Synthesis of DNA using RNA as a template is called **reverse transcription.** The telomerase reverse transcriptase, called TERT, is the catalytic subunit of the telomerase enzyme. In addition to TERC and TERT, telomerase contains a number of accessory proteins. In *Tetrahymena,* TERC contains the sequence AACCCCAAC, within which is found the complement of the repeating telomeric DNA sequence that must be synthesized (TTGGGG).

Figure 10–16 shows one model of how researchers envision the enzyme working. Part of the RNA sequence of the enzyme (shown in green) base-pairs with the ending sequence of the single-stranded overhanging DNA, while the remainder of the RNA extends beyond the overhang. Next, reverse transcription of this extending RNA sequence—synthesizing DNA on an RNA template—extends the length of the G-rich lagging strand. It is believed that the enzyme is then translocated toward the (newly formed) end of the strand, and the same events are repeated, continuing the extension process.

Once the telomere has been lengthened by telomerase, conventional DNA synthesis ensues. Primase lays down a primer near the end of the telomere, then DNA polymerase

and ligase fill most of the gap [Figure 10–16(d) and (e)]. When the primer is removed a small gap remains. However, this gap is located well beyond the original end of the chromosome, thus preventing any chromosome shortening. Another model suggests that the DNA extension, created by telomerase, acts as a primer and facilitates DNA synthesis on the opposite C-rich strand. In this model, the single-stranded extension loops back on itself, providing the 3′-OH group necessary for initiation of synthesis to fill the gap.

Telomerase function has now been found in all eukaryotes studied. As we will discuss later in the text (see Chapter 11), telomeric DNA sequences have been highly conserved throughout evolution, reflecting the critical function of telomeres. As mentioned earlier, in humans, the telomeric DNA sequence on the lagging strand that is repeated is 5′-TTAGGG-3′, differing from *Tetrahymena* by only one nucleotide.

In most eukaryotic somatic cells, telomerase is not active, and thus, with each cell division, the telomeres of each chromosome do shorten. After many divisions, the telomere may be seriously eroded, causing the cell to lose the capacity for further division. Most stem cells and malignant cells, on the other hand, maintain telomerase activity and this may contribute to their immortality. In the "Genetics, Technology, and Society" feature below, we will see that telomerase activity and telomere length have been linked to aging, cancer, and other diseases.

ESSENTIAL POINT

Replication at the ends of linear chromosomes in eukaryotes poses a special problem that can be solved by the presence of telomeres and by a unique RNA-containing enzyme called telomerase. ■

GENETICS, TECHNOLOGY, AND SOCIETY

Telomeres: The Key to Immortality?

Humans, like all multicellular organisms, grow old and die. As we age, our immune systems become less efficient, wound healing is impaired, and tissues lose resilience. Why do we go through these age-related declines, and can we reverse the march to mortality? Some recent research suggests that the answers to these questions may lie at the ends of our chromosomes.

Human cells, both those in our bodies and those growing in culture dishes, have a finite life span. When placed into tissue culture dishes, normal human fibroblasts lose their ability to grow and divide after about 50 cell divisions. Eventually, they die. Although we don't know whether cellular senescence directly causes organismal aging, the evidence is suggestive. For example, cultured cells derived from young people undergo more divisions than those from older people; cells from short-lived species stop growing after fewer divisions than those from longer-lived species; and cells from patients with premature aging syndromes undergo fewer divisions than those from normal patients.

One significant characteristic of aging cells involves their telomeres. In most mammalian somatic cells, telomeres shorten with each DNA replication because DNA polymerase cannot synthesize new DNA at the ends of each parent strand. However, cells that undergo extensive proliferation, like embryonic cells, germ cells, and adult stem cells, maintain their telomere length by using *telomerase*—a remarkable RNA-containing enzyme that adds telomeric DNA sequences onto the ends of linear chromosomes. However, most somatic cells in adult organisms do not contain telomerase and do not proliferate.

Could we gain perpetual youth and vitality by increasing our telomere lengths? Studies suggest that it may be possible to reverse senescence by artificially increasing the amount of telomerase in our cells. In one study, investigators introduced cloned telomerase genes into normal human cells in culture. The increase in telomerase activity caused the telomeres' lengths to increase, and the cells continued to grow past their typical senescence point. In another study, researchers created a strain of mice that was defective in the TERT subunit of telomerase. These mice showed the classic symptoms of aging, including tissue atrophy, neurodegeneration, and a shortened life span. When the researchers reactivated telomerase function in these prematurely aging adult mice, tissue atrophies and neurodegeneration were reversed and their life spans increased.

These studies suggest that some of the symptoms that accompany old age in humans might also be reversed by activating telomerase genes. However, before we use telomerase to achieve immortality, we need to consider a potential serious side effect—cancer.

Although normal cells shorten their telomeres and undergo senescence after a specific number of cell divisions, cancer cells do not. More than 80 percent of human tumor cells contain telomerase activity, maintain telomere lengths, and achieve immortality. Those that do not contain active telomerase use a less-well-understood mechanism known as ALT (for "alternative lengthening of telomeres").

These observations suggest that we should use caution before attempting to increase telomerase activity in humans. However, they also suggest that we might be able to devise new cancer therapies based on the idea that drugs that inhibit telomerase might destroy cancer cells by allowing telomeres to shorten, thereby forcing the cells into senescence. Because most normal human cells do not express telomerase, such a therapy might specifically affect tumor cells and be less toxic than most current anticancer drugs. Many such anti-telomerase drugs are currently under development, and some are in clinical trials.

Will a deeper understanding of telomeres allow us to both arrest cancers *and* reverse the descent into old age? Time will tell.

Your Turn

Take time, individually or in groups, to answer the following questions. Investigate the references and links to help you understand some of the research on telomeres, aging, and cancer.

1. Mutations in genes coding for telomerase subunits or telomere-binding proteins are associated with a number of human diseases. How do these mutations contribute to these diseases, which are characterized by a wide range of different phenotypes?

The role of telomeres and telomerase in human diseases is discussed in Calado, R.T. and Young, N.S. 2009. Telomere diseases. *N. Eng. J. Med*. 361: 2353–2365.

2. One anti-telomerase drug, called imetelstat (GRN163L), is being developed by Geron Corporation as a treatment for cancer. How does imetelstat work? What is the current status of imetelstat clinical trials? What are some possible side-effects of anti-telomerase drugs?

Read about this drug and its clinical trials on the Geron Web site at http://www.geron.com/imetelstat. *Search on PubMed for scientific papers dealing with GRN163L's anticancer effects.*

3. People suffering from chronic stress appear to have more health problems and to age prematurely. Is there any evidence that chronic stress, poor health, and telomere length are linked? How might stress affect telomere length or vice versa?

Some recent research suggests how these phenomena may be linked. Read about telomeres and stress in Blackburn, E.H. and Epel, E.S. 2012. Too toxic to ignore. *Nature* 490: 169–171.

4. The only treatment known to increase life span and to delay the development of age-related disease is calorie restriction (CR). CR is achieved by reducing food intake while maintaining adequate nutrition. Studies suggest that CR and telomere lengths may be related. How do CR and telomerase act, and interact, to prolong health and life span?

Begin your search for the relationships between telomerase activity and CR with Vera, E., et al. 2013. Telomerase reverse transcriptase synergizes with calorie restriction to increase health span and extend mouse longevity. *PloS One* 8: e53760.

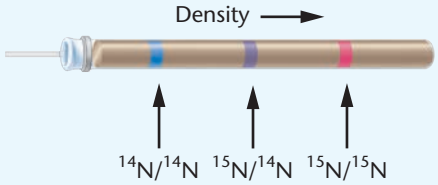
CASE STUDY | Premature aging and DNA helicases

A doctor received a patient in his early 20s with the signs of premature aging. He had gray hair, short stature, cataracts in both eyes, pinched features, beaked nose, tight and thin skin, high-pitched voice, arteriosclerosis, osteoporosis, and ulcers round his ankles. The doctor diagnosed him with Werner syndrome (WS), which is an autosomal recessive disease linked to mutations in the WRN gene on chromosome 8. The WRN protein is a member of the RecQ family of DNA helicases that unwind DNA, and is known to reactivate stalled replication forks during DNA replication. It also plays a role in telomere maintenance since WS patients exhibit accelerated telomere shortening. This disorder raises several interesting questions.

1. How do WRN mutations cause WS?
2. How many mutated WRN alleles does the patient have?
3. What treatment can be given to this patient?
4. What is the main function of a DNA helicase?

INSIGHTS AND SOLUTIONS

1. Predict the theoretical results of conservative and dispersive replication of DNA under the conditions of the Meselson–Stahl experiment. Follow the results through two generations of replication after cells have been shifted to an ^{14}N-containing medium, using the following sedimentation pattern.



2. Mutations in the *dnaA* gene of *E. coli* are lethal and can only be studied following the isolation of conditional, temperature-sensitive mutations. Such mutant strains grow nicely and replicate their DNA at the permissive temperature of 18°C, but they do not grow or replicate their DNA at the restrictive temperature of 37°C. Two observations were useful in determining the function of the DnaA protein product. First, *in vitro* studies using DNA templates that have unwound do not

Solution:

Conservative replication

Dispersive replication

(continued)

Insights and Solutions—continued

require the DnaA protein. Second, if intact cells are grown at 18°C and are then shifted to 37°C, DNA synthesis continues at this temperature until one round of replication is completed and then stops. What do these observations suggest about the role of the *dnaA* gene product?

Solution: At 18°C (the permissive temperature), the mutation is not expressed and DNA synthesis begins. Following the shift to the restrictive temperature, the already initiated DNA synthesis continues, but no new synthesis can begin. Because the DnaA protein is not required for synthesis of unwound DNA, these observations suggest that, *in vivo,* the DnaA protein plays an essential role in DNA synthesis by interacting with the intact helix and somehow facilitating the localized denaturation necessary for synthesis to proceed.

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we focused on how DNA is replicated and synthesized. In particular, we elucidated the general mechanism of replication and described how DNA is synthesized when it is copied. Based on your study of these topics, answer the following fundamental questions:
   (a) What is the experimental basis for concluding that DNA replicates semiconservatively in both prokaryotes and eukaryotes?
   (b) How was it demonstrated that DNA synthesis occurs under the direction of DNA polymerase III and not polymerase I?
   (c) How do we know that *in vivo* DNA synthesis occurs in the 5′ to 3′ direction?
   (d) How do we know that DNA synthesis is discontinuous on one of the two template strands?
   (e) What observations reveal that a "telomere problem" exists during eukaryotic DNA replication, and how did we learn of the solution to this problem?

CONCEPT QUESTION

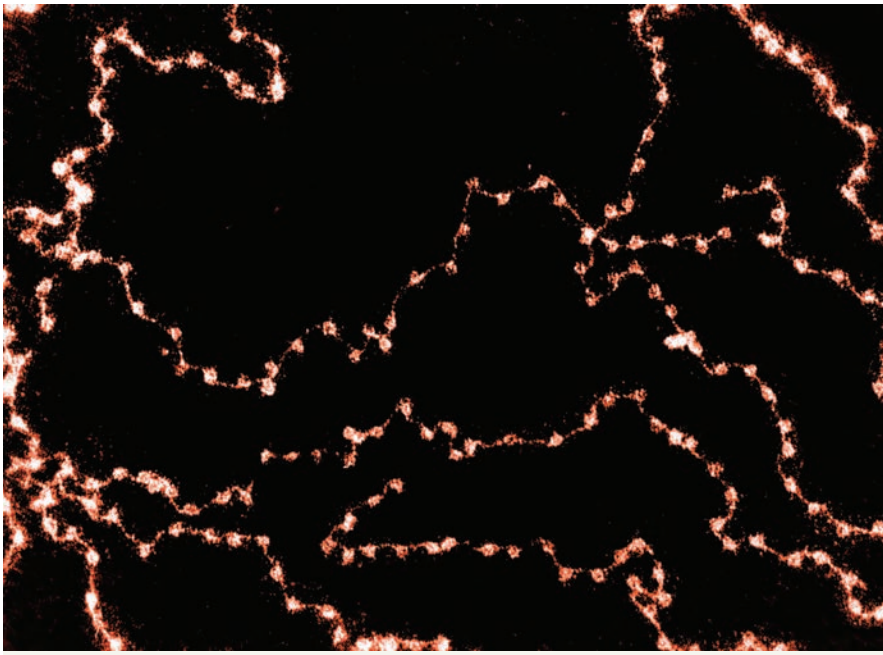
2. Review the Chapter Concepts list on p. 196. These are concerned with the replication and synthesis of DNA. Write a short essay that distinguishes between the terms *replication* and *synthesis*, as applied to DNA. Which of the two is most closely allied with the field of biochemistry? ■

3. Unlike prokaryotes, why do eukaryotes need multiple replication origins?
4. Describe the role of ^{15}N in the Meselson–Stahl experiment.
5. Predict the results of the experiment by Taylor, Woods, and Hughes if replication were (a) conservative and (b) dispersive.
6. Reconsider Problem 30 in Chapter 9. In the model you proposed, could the molecule be replicated semiconservatively? Why? Would other modes of replication work?
7. What is the end replication problem?
8. How did Kornberg assess the fidelity of DNA polymerase I in copying a DNA template?
9. Which characteristics of DNA polymerase I raised doubts that its *in vivo* function is the synthesis of DNA leading to complete replication?
10. You have two strains of bacteria, one in which the process of replication is sensitive to RNase and the other in which the presence of RNase does not affect replication. Speculate why the replications in these two strains are differentially sensitive to RNase.
11. During DNA replication, which enzyme can be disposed of in an organism with a mutant DNA polymerase that does not require a free 3'–OH?
12. Summarize and compare the properties of DNA polymerase I, II, and III.
13. List and describe the function of the ten subunits constituting DNA polymerase III. Distinguish between the holoenzyme and the core enzyme.
14. Distinguish between (a) unidirectional and bidirectional synthesis, and (b) continuous and discontinuous synthesis of DNA.
15. List the proteins that unwind DNA during *in vivo* DNA synthesis. How do they function?
16. Define and indicate the significance of (a) Okazaki fragments, (b) DNA ligase, and (c) primer RNA during DNA replication.
17. What would be the impact of the loss of processivity on DNA Pol III?
18. What are the replication origins in bacteria, yeast, and mammalian cells?
19. If the analysis of DNA from two different microorganisms demonstrated very similar base compositions, are the DNA sequences of the two organisms also nearly identical?
20. Several temperature-sensitive mutant strains of *E. coli* display the following characteristics. Predict what enzyme or function is being affected by each mutation.
   (a) Newly synthesized DNA contains many mismatched base pairs.
   (b) Okazaki fragments accumulate, and DNA synthesis is never completed.
   (c) No initiation occurs.
   (d) Synthesis is very slow.
   (e) Supercoiled strands remain after replication, which is never completed.
21. Many of the gene products involved in DNA synthesis were initially defined by studying mutant *E. coli* strains that could not synthesize DNA. (a) The *dnaE* gene encodes the α subunit of DNA polymerase III. What effect is expected from a mutation in this gene? How could the mutant strain be maintained? (b) The *dnaQ* gene encodes the ε subunit of DNA polymerase. What effect is expected from a mutation in this gene?
22. Assume a hypothetical organism in which DNA replication is conservative. Design an experiment similar to that of Taylor, Woods, and Hughes that will unequivocally establish this fact. Using the format established in Figure 10–5, draw sister chromatids and illustrate the expected results depicting this mode of replication.

11 Chromosome Structure and DNA Sequence Organization

CHAPTER CONCEPTS

- Genetic information in viruses, bacteria, mitochondria, and chloroplasts is most often contained in a short, circular DNA molecule, relatively free of associated proteins.
- Eukaryotic cells, in contrast to viruses and bacteria, contain relatively large amounts of DNA organized into nucleosomes and present during most of the cell cycle as chromatin fibers.
- Uncoiled chromatin fibers characteristic of interphase coil up and condense into chromosomes during eukaryotic cell division.
- Eukaryotic genomes are characterized by both unique and repetitive DNA sequences.
- Eukaryotic genomes consist mostly of noncoding DNA sequences.



A chromatin fiber viewed using a scanning transmission electron microscope (STEM)

Once geneticists understood that DNA houses genetic information, it became very important to determine how DNA is organized into genes and how these basic units of genetic function are organized into chromosomes. In short, the major question had to do with how the genetic material was organized as it makes up the genome of organisms. There has been much interest in this question because knowledge of the organization of the genetic material and associated molecules is important to understanding many other areas of genetics. For example, the way in which the genetic information is stored, expressed, and regulated must be related to the molecular organization of the genetic molecule DNA.

In this chapter, we focus on the various ways DNA is organized into chromosomes. These structures have been studied using numerous techniques, instruments, and approaches, including analysis by light microscopy and electron microscopy. More recently, molecular analysis has provided significant insights into chromosome organization. In the first half of the chapter, after surveying what we know about chromosomes in viruses and bacteria, we examine the large specialized eukaryotic structures called polytene and lampbrush chromosomes. Then, in the second half, we discuss how eukaryotic chromosomes are organized at the molecular level—for example, how DNA is complexed with proteins to form chromatin and how the chromatin fibers characteristic of interphase are condensed into chromosome structures visible during mitosis and meiosis. We conclude the chapter by examining certain aspects of DNA sequence organization in eukaryotic genomes.

11.1 Viral and Bacterial Chromosomes Are Relatively Simple DNA Molecules

The chromosomes of viruses and bacteria are much less complicated than those of eukaryotes. They usually consist of a single nucleic acid molecule, unlike the multiple chromosomes comprising the genome of higher forms. Compared to eukaryotes, the chromosomes contain much less genetic information and the DNA is not as extensively bound to proteins. These characteristics have greatly simplified analysis, and we now have a fairly comprehensive view of the structure of viral and bacterial chromosomes.

The chromosomes of viruses consist of a nucleic acid molecule—either DNA or RNA—that can be either single- or double-stranded. They can exist as circular structures (closed loops), or they can take the form of linear molecules. For example, the single-stranded DNA of the **ϕX174 bacteriophage** and the double-stranded DNA of the **polyoma virus** are closed loops housed within the protein coat of the mature viruses. The **bacteriophage lambda** (λ), on the other hand, possesses a linear double-stranded DNA molecule prior to infection, which closes to form a ring upon its infection of the host cell. Still other viruses, such as the T-even series of bacteriophages, have linear double-stranded chromosomes of DNA that do not form circles inside the bacterial host. Thus, circularity is not an absolute requirement for replication in viruses.

Viral nucleic acid molecules have been seen with the electron microscope. **Figure 11–1** shows a mature bacteriophage λ and its double-stranded DNA molecule in the circular configuration. One constant feature shared by viruses, bacteria, and eukaryotic cells is the ability to package an exceedingly long DNA molecule into a relatively small volume. In λ, the DNA is 17 μm long and must fit into the phage head, which is less than 0.1 μm on any side. **Table 11.1** compares the length of the chromosomes of several viruses to the size of their head structure. In each case, a similar packaging feat must be accomplished. Compare the dimensions given for phage T2 with the micrograph of both the DNA and the viral particle shown in **Figure 11–2**. Seldom does the space available in the head of a virus exceed the chromosome volume by more than a factor of two. In many cases, almost all of the space is filled, indicating nearly perfect packing. Once packed within the head, the genetic material is functionally inert until it is released into a host cell.

Bacterial chromosomes are also relatively simple in form. They generally consist of a double-stranded DNA molecule, compacted into a structure sometimes referred to as the **nucleoid.** *Escherichia coli,* the most extensively studied bacterium, has a large circular chromosome measuring approximately 1200 μm (1.2 mm) in length that may occupy up to one-third of the volume of the cell. When the cell is gently lysed and the chromosome is released, it can be visualized under the electron microscope (**Figure 11–3**).

DNA in bacterial chromosomes is found to be associated with several types of DNA-binding proteins. Two, called **HU** and **H-NS (Histone-like Nucleoid Structuring) proteins,** are small but abundant in the cell and contain a high percentage of positively charged amino acids that can bond ionically to the negative charges of the phosphate groups in DNA. These proteins function to fold and bend DNA. As such, coils are created that have the effect of compacting the DNA constituting the nucleoid. Additionally, H-NS proteins, like histones in eukaryotes, have been implicated in regulating gene activity in a nonspecific way.

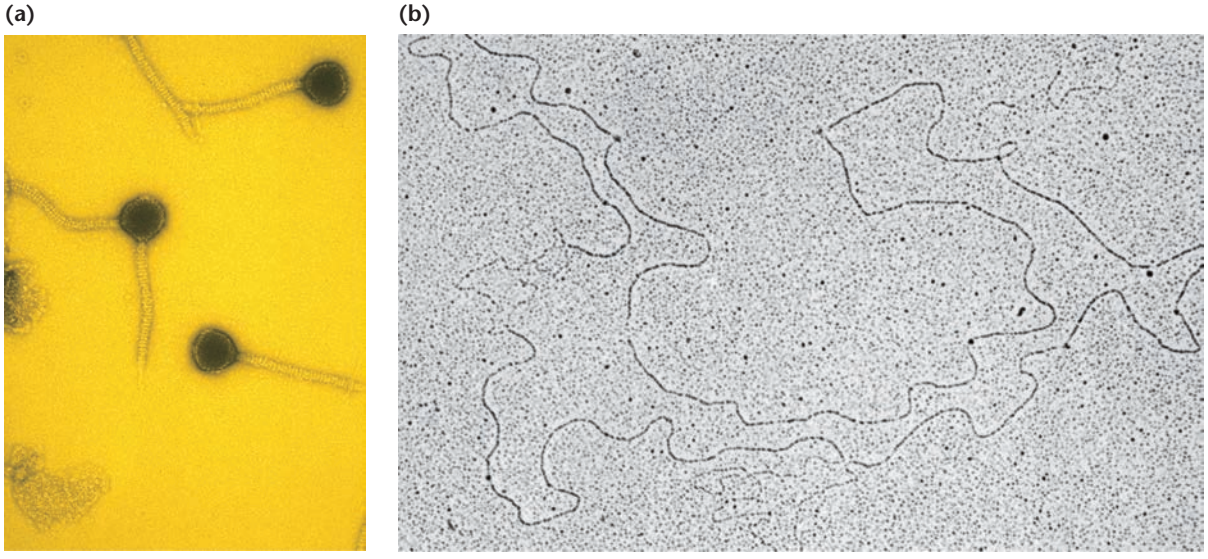


FIGURE 11–1 Electron micrographs of phage λ (a) and the DNA that was isolated from it (b). The chromosome is 17 μm long. Note that the phages are magnified about five times more than the DNA.

TABLE 11.1 The Genetic Material of Representative Viruses and Bacteria

Organism		Nucleic Acid			Overall Size of Viral Head or Bacteria (μm)
		Type	SS or DS*	Length (μm)	
Viruses	ϕX174	DNA	SS	2.0	0.025 × 0.025
	Tobacco mosaic virus	RNA	SS	3.3	0.30 × 0.02
	Phage λ	DNA	DS	17.0	0.07 × 0.07
	T2 phage	DNA	DS	52.0	0.07 × 0.10
Bacteria	*Haemophilus influenzae*	DNA	DS	832.0	1.00 × 0.30
	Escherichia coli	DNA	DS	1200.0	2.00 × 0.50

*SS = single-stranded, DS = double-stranded.

ESSENTIAL POINT

In contrast to eukaryotes, bacteriophage and bacterial chromosomes are largely devoid of associated proteins, are of much smaller size, and most often consist of circular DNA. ■

NOW SOLVE THIS

11–1 In bacteriophages and bacteria, the DNA is almost always organized into circular (closed loops) chromosomes. Phage λ is an exception, maintaining its DNA in a linear chromosome within the viral particle. However, as soon as this DNA is injected into a host cell, it circularizes before replication begins. What advantage exists in replicating circular DNA molecules compared to linear molecules, characteristic of eukaryotic chromosomes?

■ **HINT:** *This problem involves an understanding of eukaryotic DNA replication, as discussed in Chapter 10. The key to its solution is to consider why the enzyme telomerase is essential in eukaryotic DNA replication, and why bacterial and viral chromosomes can be replicated without encountering the "telomere problem."*

11.2 Mitochondria and Chloroplasts Contain DNA Similar to Bacteria and Viruses

That both **mitochondria** and **chloroplasts** contain their own DNA and a system for expressing genetic information was first suggested by the discovery of mutations and the resultant inheritance patterns in plants, yeast, and other fungi. Because both mitochondria and chloroplasts are inherited through the maternal cytoplasm in most organisms, and because each of the above-mentioned examples of mutations could be linked hypothetically to the altered function of either chloroplasts or mitochondria, geneticists set out to look for more direct evidence of DNA in these organelles. Not only was unique DNA found to be a normal component of both mitochondria and chloroplasts, but careful examination of the nature of this genetic information revealed a remarkable similarity to that found in viruses and bacteria.

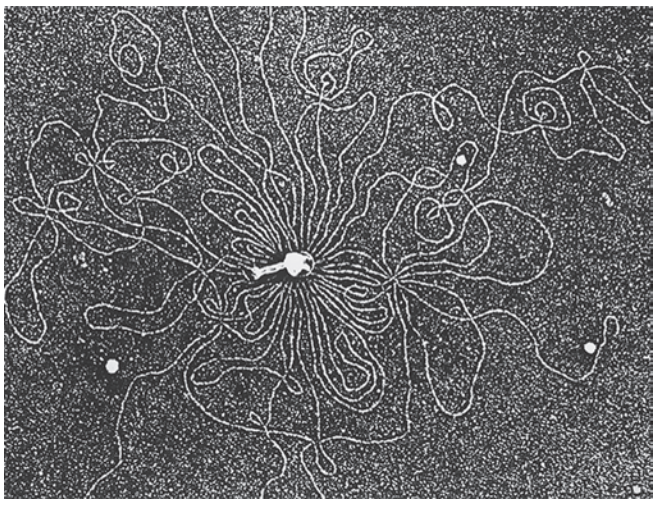


FIGURE 11–2 Electron micrograph of bacteriophage T2, which has had its DNA released by osmotic shock. The chromosome is 52 μm long.

FIGURE 11–3 Electron micrograph of the bacterium *E. coli*, which has had its DNA released by osmotic shock. The chromosome is 1200 μm long.

Molecular Organization and Gene Products of Mitochondrial DNA

Extensive information is also available concerning the structure and gene products of **mitochondrial DNA (mtDNA).** In most eukaryotes, mtDNA exists as a double-stranded, closed circle (**Figure 11–4**) that is free of the chromosomal proteins characteristic of eukaryotic chromosomal DNA. An exception is found in some ciliated protozoans, in which the DNA is linear.

In size, mtDNA varies greatly among organisms. In a variety of animals, including humans, mtDNA consists of about 16,000 to 18,000 bp (16 to 18 kb). However, yeast (*Saccharomyces*) mtDNA consists of 75 kb. Plants typically exceed this amount—367 kb is present in mitochondria in the mustard plant, *Arabidopsis*. Vertebrates have 5 to 10 such DNA molecules per organelle, whereas plants have 20 to 40 copies per organelle.

There are several other noteworthy aspects of mtDNA. With only rare exceptions, *introns* (noncoding regions within genes) are absent from mitochondrial genes, and gene repetitions are seldom present. Nor is there usually much in the way of intergenic spacer DNA. This is particularly true in species whose mtDNA is fairly small in size, such as humans. In *Saccharomyces*, with a much larger mtDNA molecule, introns and intergenic spacer DNA account for much of the excess DNA. As will be discussed in Chapter 12, the expression of mitochondrial genes uses several modifications of the otherwise standard genetic code. Also of interest is the fact that replication in mitochondria is dependent on enzymes encoded by nuclear DNA.

Another interesting observation is that in vertebrate mtDNA, the two strands vary in density, as revealed by centrifugation. This provides researchers with a way to isolate the strands for study, designating one heavy (H) and the other light (L). While most of the mitochondrial genes are encoded by the H strand, several are encoded by the complementary L strand.

Molecular Organization and Gene Products of Chloroplast DNA

Chloroplasts provide the photosynthetic function specific to plants. Like mitochondria, they contain an autonomous genetic system distinct from that found in the nucleus and cytoplasm, which has as its foundation a unique DNA molecule **(cpDNA). Chloroplast DNA,** shown in **Figure 11–5**, is fairly uniform in size among different organisms, ranging between 100 and 225 kb in length. It shares many similarities to DNA found in prokaryotic cells. It is circular and double-stranded, and it is free of the associated proteins characteristic of eukaryotic DNA.

The size of cpDNA is much larger than that of mtDNA. To some extent, this can be accounted for by a larger number of genes. However, most of the difference appears to be due to the presence in cpDNA of many long noncoding nucleotide sequences both between and within genes, the latter being called **introns.** Duplications of many DNA sequences are also present.

In the green alga *Chlamydomonas,* there are about 75 copies of the chloroplast DNA molecule per organelle. Each copy consists of a length of DNA that contains 195,000 base pairs (195 kb). In higher plants, such as the sweet pea, multiple copies of the DNA molecule are also present in each organelle, but the molecule is considerably smaller (134 kb) than that in *Chlamydomonas.*

ESSENTIAL POINT

Mitochondria and chloroplasts contain DNA that is remarkably similar in form and appearance to some bacterial and bacteriophage DNA. ■

FIGURE 11–4 Electron micrograph of mitochondrial DNA (mtDNA) derived from *Xenopus laevis.*

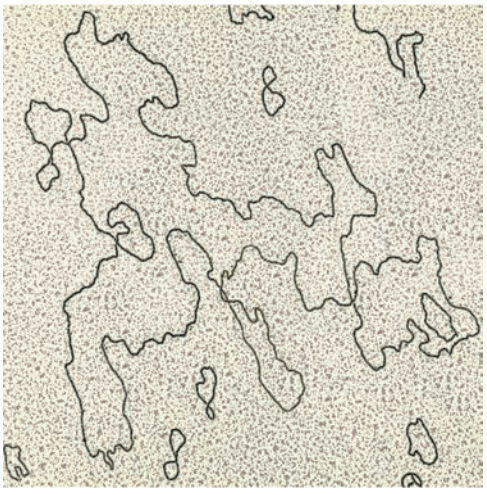


FIGURE 11–5 Electron micrograph of chloroplast DNA obtained from lettuce.

11.3 Specialized Chromosomes Reveal Variations in the Organization of DNA

We now consider two cases of genetic organization that demonstrate the specialized forms that eukaryotic chromosomes can take. Both types—*polytene chromosomes* and *lampbrush chromosomes*—are so large that their organization was discerned using light microscopy long before we understood how mitotic chromosomes form from interphase chromatin. The study of these chromosomes provided many of our initial insights into the arrangement and function of the genetic information. It is important to note that polytene and lampbrush chromosomes are unusual and not typically found in most eukaryotic cells, but the study of their structure has revealed many common themes of chromosome organization.

Polytene Chromosomes

Giant **polytene chromosomes** are found in various tissues (salivary, midgut, rectal, and malpighian excretory tubules) in the larvae of some flies and in several species of protozoans and plants. Such structures, first observed by E. G. Balbiani in 1881, provided a model system for subsequent investigations of chromosomes. What is particularly intriguing about polytene chromosomes is that they can be seen in the nuclei of interphase cells.

Each polytene chromosome is 200 to 600 μm long, and when they are observed under the light microscope, they reveal a linear series of alternating bands and interbands (**Figure 11–6**). The banding pattern is distinctive for each chromosome in any given species. Individual bands are sometimes called **chromomeres,** a generalized term describing lateral condensations of material along the axis of a chromosome.

Extensive study using electron microscopy and radioactive tracers led to an explanation for the unusual appearance of these chromosomes. First, polytene chromosomes represent paired homologs. This is highly unusual because they are present in somatic cells, where in most organisms, chromosomal material is normally dispersed as chromatin and homologs are not paired. Second, their large size and distinctiveness result from the many DNA strands that compose them. The DNA of these paired homologs undergoes many rounds of replication, *but without strand separation or cytoplasmic division.* As replication proceeds, chromosomes contain 1000 to 5000 DNA strands that remain in precise parallel alignment with one another, giving rise to the distinctive band pattern along the axis of the chromosome.

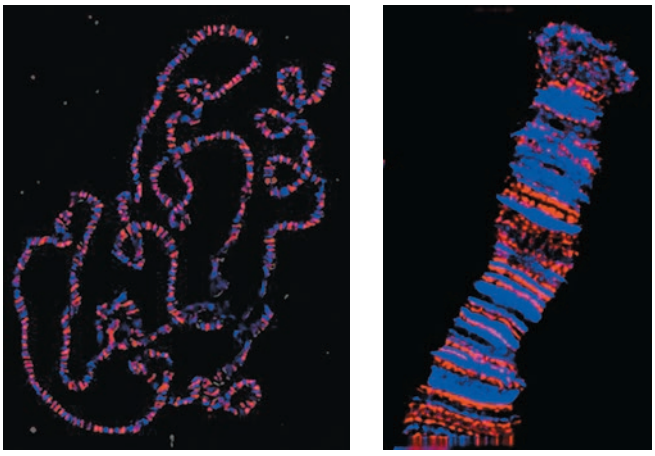


FIGURE 11–6 Polytene chromosomes derived from larval salivary gland cells of *Drosophila*.

The presence of bands on polytene chromosomes was initially interpreted as the visible manifestation of individual genes. The discovery that the strands present in bands undergo localized uncoiling during genetic activity further strengthened this view. Each such uncoiling event results in what is called a **puff** because of its appearance (**Figure 11–7**). That puffs are visible manifestations of gene activity (transcription that produces RNA) is evidenced by their high rate of incorporation of radioactively labeled RNA precursors, as assayed by autoradiography. Bands that are not extended into puffs incorporate fewer radioactive precursors or none at all.

The study of bands during development in insects such as *Drosophila* and the midge fly *Chironomus* reveals *differential gene activity*. A characteristic pattern of band formation that is equated with gene activation is observed as development

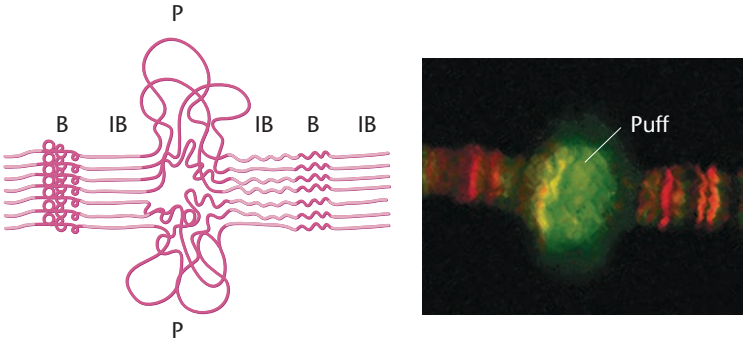


FIGURE 11–7 Photograph of a puff within a polytene chromosome. The diagram depicts the uncoiling of strands within a band (B) region to produce a puff (P) in polytene chromosomes. Interband regions (IB) are also labeled.

proceeds. Despite attempts to resolve the issue, it is not yet clear how many genes are contained in each band. However, we do know that in *Drosophila,* which contains about 15,000 genes, there are approximately 5000 bands. Interestingly, a band may contain up to 10^7 base pairs of DNA, enough to encode 50 to 100 average-size genes.

NOW SOLVE THIS

11–2 After salivary gland cells from *Drosophila* are isolated and cultured in the presence of radioactive thymidylic acid, autoradiography is performed, revealing the presence of thymidine within polytene chromosomes. Predict the distribution of the grains along the chromosomes.

■ **HINT:** *This problem involves an understanding of the organization of DNA in polytene chromosomes. The key to its solution is to be aware that ^{3}H-thymidine, as a molecular tracer, will only be incorporated into DNA during its replication.*

Lampbrush Chromosomes

Another specialized chromosome that has given us insight into chromosomal structure is the **lampbrush chromosome,** so named because it resembles the brushes used to clean kerosene-lamp chimneys in the nineteenth century. Lampbrush chromosomes were first studied in detail in 1892 in the oocytes of sharks and are now known to be characteristic of most vertebrate oocytes as well as the spermatocytes of some insects. Therefore, they are meiotic chromosomes. Most experimental work has been done with material taken from amphibian oocytes.

These unique chromosomes are easily isolated from oocytes in the first prophase stage of meiosis, where they are active in directing the metabolic activities of the developing cell. The homologs are seen as synapsed pairs held together by chiasmata. However, instead of condensing, as most meiotic chromosomes do, lampbrush chromosomes often extend to lengths of 500 to 800 μm. Later in meiosis, they revert to their normal length of 15 to 20 μm. Based on these observations, lampbrush chromosomes are interpreted as extended, uncoiled versions of the normal meiotic chromosomes.

The two views of lampbrush chromosomes in **Figure 11–8** provide significant insights into their morphology. Part (a) shows the meiotic configuration under the light microscope. The linear axis of each structure contains a large number of condensed areas, and as with polytene chromosomes, these are referred to as *chromomeres.* Emanating from each chromomere is a pair of lateral loops, which give the chromosome its distinctive appearance. In part (b), the scanning electron micrograph (SEM) reveals adjacent loops present along one of the two axes of the chromosome. As with bands in polytene chromosomes, much more DNA is present in each loop than is needed to encode a single gene. Such an SEM provides a clear view of the chromomeres and the chromosomal fibers emanating from them. Each chromosomal loop is thought to be composed of one DNA double helix, while the central axis is made up of two DNA helices. This hypothesis is consistent with the belief that each meiotic chromosome is composed of a pair of sister chromatids. Studies using radioactive RNA precursors reveal that the loops are active in the synthesis of RNA. The lampbrush loops, in a manner similar to puffs in polytene chromosomes, represent DNA that has been reeled out from the central chromomere axis during transcription.

ESSENTIAL POINT

Polytene and lampbrush chromosomes are examples of specialized structures that extended our knowledge of genetic organization and function well in advance of the technology available to the modern-day molecular biologist. ■

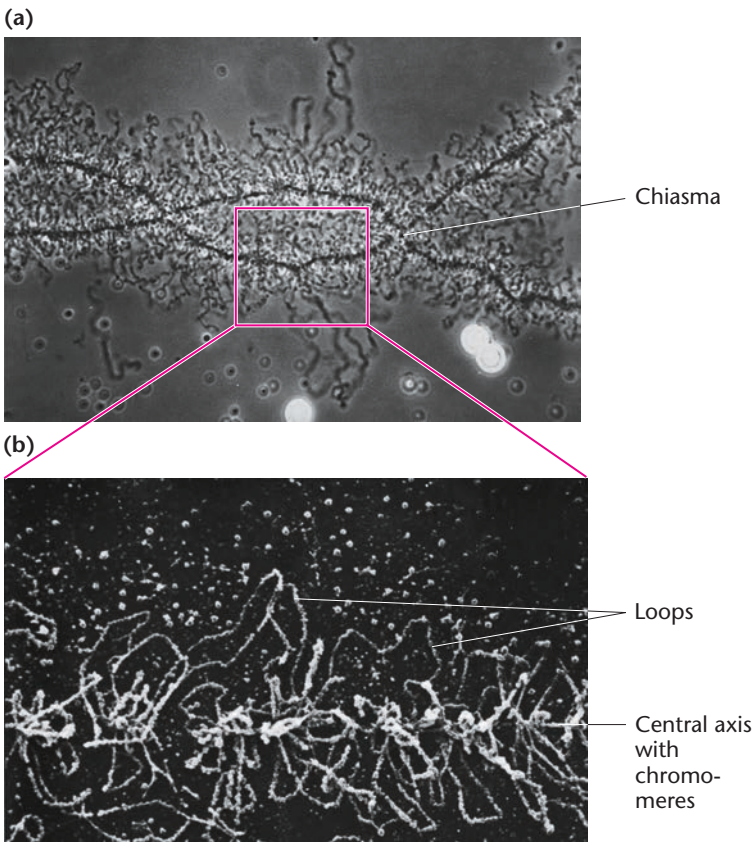


FIGURE 11–8 Lampbrush chromosomes derived from amphibian oocytes. Part (a) is a photomicrograph; part (b) is a scanning electron micrograph.

11.4 DNA Is Organized into Chromatin in Eukaryotes

We now turn our attention to the way DNA is organized in eukaryotic chromosomes. Our focus will be on eukaryotic cells, in which chromosomes are visible only during mitosis. After chromosome separation and cell division, cells enter the interphase stage of the cell cycle, during which time the components of the chromosome uncoil and are present in the form referred to as **chromatin.** While in interphase, the chromatin is dispersed in the nucleus, and the DNA of each chromosome is replicated. As the cell cycle progresses, most cells reenter mitosis, whereupon chromatin coils into visible chromosomes once again. This condensation represents a length contraction of some 10,000 times for each chromatin fiber.

The organization of DNA during the transitions just described is much more intricate and complex than in viruses or bacteria, which never exhibit a process similar to mitosis. This is due to the greater amount of DNA per chromosome, as well as the presence of a large number of proteins associated with eukaryotic DNA. For example, while DNA in the *E. coli* chromosome is 1200 μm long, the DNA in each human chromosome ranges from 19,000 to 73,000 μm in length. In a single human nucleus, all 46 chromosomes contain sufficient DNA to extend to more than 2 meters. This genetic material, along with its associated proteins, is contained within a nucleus that usually measures about 5 to 10 μm in diameter.

Chromatin Structure and Nucleosomes

As we have seen, the genetic material of viruses and bacteria consists of strands of DNA or RNA that are nearly devoid of proteins. In eukaryotic chromatin, a substantial amount of protein is associated with the chromosomal DNA in all phases of the eukaryotic cell cycle. The associated proteins are divided into basic, positively charged **histones** and less positively charged nonhistones. The histones clearly play the most essential structural role of all the proteins associated with DNA. There are five types, and they all contain large amounts of the positively charged amino acids lysine and arginine. This makes it possible for them to bond electrostatically to the negatively charged phosphate groups of nucleotides in DNA. Recall that a similar interaction has been proposed for several bacterial proteins.

The general model for chromatin structure is based on the assumption that chromatin fibers, composed of DNA and protein, undergo extensive coiling and folding as they are condensed within the cell nucleus. X-ray diffraction studies confirm that histones play an important role in chromatin structure. Chromatin produces regularly spaced diffraction rings, suggesting that repeating structural units occur along the chromatin axis. If the histone molecules are chemically removed from chromatin, the regularity of this diffraction pattern is disrupted.

A basic model for chromatin structure was worked out in the mid-1970s. Several observations were particularly relevant to the development of this model:

1. Digestion of chromatin by certain endonucleases, such as micrococcal nuclease, yields DNA fragments that are approximately 200 bp in length or multiples thereof. This demonstrates that enzymatic digestion is not random, for if it were, we would expect a wide range of fragment sizes. Thus, chromatin consists of some type of repeating unit, each of which is protected from enzymatic cleavage, except where any two units are joined. It is the area between units that is attacked and cleaved by the endonuclease.

2. Electron microscopic observations of chromatin reveal that chromatin fibers are composed of linear arrays of spherical particles (**Figure 11–9**). Discovered by Ada and Donald Olins, the particles occur regularly along the axis of a chromatin strand and resemble beads on a string. This conforms nicely to the earlier observation, which suggests the existence of repeating units. These particles, initially referred to as ν-bodies (ν is the Greek letter nu), are now called **nucleosomes.**

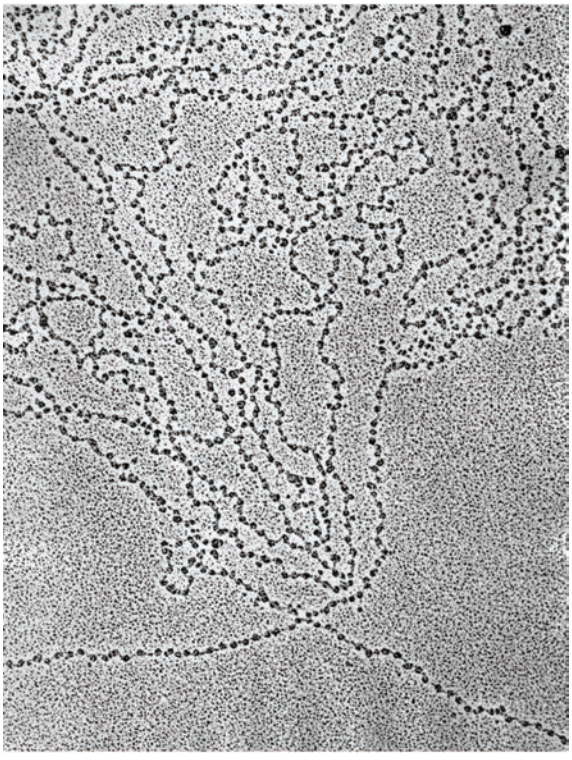


FIGURE 11–9 An electron micrograph revealing nucleosomes appearing as "beads on a string" along chromatin strands derived from *Drosophila melanogaster.*

3. Studies of precise interactions of histone molecules and DNA in the nucleosomes constituting chromatin show that histones H2A, H2B, H3, and H4 occur as two types of tetramers, $(H2A)_2 \cdot (H2B)_2$ and $(H3)_2 \cdot (H4)_2$. Roger Kornberg predicted that each repeating nucleosome unit consists of one of each tetramer (creating an octamer) in association with about 200 bp of DNA. Such a structure is consistent with previous observations and provides the basis for a model that explains the interaction of histones and DNA in chromatin.
4. When nuclease digestion time is extended, some of the 200 bp of DNA are removed from the nucleosome, creating a **nucleosome core particle** consisting of 147 bp. The DNA lost in this prolonged digestion is responsible for linking nucleosomes together. This linker DNA is associated with the fifth histone, H1.

On the basis of this information, as well as on X-ray and neutron-scattering analyses of crystallized core particles by John T. Finch, Aaron Klug, and others, a detailed model of the nucleosome was put forward in 1984, providing a basis for predicting chromatin structure and its condensation into chromosomes. In this model, illustrated in **Figure 11–10**, a 147-bp length of the 2-nm-diameter DNA molecule coils around an octamer of histones in a left-handed superhelix that completes about 1.7 turns per nucleosome. Each nucleosome, ellipsoidal in shape, measures about 11 nm at its longest point [**Figure 11–10(a)**]. Significantly, the formation of the nucleosome represents the first level of packing, whereby the DNA helix is reduced to about one-third of its original length by winding around the histones.

In the nucleus, the chromatin fiber seldom, if ever, exists in the extended form described in the previous paragraph (that is, as an extended chain of nucleosomes). Instead,

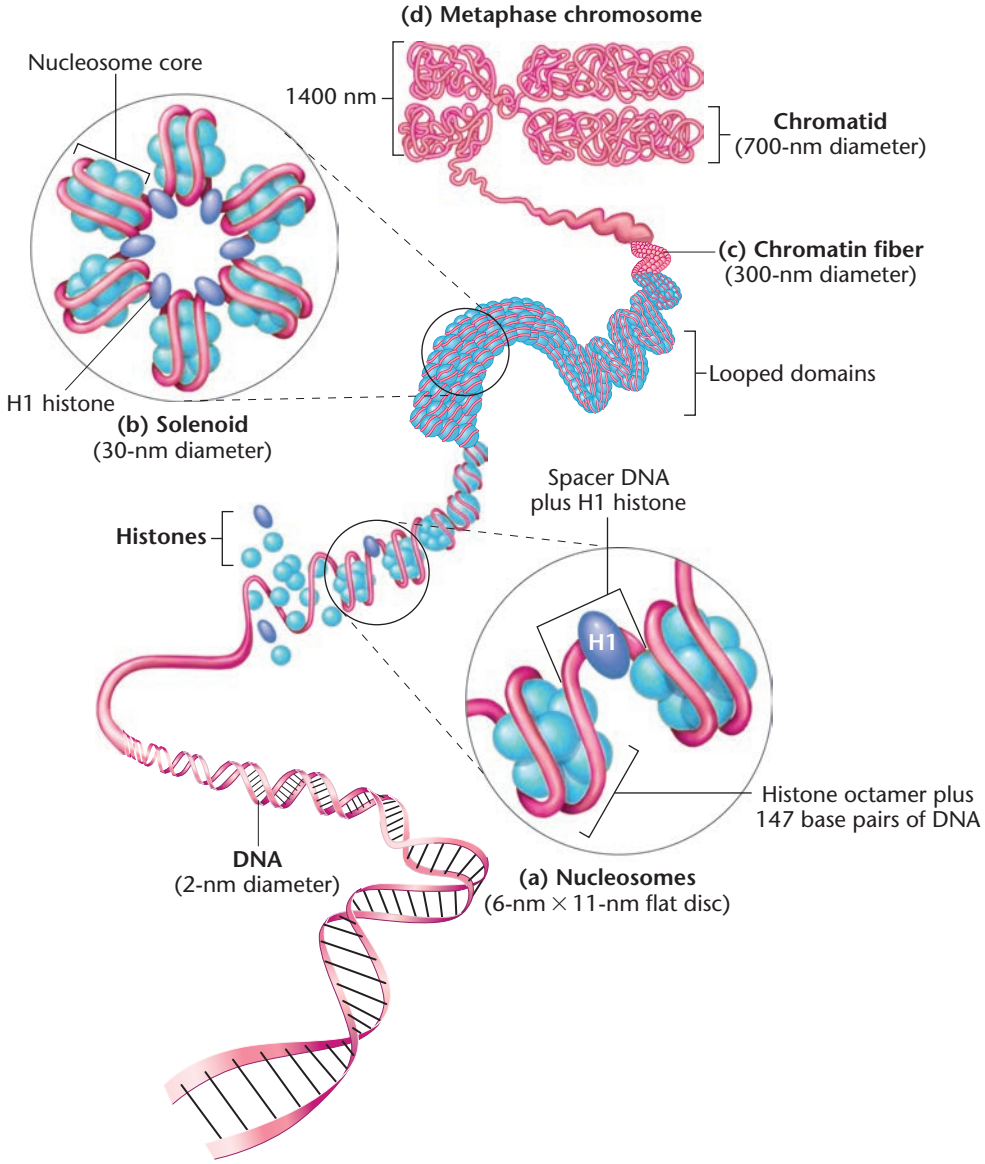


FIGURE 11–10 General model of the association of histones and DNA to form nucleosomes, illustrating the way in which each thickness of fiber may be coiled into a more condensed structure, ultimately producing a metaphase chromosome.

the 11-nm-diameter fiber is further packed into a thicker, 30-nm-diameter structure that was initially called a *solenoid* [**Figure 11–10(b)**]. This thicker structure, which is dependent on the presence of histone H1, consists of numerous nucleosomes coiled around and stacked upon one another, creating a second level of packing. This provides a six-fold increase in compaction of the DNA. It is this structure that is characteristic of an uncoiled chromatin fiber in interphase of the cell cycle. In the transition to the mitotic chromosome, still further compaction must occur. The 30-nm structures are folded into a series of *looped domains,* which further condense the chromatin fiber into a structure that is 300 nm in diameter [**Figure 11–10(c)**]. These *coiled chromatin fibers* are then compacted into the chromosome arms that constitute a chromatid, one of the longitudinal subunits of the metaphase chromosome [**Figure 11–10(d)**]. While Figure 11–10 shows the chromatid arms to be 700 nm in diameter, this value undoubtedly varies among different organisms. At a value of 700 nm, a pair of sister chromatids comprising a chromosome measures about 1400 nm.

The importance of the organization of DNA into chromatin and chromatin into mitotic chromosomes can be illustrated by considering a human cell that stores its genetic material in a nucleus that is about 5 to 10 μm in diameter. The haploid genome contains 3.2×10^9 base pairs of DNA distributed among 23 chromosomes. The diploid cell contains twice that amount. At 0.34 nm per base pair, this amounts to an enormous length of DNA (as stated earlier, to more than 2 m). One estimate is that the DNA inside a typical human nucleus is complexed with roughly 2.5×10^7 nucleosomes.

In the overall transition from a fully extended DNA helix to the extremely condensed status of the mitotic chromosome, a packing ratio (the ratio of DNA length to the length of the structure containing it) of about 500 to 1 must be achieved. In fact, our model accounts for a ratio of only about 50 to 1. Obviously, the larger fiber can be further bent, coiled, and packed to achieve even greater condensation during the formation of a mitotic chromosome.

ESSENTIAL POINT

Eukaryotic chromatin is a nucleoprotein organized into repeating units called nucleosomes, which are composed of 200 base pairs of DNA, an octamer of four types of histones, plus one linker histone. ■

Chromatin Remodeling

As with many significant findings in genetics, the study of nucleosomes has answered some important questions, but at the same time it has also led us to new ones. For example, in the preceding discussion, we established that histone proteins play an important structural role in packaging DNA into the nucleosomes that make up chromatin. While solving the structural problem of how to organize a huge amount of DNA within the eukaryotic nucleus, a new problem was apparent: *the chromatin fiber, when complexed with histones and folded into various levels of compaction, makes the DNA inaccessible to interaction with important nonhistone proteins.* For example, the proteins that function in enzymatic and regulatory roles during the processes of replication and gene expression must interact directly with DNA. To accommodate these protein–DNA interactions, chromatin must be induced to change its structure, a process called **chromatin remodeling.** In the case of replication and gene expression, chromatin must relax its compact structure but be able to reverse the process during periods of inactivity.

Insights into how different states of chromatin structure may be achieved were forthcoming in 1997, when Timothy Richmond and members of his research team were able to significantly improve the level of resolution in X-ray diffraction studies of nucleosome crystals (from 7 Å in the 1984 studies to 2.8 Å in the 1997 studies). At this resolution, most atoms are visible, thus revealing the subtle twists and turns of the superhelix of DNA that encircles the histones. Recall that the double-helical ribbon represents 147 bp of DNA surrounding four pairs of histone proteins. This configuration is repeated over and over in the chromatin fiber and is the principal packaging unit of DNA in the eukaryotic nucleus.

The work of Richmond and colleagues, extended to a resolution of 1.9 Å in 2003, has revealed the details of the location of each histone entity within the nucleosome. Of particular interest to chromatin remodeling is that unstructured **histone tails** are not packed into the folded histone domains within the core of the nucleosome. For example, tails devoid of any secondary structure extending from histones H3 and H2B protrude through the minor groove channels of the DNA helix. The tails of histone H4 appear to make a connection with adjacent nucleosomes. Histone tails also provide potential targets for a variety of chemical modifications that may be linked to genetic functions along the chromatin fiber, including the regulation of gene expression.

NOW SOLVE THIS

11–3 If a human nucleus is 10 μm in diameter, and it must hold as much as 2 m of DNA, which is complexed into nucleosomes that during full extension are 11 nm in diameter, what percentage of the volume of the nucleus does the genetic material occupy?

■ **HINT:** *This problem asks you to make some numerical calculations in order to see just how "filled" the eukaryotic nucleus is with a diploid amount of DNA. The key to its solution is the use of the formula $V = (4/3)\pi r^3$, which calculates the volume of a sphere.*

Several of these potential chemical modifications are now recognized as important to genetic function. One of the most well-studied histone modifications involves **acetylation** by the action of the enzyme *histone acetyltransferase (HAT).* The addition of an acetyl group to the positively charged amino group present on the side chain of the amino acid lysine effectively changes the net charge of the protein by neutralizing the positive charge. Lysine is in abundance in histones, and geneticists have known for some time that acetylation is linked to gene activation. It appears that high levels of acetylation open up, or remodel, the chromatin fiber, an effect that increases in regions of active genes and decreases in inactive regions. In a well-studied example, the inactivation of the X chromosome in mammals, forming a Barr body (Chapter 5), histone H4 is known to be greatly underacetylated.

Two other important chemical modifications are the **methylation** and **phosphorylation** of amino acids that are part of histones. These chemical processes result from the action of enzymes called *methyltransferases* and *kinases,* respectively. Methyl groups may be added to both arginine and lysine residues of histones, and these changes can either increase or decrease transcription depending on which amino acids are methylated. Phosphate groups can be added to the hydroxyl groups of the amino acids serine and histidine, introducing a negative charge on the protein. During the cell cycle, increased phosphorylation, particularly of histone H3, is known to occur at characteristic times. Such chemical modification is believed to be related to the cycle of chromatin unfolding and condensation that occurs during and after DNA replication. It is important to note that the above chemical modifications (acetylation, methylation, and phosphorylation) are all reversible, under the direction of specific enzymes.

Not to be confused with histone methylation, the nitrogenous base cytosine within the DNA itself can also be methylated, forming 5-methyl cytosine. Cytosine methylation is usually negatively correlated with gene activity and occurs most often when the nucleotide cytidylic acid is next to the nucleotide guanylic acid, forming what is called a **CpG island.**

The research described above has extended our knowledge of nucleosomes and chromatin organization and serves here as a general introduction to the concept of chromatin remodeling. A great deal more work must be done, however, to elucidate the specific involvement of chromatin remodeling during genetic processes. In particular, the way in which the modifications are influenced by regulatory molecules within cells will provide important insights into the mechanisms of gene expression. What is clear is that the dynamic forms in which chromatin exists are vitally important to the way that all genetic processes directly involving DNA are executed. We will return to a more detailed discussion of the role of chromatin remodeling when we consider the regulation of eukaryotic gene expression later in the text (see Chapter 15). In addition, chromatin remodeling is an important topic in the discussion of **epigenetics,** the study of modifications of an organism's genetic and phenotypic expression that are *not* attributable to alteration of the DNA sequence making up a gene. This topic is discussed in depth in a future chapter (Special Topic Chapter 1—Epigenetics).

Heterochromatin

Although we know that the DNA of the eukaryotic chromosome consists of one continuous double-helical fiber along its entire length, we also know that the whole chromosome is not structurally uniform from end to end. In the early part of the twentieth century, it was observed that some parts of the chromosome remain condensed and stain deeply during interphase, while most parts are uncoiled and do not stain. In 1928, the terms **euchromatin** and **heterochromatin** were coined to describe the parts of chromosomes that are uncoiled and those that remain condensed, respectively.

Subsequent investigation revealed a number of characteristics that distinguish heterochromatin from euchromatin. Heterochromatic areas are genetically inactive because they either lack genes or contain genes that are repressed. Also, heterochromatin replicates later during the S phase of the cell cycle than euchromatin does. The discovery of heterochromatin provided the first clues that parts of eukaryotic chromosomes do not always encode proteins. Instead, some chromosome regions are thought to be involved in maintenance of the chromosome's structural integrity and in other functions, such as chromosome movement during cell division.

The presence of heterochromatin is unique to and characteristic of the genetic material of eukaryotes. In some cases, whole chromosomes are heterochromatic. A case in point is the mammalian Y chromosome, much of which is genetically inert. And, as we discussed in Chapter 5, the inactivated X chromosome in mammalian females is condensed into an inert heterochromatic Barr body. In some species, such as mealy bugs, all chromosomes of one entire haploid set are heterochromatic.

When certain heterochromatic areas from one chromosome are translocated to a new site on the same or another nonhomologous chromosome, genetically active areas sometimes become genetically inert if they lie adjacent to the translocated heterochromatin. This influence on existing euchromatin is one example of what is more generally referred to as a **position effect.** That is, the position of a gene or group of genes relative to all other genetic material may affect their expression.

Chromosome Banding

Until about 1970, mitotic chromosomes viewed under the light microscope could be distinguished only by their

relative sizes and the positions of their centromeres. In karyotypes, two or more chromosomes are often visually indistinguishable from one another. Numerous cytological procedures, referred to as **chromosome banding,** have now made it possible to distinguish such chromosomes from one another as a result of differential staining along the longitudinal axis of mitotic chromosomes.

The most useful of these techniques, called **G-banding** (see Figure 6.2), involves the digestion of the mitotic chromosomes with the proteolytic enzyme trypsin, followed by Giemsa staining. This procedure stains regions of DNA that are rich in A═T base pairs. Another technique, called **C-banding,** uses chromosome preparations that are heat denatured. Subsequent Giemsa staining reveals only the heterochromatic regions of the centromeres.

These, and other chromosome-banding techniques, reflect the heterogeneity and complexity of the chromosome along its length. So precise is the banding pattern that when a segment of one chromosome has been translocated to another chromosome, its origin can be determined with great precision.

ESSENTIAL POINT

Heterochromatin, prematurely condensed in interphase and for the most part genetically inert, is illustrated by the centromeric and telomeric regions of eukaryotic chromosomes, the Y chromosome, and the Barr body. ■

11.5 Eukaryotic Genomes Demonstrate Complex Sequence Organization Characterized by Repetitive DNA

Thus far, we have looked at how DNA is organized into chromosomes in bacteriophages, bacteria, and eukaryotes. We now begin an examination of what we know about the organization of DNA sequences within the chromosomes making up an organism's genome, placing our emphasis on eukaryotes. Once we establish the pattern of genome organization, in a later chapter we will focus on how the genes themselves are organized within chromosomes (see Chapter 18).

We learned in Chapter 9 that, in addition to single copies of unique DNA sequences that comprise genes, a great deal of the DNA sequences within chromosomes are repetitive in nature and that various levels of repetition occur within the genome of organisms. Many studies have now provided insights into **repetitive DNA,** demonstrating various classes of these sequences and their organization within the genome. **Figure 11–11** outlines the various categories of repetitive DNA. Some functional genes are present in more than one copy and are therefore repetitive in nature. However, the majority of repetitive sequences are nongenic, and in fact, most serve no known function. We explore three main categories: (1) heterochromatin found associated with centromeres and making up telomeres, (2) tandem repeats of both short and long DNA sequences, and (3) transposable sequences that are interspersed throughout the genome of eukaryotes.

Repetitive DNA and Satellite DNA

The nucleotide composition of the DNA (e.g., the percentage of G≡C versus A═T pairs) of a particular species is reflected in its density, which can be measured with sedimentation equilibrium centrifugation. When eukaryotic DNA is analyzed in this way, the majority is present as a single main band, or peak, of fairly uniform density. However, one or more additional peaks represent DNA that differs slightly in density. This component, called **satellite DNA,** represents a variable proportion of the total DNA, depending on the species. A profile of main-band and satellite DNA from the mouse is shown in **Figure 11–12**. By contrast, prokaryotes contain only main-band DNA.

The significance of satellite DNA remained an enigma until the mid-1960s, when Roy Britten and David Kohne developed the technique for measuring the reassociation kinetics of DNA that had previously been dissociated into single strands. They demonstrated that certain portions of DNA reassociated more rapidly than others (see Chapter 9).

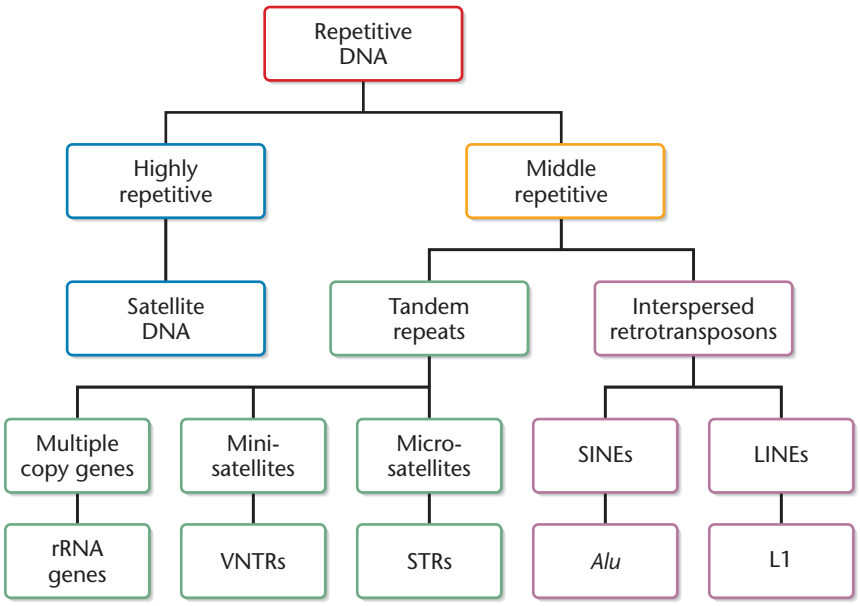


FIGURE 11–11 An overview of the categories of repetitive DNA.

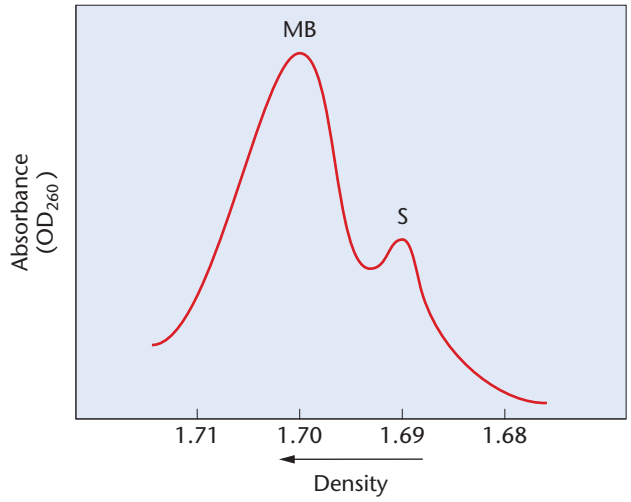


FIGURE 11–12 Separation of main-band (MB) and satellite (S) DNA from the mouse, using ultracentrifugation in a CsCl gradient.

They concluded that rapid reassociation was characteristic of multiple DNA fragments composed of identical or nearly identical nucleotide sequences—the basis for the descriptive term repetitive DNA.

When satellite DNA is subjected to analysis by reassociation kinetics, it falls into the category of **highly repetitive DNA,** which is known to consist of relatively short sequences repeated a large number of times. Further evidence suggests that these sequences are present as tandem repeats clustered in very specific chromosomal areas known to be heterochromatic—the regions flanking centromeres. This was discovered in 1969 when several researchers, including Mary Lou Pardue and Joe Gall, applied ***in situ* molecular hybridization** to the study of satellite DNA. This technique involves the molecular hybridization between an isolated fraction of radioactively labeled DNA or RNA probes and the DNA contained in the chromosomes of a cytological preparation. Following the hybridization procedure, autoradiography is performed to locate the chromosome areas complementary to the fraction of DNA or RNA.

Pardue and Gall demonstrated that radioactive probes made from mouse satellite DNA hybridize with the DNA of centromeric regions of mouse mitotic chromosomes, which are all telocentric (**Figure 11–13**). Several conclusions were drawn: Satellite DNA differs from main-band DNA in its molecular composition, as established by buoyant density studies. It is composed of repetitive sequences. Finally, satellite DNA is found in the heterochromatic centromeric regions of chromosomes.

Centromeric DNA Sequences

The separation of homologs during mitosis and meiosis depends on **centromeres,** described cytologically as the *primary constrictions* along eukaryotic chromosomes (see Chapter 2). In this role, it is believed that the DNA sequence contained within the centromere is critical. Careful analysis has confirmed this prediction. The minimal region of the centromere that supports the function of chromosomal segregation is designated the **CEN region.** Within this heterochromatic region of the chromosome, the DNA binds a platform of proteins, which in multicellular organisms includes the **kinetochore** that binds to the spindle fiber during division (see Figure 2–8).

The CEN regions of the yeast *Saccharomyces cerevisiae* were the first to be studied. Each centromere serves an identical function, so it is not surprising that CENs from different chromosomes were found to be remarkably similar in their organization. The CEN region of yeast chromosomes consists of about 120 bp. Mutational analysis suggests that portions near the 3′-end of this DNA region are most critical to centromere function since mutations in them, but not those nearer the 5′-end, disrupt centromere function. Thus, the DNA of this region appears to be essential to the eventual binding to the spindle fiber.

Centromere sequences of multicellular eukaryotes are much more extensive than in yeast and vary considerably in size. For example, in *Drosophila* the CEN region is found within some 200 to 600 kb of DNA, much of which is highly repetitive. Recall from our prior discussion that highly repetitive satellite DNA is localized in the centromere regions of mice. In humans, one of the most recognized satellite DNA sequences is the **alphoid family,** found mainly in the centromere regions. Alphoid sequences, each about 170 bp in length, are present in tandem arrays of up to 1 million base pairs. Embedded within this repetitive DNA are more specific sequences that are critical to centromere function.

FIGURE 11–13 *In situ* molecular hybridization between RNA transcribed from mouse satellite DNA and mitotic chromosomes. The grains in the autoradiograph localize the chromosome regions (the centromeres) containing satellite DNA sequences.

One final observation of interest is that the H3 histone, a normal part of most all eukaryotic nucleosomes, is substituted by a variant histone designated CENP-A in centromeric heterochromatin. It is believed that the N-terminal protein tails that make CEN-P unique are involved in the binding of kinetechore proteins that are essential to the microtubules of spindle fibers. This finding supports the supposition that the DNA sequence found only in centromeres is related to the function of this unique chromosomal structure.

Middle Repetitive Sequences: VNTRs and STRs

A brief look at still another prominent category of repetitive DNA sheds additional light on the organization of the eukaryotic genome. In addition to highly repetitive DNA, which constitutes about 5 percent of the human genome (and 10 percent of the mouse genome), a second category, **middle** (or **moderately**) **repetitive DNA,** is fairly well characterized. Because we now know a great deal about the human genome, we will use our own species to illustrate this category of DNA in genome organization.

Although middle repetitive DNA does include some duplicated genes (such as those encoding ribosomal RNA), most prominent in this category are either noncoding tandemly repeated sequences or noncoding interspersed sequences. No function has been ascribed to these components of the genome. An example is DNA described as **variable number tandem repeats (VNTRs).** These repeating DNA sequences may be 15 to 100 bp long and are found within and between genes. Many such clusters are dispersed throughout the genome and are often referred to as **minisatellites.**

The number of tandem copies of each specific sequence at each location varies from one individual to the next, creating localized regions of 1000 to 20,000 bp (1–20 kb) in length. As we will see in Special Topics Chapter 3—DNA Forensics, the variation in size (length) of these regions between individual humans was originally the basis for the forensic technique referred to as **DNA fingerprinting.**

Another group of tandemly repeated sequences consists of di-, tri-, tetra-, and pentanucleotides, also referred to as **microsatellites** or **short tandem repeats (STRs).** Like VNTRs, they are dispersed throughout the genome and vary among individuals in the number of repeats present at any site. For example, in humans, the most common microsatellite is the dinucleotide $(CA)_n$, where n equals the number of repeats. Most commonly, n is between 5 and 50. These clusters have served as useful molecular markers for genome analysis.

Repetitive Transposed Sequences: SINEs and LINEs

Still another category of repetitive DNA consists of sequences that are interspersed individually throughout the genome, rather than being tandemly repeated. They can be either short or long, and many have the added distinction of being similar to **transposable sequences,** which are mobile and can potentially relocate within the genome. A large portion of the human genome is composed of such sequences.

For example, **short interspersed elements,** called **SINEs,** are less than 500 base pairs long and may be present 500,000 times or more in the human genome. The best characterized human SINE is a set of closely related sequences called the ***Alu* family** (the name is based on the presence of DNA sequences recognized by the restriction endonuclease *Alu*I). Members of this DNA family, also found in other mammals, are 200 to 300 base pairs long and are dispersed rather uniformly throughout the genome, both between and within genes. In humans, this family encompasses more than 5 percent of the entire genome.

Alu sequences are of particular interest because some members of the *Alu* family are transcribed into RNA, although the specific role of this RNA is not certain. Even so, the consequence of *Alu* sequences is their potential for transposition within the genome, which is related to chromosome rearrangements during evolution. *Alu* sequences are thought to have arisen from an RNA element whose DNA complement was dispersed throughout the genome as a result of the activity of reverse transcriptase (an enzyme that synthesizes DNA on an RNA template).

The group of **long interspersed elements (LINEs)** represents yet another category of repetitive transposable DNA sequences. LINEs are usually about 6 kb in length and in the human genome are present approximately 850,000 times. The most prominent example in humans is the **L1 family.** Members of this sequence family are about 6400 base pairs long and are present up to 500,000 times. Their 5′-end is highly variable, and their role within the genome has yet to be defined.

The general mechanism for transposition of L1 elements is now clear. The L1 DNA sequence is first transcribed into an RNA molecule. The RNA then serves as the template for the synthesis of the DNA complement using the enzyme *reverse transcriptase.* This enzyme is encoded by a portion of the L1 sequence. The new L1 copy then integrates into the DNA of the chromosome at a new site. Because of the similarity of this transposition mechanism to that used by retroviruses, LINEs are referred to as **retrotransposons.**

SINEs and LINEs represent a significant portion of human DNA. SINEs constitute about 13 percent of the human genome, whereas LINEs constitute up to 21 percent. Within both types of elements, repeating sequences of DNA are present in combination with unique sequences.

Middle Repetitive Multiple-Copy Genes

In some cases, middle repetitive DNA includes functional genes present tandemly in multiple copies. For example,

many copies exist of the genes encoding ribosomal RNA. *Drosophila* has 120 copies per haploid genome. Single genetic units encode a large precursor molecule that is processed into the 5.8*S*, 18*S*, and 28*S* rRNA components. In humans, multiple copies of this gene are clustered on the p arm of the acrocentric chromosomes 13, 14, 15, 21, and 22. Multiple copies of the genes encoding 5*S* rRNA are transcribed separately from multiple clusters found together on the terminal portion of the p arm of chromosome 1.

ESSENTIAL POINT

Eukaryotic genomes demonstrate complex sequence organization characterized by numerous categories of repetitive DNA, consisting of either tandem repeats clustered in various regions of the genome or single sequences repeatedly interspersed at random throughout the genome. ■

11.6 The Vast Majority of a Eukaryotic Genome Does Not Encode Functional Genes

Given the preceding information concerning various forms of repetitive DNA in eukaryotes, it is of interest to pose an important question: *What proportion of the eukaryotic genome actually encodes functional genes?*

We have seen that, taken together, the various forms of highly repetitive and moderately repetitive DNA comprise a substantial portion of the human genome. In addition to repetitive DNA, a large amount of the DNA consists of single-copy sequences as defined by reassociation kinetic analysis (Chapter 9) that appear to be noncoding. Included are many instances of what we call **pseudogenes.** These are DNA sequences representing evolutionary vestiges of duplicated copies of genes that have undergone significant mutational alteration. As a result, although they show some homology to their parent gene, they are usually not transcribed because of insertions and deletions throughout their structure.

Although the proportion of the genome consisting of repetitive DNA varies among organisms, one feature seems to be shared: *Only a very small part of the genome actually codes for proteins.* For example, the 20,000 to 30,000 genes encoding proteins in the sea urchin occupy less than 10 percent of the genome. In *Drosophila,* only 5 to 10 percent of the genome is occupied by genes coding for proteins. In humans, it appears that the estimated 20,000 protein-coding genes genes occupy less than 2 percent of the total DNA sequence making up the genome.

Related to the above observation, we are currently discovering many cases where DNA sequences are transcribed into RNA molecules that are not translated into proteins, and that play important cellular roles, such as regulation of genetic activity. This topic is explored in more depth later in the text; see Special Topics Chapter 2—Emerging Roles of RNA.

Database of Genomic Variants: Structural Variations in the Human Genome

MasteringGenetics™ Visit the Study Area: Exploring Genomics

In this chapter, we focused on structural details of chromosomes and DNA sequence organization in chromosomes. A related finding is that large segments of DNA and a number of genes can vary greatly in copy number due to duplications, creating ***copy number variations (CNVs).*** Many studies are underway to identify and map CNVs and to find possible disease conditions associated with them.

Several thousand CNVs have been identified in the human genome, and estimates suggest there may be thousands more within human populations.

In this Exploring Genomics exercise we will visit the **Database of Genomic Variants (DGV)**, which provides a quickly expanding summary of structural variations in the human genome including CNVs.

- **Exercise I – Database of Genomic Variants**

1. Access the DGV at http://dgv.tcag.ca/dgv/app/home. Click the "About the Project" tab to learn more about the purpose of the DGV.
2. Information in the DGV is easily viewed by clicking on a chromosome of interest using the "Find DGV Variants by Chromosome" feature. Using this feature, click on a chromosome of interest to you. A table will appear under the "Variants" tab showing several columns of data including:

- Start and Stop: Shows the locus for the CNV, including the base pairs that span the variation.
- Variant Accession: Provides a unique identifying number for each variation. Click on the variant accession number to reveal a separate page of specific details about each CNV, including the chromosomal

banding location for the variation and known genes that are located in the CNV.

- Variant Type: Most variations in this database are CNVs. Variant subtypes, such as deletions or insertions, and duplications, are shown in an adjacent column.

3. Let's analyze a particular group of CNVs. Many CNVs are unlikely to affect phenotype because they involve large areas of non–protein-coding or nonregulatory sequences. But gene-containing CNVs have been identified, including variants containing genes associated with Alzheimer's disease, Parkinson's disease, and other conditions.

   Defensin (*DEF*) genes are part of a large family of highly duplicated genes. To learn more about *DEF* genes and CNVs, use the Keyword Search box and search for *DEF*. A results page for the search will appear with a listing of relevant CNVs. Click on the name for any of the different *DEF* genes listed, which will take you to a wealth of information (including links to Online Mendelian Inheritance in Man, OMIM) about these genes so that you can answer the following questions. Do this for several *DEF*-containing CNVs on different chromosomes to find the information you will need for your answers.

   a. On what chromosome(s) did you find CNVs containing *DEF* genes?

   b. What did you learn about the function of *DEF* gene products? What do DEF proteins do?

   c. Variations in *DEF* genotypes and *DEF* gene expression in humans have been implicated in a number of different human disease conditions. Give examples of the kinds of disorders affected by variations in *DEF* genotypes.

   d. Explore the DGV to search a chromosome of interest to you and learn more about CNVs that have been mapped to that chromosome. Try the Genome Browser feature that will show you maps of each chromosome indicating different variations. For CNVs (shown in blue), clicking on the CNV will take you to its locus on the chromosome.

CASE STUDY | Art inspires learning

A genetics student visiting a museum saw a painting by Goya showing a woman with a newborn baby in her lap that had very short arms and legs along with some facial abnormalities. Wondering whether this condition might be a genetic disorder, the student went online, learning that the baby might have Roberts syndrome (RBS), a rare autosomal recessive trait. She read that cells in RBS have mitotic errors, including the premature separation of centromeres and other heterochromatic regions of homologs in metaphase instead of anaphase. As a result, metaphase chromosomes have a rigid or "railroad track" appearance. RBS has been shown to be caused by mutant alleles of the *ESCO2* gene, which functions during cell division.

The student wrote a list of questions to investigate in an attempt to better understand this condition. How would you answer these questions?

1. What do centromeres and other heterochromatic regions have in common that might cause this appearance?
2. What might be the role of the protein encoded by *ESCO2*, which in mutant form could cause these changes in mitotic chromosomes?
3. How could premature separation of centromeres cause the problems seen in RBS?

INSIGHTS AND SOLUTIONS

A previously undiscovered single-cell organism was found living at a great depth on the ocean floor. Its nucleus contained only a single linear chromosome with 7×10^6 nucleotide pairs of DNA coalesced with three types of histonelike proteins. Consider the following questions:

1. A short micrococcal nuclease digestion yielded DNA fractions of 700, 1400, and 2100 bp. Predict what these fractions represent. What conclusions can be drawn?

   Solution: The chromatin fiber may consist of a variation of nucleosomes containing 700 bp of DNA. The 1400- and 2100-bp fractions, respectively, represent two and three linked nucleosomes. Enzymatic digestion may have been incomplete, leading to the latter two fractions.

2. The analysis of individual nucleosomes reveals that each unit contained one copy of each protein and that the short linker DNA contained no protein bound to it. If the entire chromosome consists of nucleosomes (discounting any linker DNA), how many are there, and how many total proteins are needed to form them?

   Solution: Since the chromosome contains 7×10^6 bp of DNA, the number of nucleosomes, each containing 7×10^2 bp, is equal to

   $$7 \times 10^6 / 7 \times 10^2 = 10^4 \text{ nucleosomes}$$

   The chromosome contains 10^4 copies of each of the three proteins, for a total of 3×10^4 proteins.

3. Further analysis revealed the organism's DNA to be a double helix similar to the Watson–Crick model but containing 20 bp per complete turn of the right-handed helix. The physical size of the nucleosome was exactly double the volume occupied by that found in all other known eukaryotes, by virtue of increasing the distance along the fiber axis

(continued)

Insights and Solutions—continued

by a factor of two. Compare the degree of compaction of this organism's nucleosome to that found in other eukaryotes.

Solution: The unique organism compacts a length of DNA consisting of 35 complete turns of the helix (700 bp per nucleosome/20 bp per turn) into each nucleosome. The normal eukaryote compacts a length of DNA consisting of 20 complete turns of the helix (200 bp per nucleosome/10 bp per turn) into a nucleosome one-half the volume of that in the unique organism. The degree of compaction is therefore less in the unique organism.

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we focused on how DNA is organized at the chromosomal level. Along the way, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions:
   (a) How do we know that viral and bacterial chromosomes most often consist of circular DNA molecules devoid of protein?
   (b) What is the experimental basis for concluding that puffs in polytene chromosomes and loops in lampbrush chromosomes are areas of intense transcription of RNA?
   (c) How did we learn that eukaryotic chromatin exists in the form of repeating nucleosomes, each consisting of about 200 base pairs and an octamer of histones?
   (d) How do we know that satellite DNA consists of repetitive sequences and has been derived from regions of the centromere?

CONCEPT QUESTION

2. Review the Chapter Concepts list on p. 215. These all relate to how DNA is organized in viral, prokaryote, and eukaryote chromosomes. Write a short essay that contrasts the major differences between the organization of DNA in viruses and bacteria versus eukaryotes. ■

3. How does the 1200-μm-long chromosome of *E. coli* fit into a bacterial cell 2.0 × 0.5 μm long?
4. Describe how giant polytene chromosomes are formed.
5. What genetic process is occurring in a puff of a polytene chromosome?
6. Describe the structure of lampbrush chromosomes. Where are they located?
7. What chemical and structural properties of histones enable them to successfully package eukaryotic DNA? What is chromatin remodeling, and how is it controlled within eukaryotic cells?
8. Describe the sequence of research findings that led to the development of the model of chromatin structure.
9. Describe the basic structure of a nucleosome. What is the role of histone H1?
10. Describe the transitions that occur as nucleosomes are coiled and folded, ultimately forming a chromatid.
11. In instances in the eukaryotic genome, DNA sequences represent evolutionary vestiges of duplicated copies of genes. What are such regions called and what are their characteristics?
12. Contrast the various categories of repetitive DNA.
13. Define satellite DNA. Describe where it is found in the genome of eukaryotes and its role as part of chromosomes.
14. What do SINE and LINE mean in terms of chromosome structure? Why are they called "repetitive"?
15. Mammals contain a diploid genome consisting of at least 10^9 bp. If this amount of DNA is present as chromatin fibers, where each group of 200 bp of DNA is combined with 9 histones into a nucleosome and each group of 6 nucleosomes is combined into a solenoid, achieving a final packing ratio of 50, determine (a) the total number of nucleosomes in all fibers, (b) the total number of histone molecules combined with DNA in the diploid genome, and (c) the combined length of all fibers.
16. Supercoiled DNA is slightly unwound compared to relaxed DNA and this enables it to assume a more compact structure with enhanced physical stability. Describe the enzymes that control the number of supercoils present in the *E. coli* chromosome. How much would you have to reduce the linking number to increase the number of supercoils by five?
17. A particular variant of the lambda bacteriophage has a DNA double-stranded genome of 51,365 base pairs. How long would this DNA be?
18. While much remains to be learned about the role of nucleosomes and chromatin structure and function, recent research indicates that *in vivo* chemical modification of histones is associated with changes in gene activity. For example, Bernstein and others (2000. *Proc. Natl. Acad. Sci. USA* 97: 5340–5345) determined that acetylation of H3 and H4 is associated with 21.1 percent and 13.8 percent increase in yeast gene activity, respectively, and that yeast heterochromatin is hypomethylated relative to the genome average. Speculate on the significance of these findings in terms of nucleosome–DNA interactions and gene activity.
19. In an article entitled "*Nucleosome Positioning at the Replication Fork,*" Lucchini and others (2002. *EMBO J.* 20: 7294–7302) state, "both the 'old' randomly segregated nucleosomes as well as the 'new' assembled histone octamers rapidly position themselves (within seconds) on the newly replicated DNA strands." Given this statement, how would one compare the distribution of nucleosomes and DNA in newly replicated chromatin? How could one experimentally test the distribution of nucleosomes on newly replicated chromosomes?
20. The human genome contains approximately 10^6 copies of an *Alu* sequence, one of the best-studied classes of short interspersed elements (SINEs), per haploid genome. Individual *Alu*s share a 282-nucleotide consensus sequence followed by a 3′-adenine-rich tail region (Schmid, 1998. *Nucl. Acids Res.* 26: 4541–4550). Given that there are approximately 3×10^9 bp per human haploid genome, about how many base pairs are spaced between each *Alu* sequence?
21. Below is a diagram of the general structure of the bacteriophage λ chromosome. Speculate on the mechanism by which it forms a closed ring upon infection of the host cell.

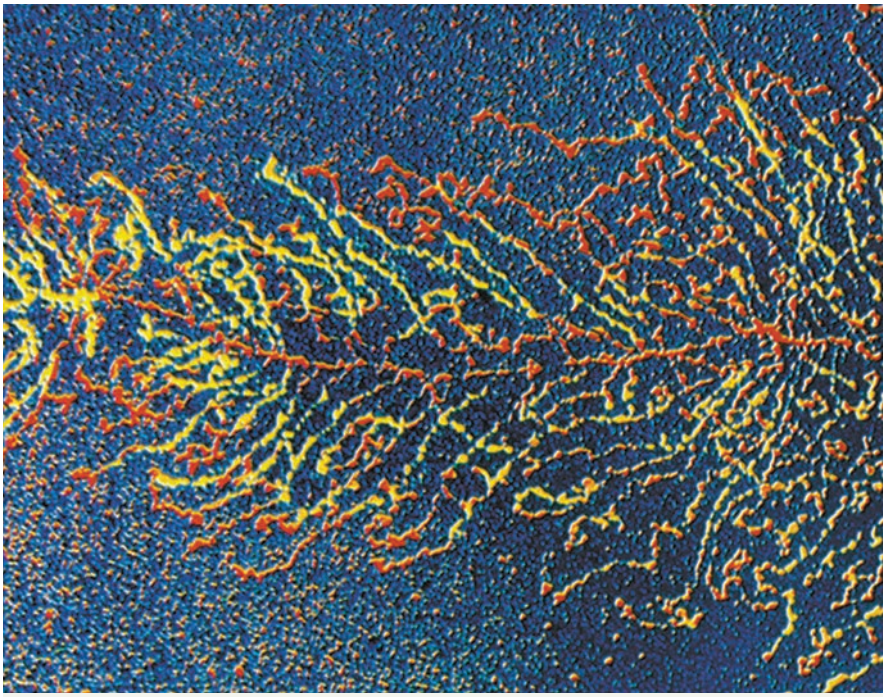
5′GGGCGGCGACCT—double-stranded region—3′

3′—double-stranded region—CCCGCCGCTGGA5′

12 The Genetic Code and Transcription

CHAPTER CONCEPTS

- Genetic information is stored in DNA using a triplet code that is nearly universal to all living things on Earth.
- The encoded genetic information stored in DNA is initially copied into an RNA transcript.
- Once transferred from DNA to RNA, the genetic code exists as triplet codons, using the four ribonucleotides in RNA as the letters composing it.
- By using four different letters taken three at a time, 64 triplet sequences are possible. Most encode one of the 20 amino acids present in proteins, which serve as the end products of most genes.
- Several codons provide signals that initiate or terminate protein synthesis.
- The process of transcription is similar but more complex in eukaryotes compared to prokaryotes and bacteriophages that infect them.



Electron micrograph visualizing the process of transcription.

The linear sequence of deoxyribonucleotides making up DNA ultimately dictates the components constituting proteins, the end product of most genes. The central question is how such information stored as a nucleic acid is decoded into a protein. **Figure 12–1** gives a simplified overview of how this transfer of information occurs. In the first step in gene expression, information on one of the two strands of DNA (the template strand) is copied into an RNA complement through transcription. Once synthesized, this RNA acts as a "messenger" molecule bearing the coded information—hence its name, *messenger RNA (mRNA)*. The mRNAs then associate with ribosomes, where decoding into proteins takes place.

In this chapter, we focus on the initial phases of gene expression by addressing two major questions. First, how is genetic information encoded? Second, how does the transfer from DNA to RNA occur, thus defining the process of transcription? As you shall see, ingenious analytical research established that the genetic code is written in units of three letters—ribonucleotides present in mRNA that reflect the stored information in genes. Most all triplet code words direct the incorporation of a specific amino acid into a protein as it is synthesized. As we can predict based on our prior discussion of the replication of DNA, transcription is also a complex process dependent on a major polymerase enzyme and a cast of supporting proteins. We will explore what is known about transcription in bacteria and then contrast this prokaryotic model with the differences found in eukaryotes. Together, the information in this and the next chapter provides a

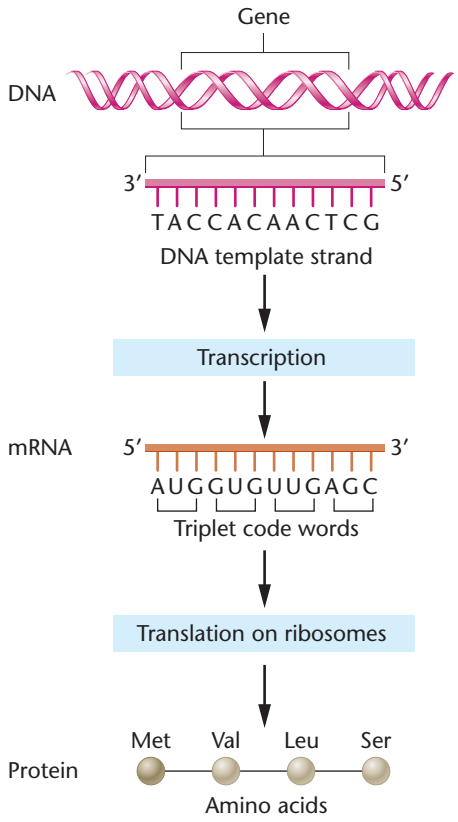


FIGURE 12–1 Flowchart illustrating how genetic information encoded in DNA produces protein.

comprehensive picture of molecular genetics, which serves as the most basic foundation for understanding living organisms. In Chapter 13, we will address how translation occurs and discuss the structure and function of proteins.

12.1 The Genetic Code Exhibits a Number of Characteristics

Before we consider the various analytical approaches that led to our current understanding of the genetic code, let's summarize the general features that characterize it.

1. The genetic code is written in linear form, using the ribonucleotide bases that compose mRNA molecules as "letters." The ribonucleotide sequence is derived from the complementary nucleotide bases in DNA.
2. Each "word" within the mRNA consists of three ribonucleotide letters, thus representing a triplet code. With several exceptions, each group of *three* ribonucleotides, called a codon, specifies *one* amino acid.
3. The code is **unambiguous**—each triplet specifies only a single amino acid.
4. The code is **degenerate;** that is, a given amino acid can be specified by more than one triplet codon. This is the case for 18 of the 20 amino acids.
5. The code contains one "start" and three "stop" signals, triplets that **initiate** and **terminate** translation.
6. No internal punctuation (such as a comma) is used in the code. Thus, the code is said to be **commaless.** Once translation of mRNA begins, the codons are read one after the other, with no breaks between them.
7. The code is **nonoverlapping.** Once translation commences, any single ribonucleotide at a specific location within the mRNA is part of only one triplet.
8. The code is nearly **universal.** With only minor exceptions, almost all viruses, prokaryotes, archaea, and eukaryotes use a single coding dictionary.

12.2 Early Studies Established the Basic Operational Patterns of the Code

In the late 1950s, before it became clear that mRNA is the intermediate that transfers genetic information from DNA to proteins, researchers thought that DNA itself might directly encode proteins during their synthesis. Because ribosomes had already been identified, the initial thinking was that information in DNA was transferred in the nucleus to the RNA of the ribosome, which served as the template for protein synthesis in the cytoplasm. This concept soon became untenable as accumulating evidence indicated the existence of an unstable intermediate template. The RNA of ribosomes, on the other hand, was extremely stable. As a result, in 1961 François Jacob and Jacques Monod postulated the existence of **messenger RNA (mRNA).** Once mRNA was discovered, it was clear that even though genetic information is stored in DNA, the code that is translated into proteins resides in RNA. The central question then was how only four letters—the four ribonucleotides—could specify 20 words (the amino acids).

The Triplet Nature of the Code

In the early 1960s, Sidney Brenner argued on theoretical grounds that the code had to be a triplet since three-letter words represent the minimal use of four letters to specify 20 amino acids. A code of four nucleotides, taken two at a time, for example, provides only 16 unique code words (4^2). A triplet code yields 64 words (4^3)—clearly more than the 20 needed—and is much simpler than a four-letter code (4^4), which specifies 256 words.

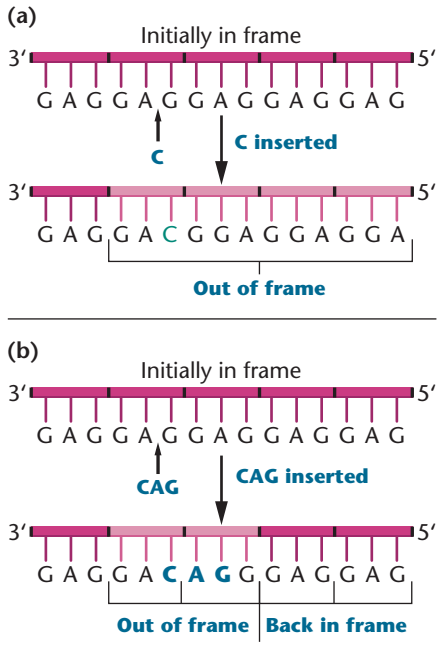


FIGURE 12–2 The effect of frameshift mutations on a DNA sequence with the repeating triplet sequence GAG. (a) The insertion of a single nucleotide shifts all subsequent triplet reading frames. (b) The insertion of three nucleotides changes only two triplets, but the frame of reading is then reestablished to the original sequence.

Experimental evidence supporting the triplet nature of the code was subsequently derived from research by Francis Crick and his colleagues. Using phage T4, they studied **frameshift mutations,** which result from the addition or deletion of one or more nucleotides within a gene and subsequently the mRNA transcribed by it. The gain or loss of letters shifts the *frame of reading* during translation. Crick and his colleagues found that the gain or loss of one or two nucleotides caused a frameshift mutation, but when three nucleotides were involved, the frame of reading was reestablished (**Figure 12–2**). This would not occur if the code was anything other than a triplet. This work also suggested that most triplet codes are not blank, but rather encode amino acids, supporting the concept of a degenerate code.

12.3 Studies by Nirenberg, Matthaei, and Others Deciphered the Code

In 1961, Marshall Nirenberg and J. Heinrich Matthaei deciphered the first specific coding sequences, which served as a cornerstone for the complete analysis of the genetic code. Their success, as well as that of others who made important contributions to breaking the code, was dependent on the use of two experimental tools—an *in vitro (cell-free) protein-synthesizing system* and an enzyme, **polynucleotide phosphorylase,** which enabled the production of synthetic mRNAs. These mRNAs are templates for polypeptide synthesis in the cell-free system.

Synthesizing Polypeptides in a Cell-Free System

In the cell-free system, amino acids are incorporated into polypeptide chains. This *in vitro* mixture must contain the essential factors for protein synthesis in the cell: ribosomes, tRNAs, amino acids, and other molecules essential to translation (see Chapter 13). In order to follow (or trace) protein synthesis, one or more of the amino acids must be radioactive. Finally, an mRNA must be added, which serves as the template that will be translated.

In 1961, mRNA had yet to be isolated. However, use of the enzyme polynucleotide phosphorylase allowed the artificial synthesis of RNA templates, which could be added to the cell-free system. This enzyme, isolated from bacteria, catalyzes the reaction shown in **Figure 12–3**. Discovered in 1955 by Marianne Grunberg-Manago and Severo Ochoa, the enzyme functions metabolically in bacterial cells to degrade RNA. However, *in vitro,* with high concentrations of ribonucleoside diphosphates, the reaction can be "forced" in the opposite direction to synthesize RNA, as shown.

In contrast to RNA polymerase, polynucleotide phosphorylase does not require a DNA template. As a result, each addition of a ribonucleotide is random, based on the relative concentration of the four ribonucleoside diphosphates added to the reaction mixtures. The probability of the insertion of a specific ribonucleotide is proportional to the availability of

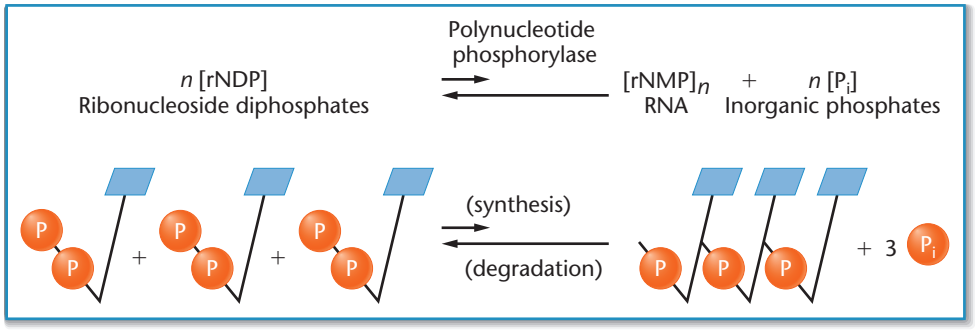


FIGURE 12–3 The reaction catalyzed by the enzyme polynucleotide phosphorylase. Note that the equilibrium of the reaction favors the degradation of RNA but can be "forced" in the direction favoring synthesis.

that molecule, relative to other available ribonucleotides. *This point is absolutely critical to understanding the work of Nirenberg and others in the ensuing discussion.*

Together, the cell-free system and the availability of synthetic mRNAs provided a means of deciphering the ribonucleotide composition of various triplets encoding specific amino acids.

The Use of Homopolymers

In their initial experiments, Nirenberg and Matthaei synthesized **RNA homopolymers,** each with only one type of ribonucleotide. Therefore, the mRNA added to the *in vitro* system was UUUUUU . . . , AAAAAA . . . , CCCCCC . . . , or GGGGGG . . . . They tested each mRNA and were able to determine which, if any, amino acids were incorporated into newly synthesized proteins. To do this, the researchers labeled 1 of the 20 amino acids added to the *in vitro* system and conducted a series of experiments, each with a different radioactively labeled amino acid.

For example, in experiments using ^{14}C-phenylalanine (Table 12.1), Nirenberg and Matthaei concluded that the message poly U (polyuridylic acid) directed the incorporation of only phenylalanine into the homopolymer polyphenylalanine. Assuming the validity of a triplet code, they determined the first specific codon assignment—UUU codes for phenylalanine. Using similar experiments, they quickly found that AAA codes for lysine and CCC codes for proline. Poly G was not an adequate template, probably because the molecule folds back upon itself. Thus, the assignment for GGG had to await other approaches.

Note that the *specific triplet codon assignments* were possible only because homopolymers were used. This method yields only the *composition of triplets,* but since three identical letters can have only one possible sequence (e.g., UUU), the actual codons were identified.

Mixed Copolymers

With these techniques in hand, Nirenberg and Matthaei, and Ochoa and coworkers turned to the use of **RNA heteropolymers.** In this type of experiment, two or more different ribonucleoside diphosphates are added in combination to form the synthetic mRNA. The researchers reasoned that if they knew the relative proportion of each type of ribonucleoside diphosphate, they could predict the frequency of any particular triplet codon occurring in the synthetic mRNA. If they then added the mRNA to the cell-free system and ascertained the percentage of any particular amino acid present in the new protein, they could analyze the results and predict the composition (not the specific sequence) of triplets specifying particular amino acids.

This approach is shown in Figure 12–4. Suppose that A and C are added in a ratio of 1A:5C. The insertion of a ribonucleotide at any position along the RNA molecule during its synthesis is determined by the ratio of A:C. Therefore, there is a 1/6 chance for an A and a 5/6 chance for a C to occupy each position. On this basis, we can calculate the frequency of any given triplet appearing in the message.

For AAA, the frequency is $(1/6)^3$, or about 0.4 percent. For AAC, ACA, and CAA, the frequencies are identical—that is, $(1/6)^2(5/6)$, or about 2.3 percent for each triplet. Together, all three 2A:1C triplets account for 6.9 percent of the total three-letter sequences. In the same way, each of three 1A:2C triplets accounts for $(1/6)(5/6)^2$, or 11.6 percent (or a total of 34.8 percent); CCC is represented by $(5/6)^3$, or 57.9 percent of the triplets.

By examining the percentages of any given amino acid incorporated into the protein synthesized under the direction of this message, we can propose probable base compositions for each amino acid (Figure 12–4). Since proline appears 69 percent of the time, we could propose that proline is encoded by CCC (57.9 percent) and one triplet of 2C:1A (11.6 percent). Histidine, at 14 percent, is probably coded by one 2C:1A (11.6 percent) and one 1C:2A (2.3 percent). Threonine, at 12 percent, is likely coded by only one 2C:1A. Asparagine and glutamine each appear to be coded by one of the 1C:2A triplets, and lysine appears to be coded by AAA.

Using as many as all four ribonucleotides to construct the mRNA, the researchers conducted many similar experiments. Although determining the *composition* of the triplet code words for all 20 amino acids represented a significant breakthrough, the *specific sequences* of triplets were still unknown—other approaches were needed.

TABLE 12.1 Incorporation of ^{14}C-phenylalanine into Protein

Artificial mRNA	Radioactivity (counts/min)
None	44
Poly U	39,800
Poly A	50
Poly C	38

Source: After Nirenberg and Matthaei (1961).

ESSENTIAL POINT

The use of RNA homopolymers and mixed copolymers in a cell-free system allowed the determination of the composition, but not the sequence, of triplet codons designating specific amino acids. ■

The Triplet Binding Assay

It was not long before more advanced techniques were developed. In 1964, Nirenberg and Philip Leder developed the **triplet binding assay,** which led to specific

Possible compositions	Possible triplets	Probability of occurrence of any triplet	Final %
3A	AAA	$(1/6)^3 = 1/216 = 0.4\%$	0.4
1C:2A	AAC ACA CAA	$(1/6)^2(5/6) = 5/216 = 2.3\%$	$3 \times 2.3 = 6.9$
2C:1A	ACC CAC CCA	$(1/6)(5/6)^2 = 25/216 = 11.6\%$	$3 \times 11.6 = 34.8$
3C	CCC	$(5/6)^3 = 125/216 = 57.9\%$	57.9
			100.0

Chemical synthesis of message

RNA

C C C C C C C C A C C C C C C A A C C A C C C C C A C C C C C A C C C A A

Translation of message

Percentage of amino acids in protein		Probable base-composition assignments
Lysine	<1	AAA
Glutamine	2	1C:2A
Asparagine	2	1C:2A
Threonine	12	2C:1A
Histidine	14	2C:1A, 1C:2A
Proline	69	CCC, 2C:1A

FIGURE 12–4 Results and interpretation of a mixed copolymer experiment where a ratio of 1A:5C is used (1/6A:5/6C).

NOW SOLVE THIS

12–1 In a mixed copolymer experiment using polynucleotide phosphorylase, 3/4G:1/4C was used to form the synthetic message. The amino acid composition of the resulting protein was determined to be:

Glycine	36/64	(56 percent)
Alanine	12/64	(19 percent)
Arginine	12/64	(19 percent)
Proline	4/64	(6 percent)

From this information,

(a) indicate the percentage (or fraction) of the time each possible codon will occur in the message.

(b) determine one consistent base-composition assignment for the amino acids present.

(c) Once the wobble hypothesis has been discussed (p. 238), return to this problem and predict as many specific codon assignments as possible.

HINT: *This problem asks you to analyze a mixed copolymer experiment and to predict codon composition assignments for the amino acids encoded by the synthetic message. The key to its solution is to first calculate the proportion of each triplet codon in the synthetic RNA and then match these to the proportions of amino acids that are synthesized.*

assignments of triplets. The technique took advantage of the observation that ribosomes, when presented *in vitro* with an RNA sequence as short as three ribonucleotides, will bind to it and form a complex similar to that found *in vivo*. The triplet acts like a codon in mRNA, attracting the complementary sequence within tRNA (**Figure 12–5**). The triplet sequence in tRNA that is complementary to a codon of mRNA is an **anticodon.**

Although it was not yet feasible to chemically synthesize long stretches of RNA, triplets of known sequence could be synthesized in the laboratory to serve as templates. All that was needed was a method to determine which tRNA–amino acid was bound to the triplet RNA–ribosome complex. The test system Nirenberg and Leder devised was quite simple. The amino acid to be tested was made radioactive, and a charged tRNA was produced. Because codon compositions were known, researchers could narrow the range of amino acids that should be tested for each specific triplet.

The radioactively charged tRNA, the RNA triplet, and ribosomes were incubated together and then passed through a nitrocellulose filter, which retains the larger ribosomes but not the other smaller components, such as unbound charged tRNA. If radioactivity is not retained on the filter, an incorrect amino acid has been tested. But if radioactivity remains on the filter, it is retained because the charged tRNA has bound to

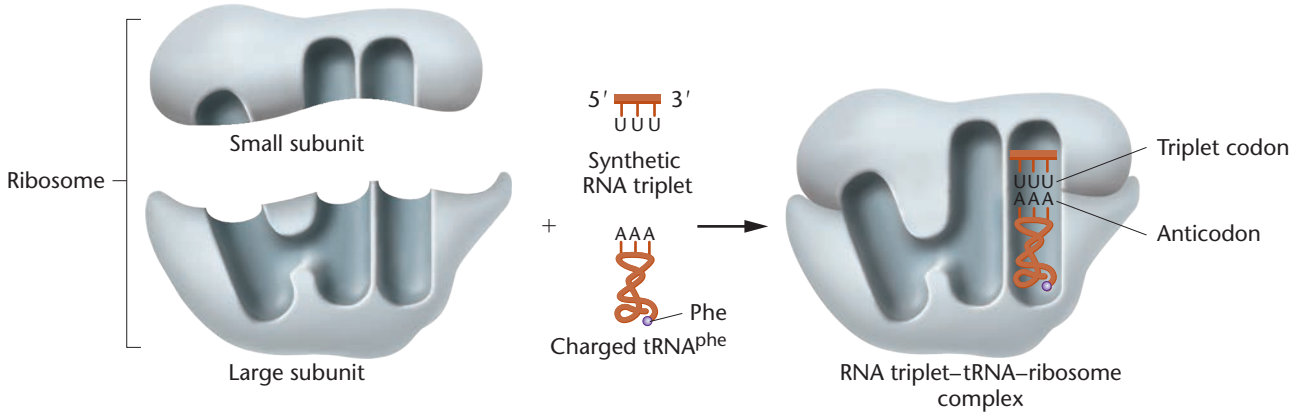


FIGURE 12–5 Illustration of the behavior of the components during the triplet binding assay. The synthetic UUU triplet RNA sequence acts as a codon, attracting the complementary AAA anticodon of the charged $tRNA^{phe}$, which together are bound by the subunits of the ribosome.

the triplet associated with the ribosome. When this occurs, a specific codon assignment can be made.

Work proceeded in several laboratories, and in many cases clear-cut, unambiguous results were obtained. Table 12.2, for example, shows 26 triplets assigned to 9 amino acids. However, in some cases, the degree of triplet binding was inefficient and assignments were not possible. Eventually, about 50 of the 64 triplets were assigned. These specific assignments of triplets to amino acids led to two major conclusions. First, the genetic code is **degenerate;** that is, one amino acid can be specified by more than one triplet. Second, the code is **unambiguous.** That is, a single triplet specifies only one amino acid. As you shall see later in this chapter, these conclusions have been upheld with only minor exceptions. The triplet binding technique was a major innovation in deciphering the genetic code.

Repeating Copolymers

Yet another innovative technique used to decipher the genetic code was developed in the early 1960s by Gobind Khorana, who chemically synthesized long RNA molecules consisting of short sequences repeated many times. First, he created shorter sequences (e.g., di-, tri-, and tetranucleotides), which were then replicated many times and finally joined enzymatically to form the long polynucleotides. As shown in Figure 12–6, a dinucleotide made in this way is converted to an mRNA with two repeating triplets. A trinucleotide is converted to an mRNA with three potential triplets, depending on the point at which initiation occurs, and a tetranucleotide creates four repeating triplets.

When synthetic mRNAs were added to a cell-free system, the predicted number of amino acids incorporated into polypeptides was upheld. Several examples are shown in Table 12.3. When the data were combined with those on composition assignment and triplet binding, specific assignments were possible.

One example of specific assignments made in this way will illustrate the value of Khorana's approach. Consider the following experiments in concert with one another: (1) The repeating *trinucleotide sequence* UUCUUCUUC . . . can be read as three possible repeating triplets—UUC, UCU, and CUU—depending on the initiation point. When placed in a cell-free translation system, three different polypeptide homopolymers—containing either phenylalanine (phe), serine (ser), or leucine (leu)—are produced. Thus, we know that each of the three triplets encodes one of the three amino acids, but we do not know which codes which; (2) On the other hand, the *repeating dinucleotide sequence* UCUCUCUC . . . produces the triplets UCU and CUC and, when used in an experiment, leads to the incorporation of leucine and serine into a polypeptide. Thus, the triplets UCU and CUC specify leucine and serine, but we still do not know which triplet specifies which amino acid. However, when considering both sets of results in concert, we can conclude that UCU, which is common to both experiments, must encode either leucine or serine but not phenylalanine.

TABLE 12.2 Amino Acid Assignments to Specific Trinucleotides Derived from the Triplet Binding Assay

Trinucleotides	Amino Acid
AAA AAG	Lysine
AUG	Methionine
AUU AUC AUA	Isoleucine
CCG CCA CCU CCC	Proline
CUC CUA CUG CUU	Leucine
GAA GAG	Glutamic acid
UCA UCG UCU UCC	Serine
UGU UGC	Cysteine
UUA UUG	Leucine
UUU UUC	Phenylalanine

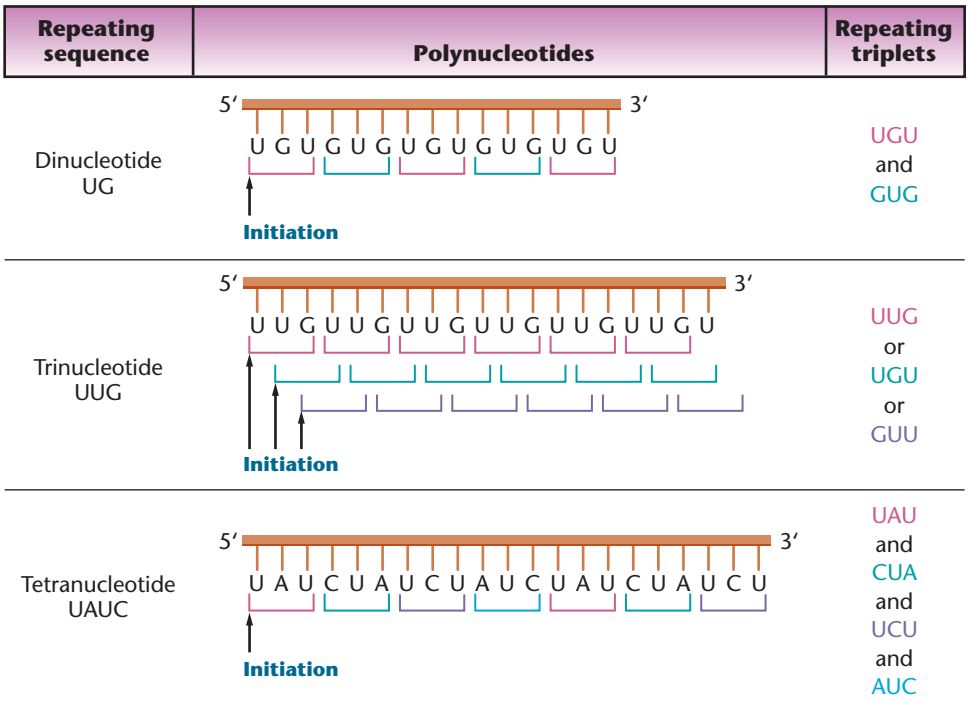


FIGURE 12–6 The conversion of di-, tri-, and tetranucleotides into repeating RNA copolymers. The triplet codons that are produced in each case are shown.

Thus, either CUU *or* UUC encodes leucine *or* serine, while the other encodes phenylalanine; (3) To derive more specific information, we can examine the results of using the repeating tetranucleotide sequence UUAC, which produces the triplets UUA, UAC, ACU, and CUU. The CUU triplet is one of the two in which we are interested. Three amino acids are incorporated by this experiment: leucine, threonine, and tyrosine. Because CUU must specify only serine or leucine, and because, of these two, only leucine appears in the resulting polypeptide, we may conclude that CUU specifies leucine. Once this assignment is established, we can logically determine all others. Of the two triplet pairs remaining (UUC and UCU from the first experiment *and* UCU and CUC from the second experiment), whichever triplet is common to both must encode serine. This is UCU. By elimination, UUC is determined to encode phenylalanine and CUC is determined to encode leucine. Thus, through painstaking logical analysis, four specific triplets encoding three different amino acids have been assigned from these experiments.

TABLE 12.3 Amino Acids Incorporated Using Repeated Synthetic Copolymers of RNA

Repeating Copolymer	Codons Produced	Amino Acids in Polypeptides
UG	UGU	Cysteine
	GUG	Valine
AC	ACA	Threonine
	CAC	Histidine
UUC	UUC	Phenylalanine
	UCU	Serine
	CUU	Leucine
AUC	AUC	Isoleucine
	UCA	Serine
	CAU	Histidine
UAUC	UAU	Tyrosine
	CUA	Leucine
	UCU	Serine
	AUC	Isoleucine
GAUA	GAU	None
	AGA	None
	UAG	None
	AUA	None

From these interpretations, Khorana reaffirmed the identity of triplets that had already been deciphered and filled in gaps left from other approaches. A number of examples are shown in Table 12.3. For example, the use of two tetranucleotide sequences, GAUA and GUAA, suggested that at least two triplets were *termination codons.* Khorana reached this conclusion because neither of these repeating sequences directed the incorporation of more than a few amino acids into a polypeptide, too few for him to detect. There are no triplets common to both messages, and both seemed to contain at least one triplet that terminates protein synthesis. Of the possible triplets in the poly–(GAUA) sequence shown in Table 12.3, UAG was later shown to be a termination codon.

ESSENTIAL POINT

Use of the triplet binding assay and of repeating copolymers allowed the determination of the specific sequences of triplet codons designating specific amino acids. ■

NOW SOLVE THIS

12–2 When repeating copolymers are used to form synthetic mRNAs, dinucleotides produce a single type of polypeptide that contains only two different amino acids. On the other hand, using a trinucleotide sequence produces three different polypeptides, each consisting of only a single amino acid. Why? What will be produced when a repeating tetranucleotide is used?

■ **HINT:** *This problem asks you to consider different outcomes of repeating copolymer experiments. The key to its solution is to be aware that when using a repeating copolymer of RNA, translation can be initiated at different ribonucleotides. You must simply determine the number of triplet codons produced by initiation at each of the different ribonucleotides.*

12.4 The Coding Dictionary Reveals the Function of the 64 Triplets

The various techniques used to decipher the genetic code have yielded a dictionary of 61 triplet codons assigned to amino acids. The remaining three triplets are termination signals and do not specify any amino acid.

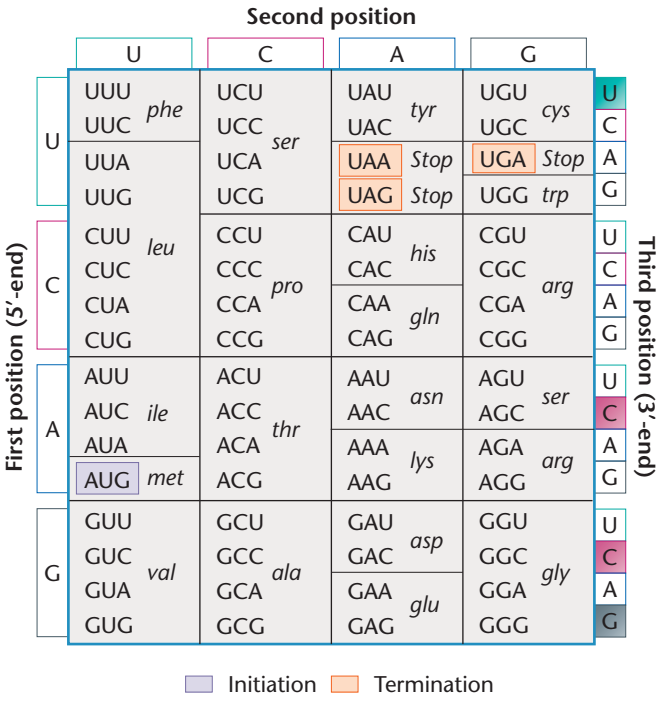


FIGURE 12–7 The coding dictionary. AUG encodes methionine, which initiates most polypeptide chains. All other amino acids except tryptophan, which is encoded only by UGG, are encoded by two to six triplets. The triplets UAA, UAG, and UGA are termination signals and do not encode any amino acids.

Degeneracy and the Wobble Hypothesis

A general pattern of triplet codon assignments becomes apparent when we look at the genetic coding dictionary. Figure 12–7 designates the assignments in a particularly illustrative form first suggested by Francis Crick.

Most evident is that the code is degenerate, as the early researchers predicted. That is, almost all amino acids are specified by two, three, or four different codons. Three amino acids (serine, arginine, and leucine) are each encoded by six different codons. Only tryptophan and methionine are encoded by single codons.

Also evident is the *pattern of degeneracy.* Most often, in a set of codons specifying the same amino acid, the first two letters are the same, with only the third differing. Crick discerned a pattern in the degeneracy at the third position, and in 1966, he postulated the **wobble hypothesis.**

Crick's hypothesis first predicted that the initial two ribonucleotides of triplet codes are more critical than the third in attracting the correct tRNA during translation. He postulated that hydrogen bonding at the third position of the codon–anticodon interaction is less constrained and need not adhere as specifically to the established base-pairing rules. The wobble hypothesis thus proposes a more flexible set of base-pairing rules at the third position of the codon (Table 12.4).

This relaxed base-pairing requirement, or "wobble," allows the anticodon of a single form of tRNA to pair with more than one triplet in mRNA. Consistent with the wobble hypothesis and degeneracy, U at the first position (the 5′-end) of the tRNA anticodon may pair with A or G at the third position (the 3′-end) of the mRNA codon, and G may likewise pair with U or C. Inosine (I), one of the modified bases found in tRNA, may pair with C, U, or A. Applying these wobble rules, a minimum of about 30 different tRNA species is necessary to accommodate the 61 triplets specifying an amino acid. If nothing more, wobble can be considered a potential economy measure, provided that the fidelity of translation is not compromised. Current estimates are that 30 to 40 tRNA species are present in bacteria and up to 50 tRNA species exist in animal and plant cells.

TABLE 12.4 Codon–Anticodon Base-Pairing Rules

Base at First Position (5′-end) of tRNA	Base at Third Position (3′-end) of mRNA
A	U
C	G
G	C or U
U	A or G
I	A, U, or C

The Ordered Nature of the Code

Still another observation has become apparent in the pattern of codon sequences and their corresponding amino acids, leading to the description referred to as an **ordered genetic code.** Chemically similar amino acids often share one or two "middle" bases in the different triplets encoding them. For example, either U or C is often present in the second position of triplets that specify hydrophobic amino acids, including valine and alanine, among others. Two codons (AAA and AAG) specify the positively charged amino acid lysine. If only the middle letter of these codons is changed from A to G (AGA and AGG), the positively charged amino acid arginine is specified. Hydrophilic amino acids, such as serine and threonine, are specified by triplet codons, with G or C in the second position.

The chemical properties of amino acids will be discussed in more detail in Chapter 13. The end result of an "ordered" code is that it buffers the potential effect of mutation on protein function. While many mutations of the second base of triplet codons result in a change of one amino acid to another, the change is often to an amino acid with similar chemical properties. In such cases, protein function may not be noticeably altered.

Initiation and Termination

In contrast to the *in vitro* experiments discussed earlier, initiation of protein synthesis *in vivo* is a highly specific process. In bacteria, the initial amino acid inserted into all polypeptide chains is a modified form of methionine—**N-formylmethionine (fmet).** Only one codon, AUG, codes for methionine, and it is sometimes called the **initiator codon.** However, when AUG appears internally in mRNA, rather than at an initiating position, unformylated methionine is inserted into the polypeptide chain. Rarely, another codon, GUG, specifies methionine during initiation, though it is not clear why this happens, since GUG normally encodes valine.

In bacteria, either the formyl group is removed from the initial methionine upon the completion of protein synthesis or the entire formylmethionine residue is removed. In eukaryotes, methionine is also the initial amino acid during polypeptide synthesis. However, it is not formylated.

As mentioned in the preceding section, three other triplets (UAG, UAA, and UGA) serve as **termination codons,** punctuation signals that do not code for any amino acid. They are not recognized by a tRNA molecule, and translation terminates when they are encountered. Mutations that produce any of the three triplets internally in a gene will also result in termination. Consequently, only a partial polypeptide has been synthesized when it is prematurely released from the ribosome. When such a change occurs in the DNA, it is called a **nonsense mutation.**

ESSENTIAL POINT

The complete coding dictionary reveals that of the 64 possible triplet codons, 61 encode the 20 amino acids found in proteins, while three triplets terminate translation. ■

12.5 The Genetic Code Has Been Confirmed in Studies of Bacteriophage MS2

The various aspects of the genetic code discussed thus far yield a fairly complete picture, suggesting that it is triplet in nature, degenerate, unambiguous, and commaless, and that it contains punctuation start and stop signals. That these features are correct was confirmed by analysis of the RNA-containing **bacteriophage MS2** by Walter Fiers and his coworkers.

MS2 is a bacteriophage that infects *E. coli.* Its nucleic acid (RNA) contains only about 3500 ribonucleotides, making up only three genes, specifying a coat protein, an RNA replicase, and a maturation protein. The small genome and a few gene products enabled Fiers and his colleagues to sequence the genes and their products. When the chemical constitution of these genes and their encoded proteins were compared, they were found to exhibit **colinearity.** That is, based on the coding dictionary, *the linear sequence of triplet codons corresponds precisely with the linear sequence of amino acids in each protein.* Punctuation was also confirmed. For example, in the coat protein gene, the codon for the first amino acids is AUG, the common initiator codon. The codon for the last amino acid is followed by two consecutive termination codons, UAA and UAG. The analysis clearly showed that the genetic code in this virus was identical to that established experimentally in bacterial systems.

12.6 The Genetic Code Is Nearly Universal

Between 1960 and 1978, it was generally assumed that the genetic code would be found to be universal, applying equally to viruses, bacteria, archaea, and eukaryotes. Certainly, the nature of mRNA and the translation machinery seemed to be very similar in these organisms. For example, cell-free systems derived from bacteria can translate eukaryotic mRNAs. Poly U stimulates synthesis of polyphenylalanine in cell-free systems when the components are derived from eukaryotes. Many recent studies involving recombinant DNA technology (see Chapter 17) reveal

that eukaryotic genes can be inserted into bacterial cells, which are then transcribed and translated. Within eukaryotes, mRNAs from mice and rabbits have been injected into amphibian eggs and efficiently translated. For the many eukaryotic genes that have been sequenced, notably those for hemoglobin molecules, the amino acid sequence of the encoded proteins adheres to the coding dictionary established from bacterial studies.

However, several 1979 reports on the coding properties of DNA derived from mitochondria (**mtDNA**) of yeast and humans undermined the principle of the universality of the genetic language. Since then, mtDNA has been examined in many other organisms.

Cloned mtDNA fragments have been sequenced and compared with the amino acid sequences of various mitochondrial proteins, revealing several exceptions to the coding dictionary (Table 12.5). Most surprising is that the codon UGA, normally specifying termination, encodes tryptophan during translation in yeast and human mitochondria. In yeast mitochondria, threonine is inserted instead of leucine when CUA is encountered in mRNA. In human mitochondria, AUA, which normally specifies isoleucine, directs the internal insertion of methionine.

In 1985, several other exceptions to the standard coding dictionary were discovered in the bacterium *Mycoplasma capricolum* and in the nuclear genes of the protozoan ciliates *Paramecium, Tetrahymena,* and *Stylonychia.* For example, as shown in Table 12.5, one alteration converts the termination codon UGA to tryptophan, yet several others convert the normal termination codons UAA and UAG to glutamine. These changes are significant because both a prokaryote and several eukaryotes are involved, representing distinct species that have evolved separately over a long period of time.

Note the apparent pattern in several of the altered codon assignments. The change in coding capacity involves only a shift in recognition of the third, or wobble, position. For example, AUA specifies isoleucine in the cytoplasm and methionine in the mitochondrion, but in the cytoplasm, methionine is specified by AUG. Similarly, UGA calls for termination in the cytoplasm, but it specifies tryptophan in the mitochondrion; in the cytoplasm, tryptophan is specified by UGG. It has been suggested that such changes in codon recognition may represent an evolutionary trend toward reducing the number of tRNAs needed in mitochondria; only 22 tRNA species are encoded in human mitochondria, for example. However, until more examples are found, the differences must be considered to be exceptions to the previously established general coding rules.

TABLE 12.5 **Exceptions to the Universal Code**

Triplet	Normal Code Word	Altered Code Word	Source
UGA	Termination	Tryptophan	Human and yeast mitochondria; *Mycoplasma*
CUA	Leucine	Threonine	Yeast mitochondria
AUA	Isoleucine	Methionine	Human mitochondria
AGA	Arginine	Termination	Human mitochondria
AGG	Arginine	Termination	Human mitochondria
UAA	Termination	Glutamine	*Paramecium; Tetrahymena; Stylonychia*
UAG	Termination	Glutamine	*Paramecium*

12.7 Different Initiation Points Create Overlapping Genes

Earlier we stated that the genetic code is nonoverlapping—each ribonucleotide in an mRNA is part of only one codon. However, this characteristic of the code does not rule out the possibility that a single mRNA may have multiple initiation points for translation. If so, these points could theoretically create several different reading frames within the same mRNA, thus specifying more than one polypeptide and leading to the concept of **overlapping genes.**

That this might actually occur in some viruses was suspected when phage ϕX174 was carefully investigated. The circular DNA chromosome consists of 5386 nucleotides, which should encode a maximum of 1795 amino acids, sufficient for five or six proteins. However, this small virus in fact synthesizes 11 proteins consisting of more than 2300 amino acids. A comparison of the nucleotide sequence of the DNA and the amino acid sequences of the polypeptides synthesized has clarified the apparent paradox. At least four cases of multiple initiation have been discovered, creating overlapping genes.

For example, in one case, the coding sequences for the initiation of two polypeptides are found at separate positions within the reading frame that specifies the sequence of a third polypeptide. In one case, seven different polypeptides may be created from a DNA sequence that might otherwise have specified only three polypeptides.

A similar situation has been observed in other viruses, including phage G4 and the animal virus SV40. Like ϕX174, phage G4 contains a circular single-stranded DNA molecule. The use of overlapping reading frames optimizes the use of a limited amount of DNA present in these small viruses. However, such an approach to storing information has a distinct disadvantage in that a single mutation may affect more than one protein and thus increase the chances that the change will be deleterious or lethal.

12.8 Transcription Synthesizes RNA on a DNA Template

Even while the genetic code was being studied, it was quite clear that proteins were the end products of many genes. Thus, while some geneticists attempted to elucidate the code, other research efforts focused on the nature of genetic expression. The central question was how DNA, a nucleic acid, could specify a protein composed of amino acids.

The complex multistep process begins with the transfer of genetic information stored in DNA to RNA. The process by which RNA molecules are synthesized on a DNA template is called **transcription.** It results in an mRNA molecule complementary to the gene sequence of one of the double helix's two strands. Each triplet codon in the mRNA is, in turn, complementary to the anticodon region of its corresponding tRNA as the amino acid is correctly inserted into the polypeptide chain during translation. The significance of transcription is enormous, for it is the initial step in the process of **information flow** within the cell. The idea that RNA is involved as an intermediate molecule in the process of information flow between DNA and protein was suggested by the following observations:

1. DNA is, for the most part, associated with chromosomes in the nucleus of the eukaryotic cell. However, protein synthesis occurs in association with ribosomes located outside the nucleus in the cytoplasm. Therefore, DNA does not appear to participate directly in protein synthesis.
2. RNA is synthesized in the nucleus of eukaryotic cells, where DNA is found, and is chemically similar to DNA.
3. Following its synthesis, most RNA migrates to the cytoplasm, where protein synthesis (translation) occurs.
4. The amount of RNA is generally proportional to the amount of protein in a cell.

Collectively, these observations suggested that genetic information, stored in DNA, is transferred to an RNA intermediate, which directs the synthesis of proteins. As with most new ideas in molecular genetics, the initial supporting evidence was based on experimental studies of bacteria and their phages. It was clearly established that during initial infection, RNA synthesis preceded phage protein synthesis and that the RNA is complementary to phage DNA.

The results of these experiments agree with the concept of a messenger RNA (mRNA) being made on a DNA template and then directing the synthesis of specific proteins in association with ribosomes. This concept was formally proposed by François Jacob and Jacques Monod in 1961 as part of a model for gene regulation in bacteria. Since then, mRNA has been isolated and studied thoroughly. There is no longer any question about its role in genetic processes.

12.9 RNA Polymerase Directs RNA Synthesis

To establish that RNA can be synthesized on a DNA template, it was necessary to demonstrate that there is an enzyme capable of directing this synthesis. By 1959, several investigators, including Samuel Weiss, had independently isolated such a molecule from rat liver. Called **RNA polymerase,** it has the same general substrate requirements as does DNA polymerase, the major exception being that the substrate nucleotides contain the ribose rather than the deoxyribose form of the sugar. Unlike DNA polymerase, no primer is required to initiate synthesis; the initial base remains as a nucleoside triphosphate (NTP). The overall reaction summarizing the synthesis of RNA on a DNA template can be expressed as

$$n(\text{NTP}) \xrightarrow[\text{enzyme}]{\text{DNA}} (\text{NMP})_n + n(\text{PP}_\text{i})$$

As this equation reveals, nucleoside triphosphates (NTPs) are substrates for the enzyme, which catalyzes the polymerization of nucleoside monophosphates (NMPs), or nucleotides, into a polynucleotide chain $(\text{NMP})_n$. Nucleotides are linked during synthesis by 3′-to-5′ phosphodiester bonds (see Figure 9–10). The energy created by cleaving the triphosphate precursor into the monophosphate form drives the reaction, and inorganic pyrophosphates (PP_i) are produced.

A second equation summarizes the sequential addition of each ribonucleotide as the process of transcription progresses

$$(\text{NMP})_\text{n} + \text{NTP} \xrightarrow[\text{enzyme}]{\text{DNA}} (\text{NMP})_{n+1} + \text{PP}_\text{i}$$

As this equation shows, each step of transcription involves the addition of one ribonucleotide (NMP) to the growing polyribonucleotide chain $(\text{NMP})_{n+1}$, using a nucleoside triphosphate (NTP) as the precursor.

RNA polymerase from *E. coli* has been extensively characterized and shown to consist of subunits designated α, β, β', and σ. The active form of the enzyme, the **holoenzyme,** contains the subunits α_2, β, β', and σ and has a molecular weight of almost 500,000 Da. Of these subunits, it is the β and β' polypeptides that provide the catalytic basis and active site for transcription. Still another subunit, the σ **(sigma) factor,** plays a regulatory function in the initiation of RNA transcription.

Although there is but a single form of the enzyme in *E. coli,* there are several different σ factors, creating variations of the polymerase holoenzyme. On the other hand, eukaryotes display three distinct forms of RNA polymerase, each consisting of a greater number of polypeptide subunits than in bacteria.

ESSENTIAL POINT

Transcription—the initial step in gene expression—is the synthesis, under the direction of RNA polymerase, of a strand of RNA complementary to a DNA template. ■

NOW SOLVE THIS

12–3 The following represent deoxyribonucleotide sequences in the template strand of DNA:

Sequence 1: 5′-CTTTTTTGCCAT-3′
Sequence 2: 5′-ACATCAATAACT-3′
Sequence 3: 5′-TACAAGGGTTCT-3′

(a) For each strand, determine the mRNA sequence that would be derived from transcription.
(b) Using Figure 12–7, determine the amino acid sequence that is encoded by these mRNAs.
(c) For Sequence 1, what is the sequence of the partner DNA strand?

■ **HINT:** *This problem asks you to consider the outcome of the transfer of complementary information from DNA to RNA and to determine the amino acids encoded by this information. The key to its solution is to remember that in RNA, uracil is complementary to adenine, and that while DNA stores genetic information in the cell, the code that is translated is contained in the RNA complementary to the template strand of DNA making up a gene.*

Promoters, Template Binding, and the σ Subunit

Transcription results in the synthesis of a single-stranded RNA molecule complementary to a region along only one strand of the DNA double helix. For simplicity, let's call the transcribed DNA strand the **template strand** and its complement the **partner strand.**

The initial step is **template binding** (Figure 12–8). In bacteria, the site of this initial binding is established when the RNA polymerase σ subunit recognizes specific DNA sequences called **promoters.** These regions are located in the region upstream (5′) from the point of initial transcription of a gene. It is believed that the enzyme "explores" a length of DNA until it recognizes the promoter region and binds to about 60 nucleotide pairs of the helix, 40 of which are upstream from the point of initial transcription. Once this occurs, the helix is denatured or unwound locally, making the DNA template accessible to the action of the enzyme. The point at which transcription actually begins is called the **transcription start site.**

The importance of promoter sequences cannot be overemphasized. They govern the efficiency of the initiation of transcription. In bacteria, both strong promoters and weak promoters have been discovered. Because the interaction of promoters with RNA polymerase governs transcription, the nature of the binding between them is at the heart of discussions concerning genetic regulation, the subject of Chapter 15. While we will later pursue more detailed information involving promoter–enzyme interactions, we must address two points here.

The first point is the concept of **consensus sequences** of DNA. These sequences are similar (homologous) in different genes of the same organism or in one or more genes of related organisms. Their conservation throughout evolution attests to the critical nature of their role in biological processes. Two such sequences have been found in bacterial promoters. One, TATAAT, is located 10 nucleotides upstream from the site of initial transcription (the −10 region, or **Pribnow box**). The other, TTGACA, is located 35 nucleotides upstream (the −35 region). Mutations in either region diminish transcription, often severely.

Sequences such as these are said to be ***cis*-acting elements.** Use of the term *cis* is drawn from organic chemistry nomenclature, meaning "next to" or on the same side as, in contrast to being "across from," or *trans,* to other functional groups. In molecular genetics, then, *cis*-elements are adjacent parts of the same DNA molecule. This is in contrast to ***trans*-acting factors,** molecules that bind to these DNA elements. As we will soon learn, in most eukaryotic genes studied, a consensus sequence comparable to that in the −10 region has been recognized. Because it is rich in adenine and thymine residues, it is called the **TATA box.**

The second point involves the σ subunit in bacteria. The major form is designated as σ^{70}, based on its molecular weight of 70 kilodaltons (kDa). The promoters of most bacterial genes are recognized by this form; however, several alternative forms of RNA polymerase in *E. coli* have unique σ subunits associated with them (e.g., σ^{32}, σ^{54}, σ^{S}, and σ^{E}). Each form recognizes different promoter sequences, which in turn provides specificity to the initiation of transcription.

Initiation, Elongation, and Termination of RNA Synthesis

Once it has recognized and bound to the promoter [Figure 12–8(b)], RNA polymerase catalyzes **initiation,** the insertion of the first 5′-ribonucleoside triphosphate, which is complementary to the first nucleotide at the start site of the DNA template strand. As we noted earlier, no primer is required. Subsequent ribonucleotide complements are inserted and linked by phosphodiester bonds as RNA polymerization proceeds. This process of **chain elongation** [Figure 12–8(c)] continues in a 5′ to 3′ extension, creating a temporary DNA/RNA duplex whose chains run antiparallel to one another.

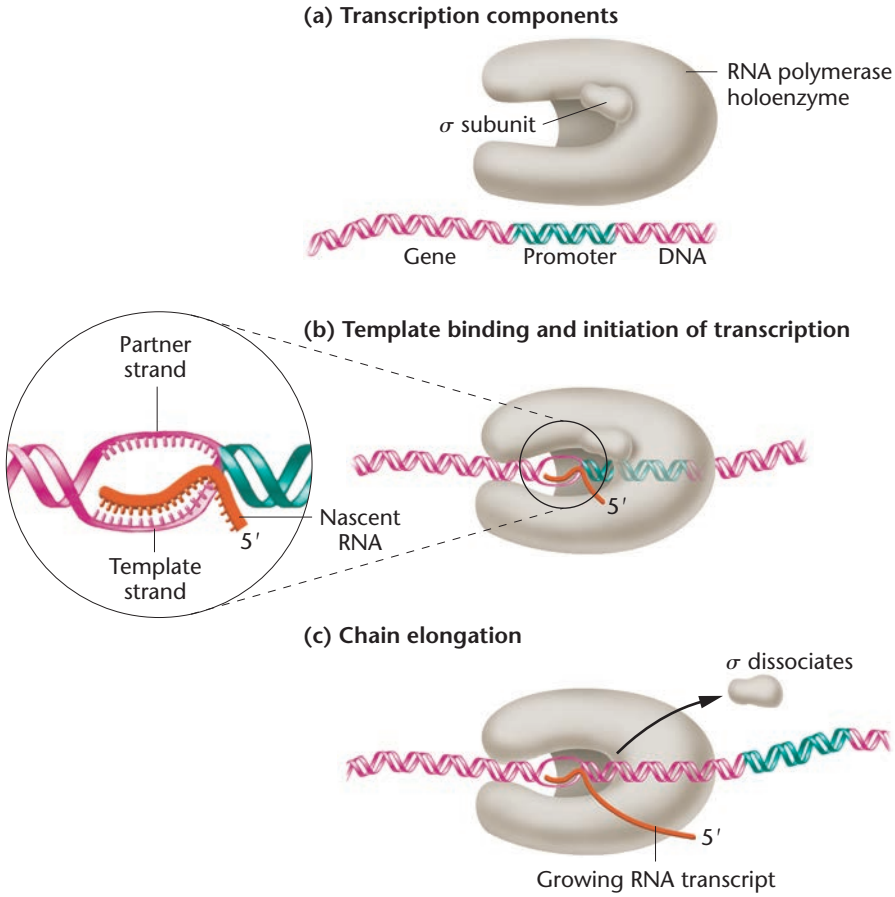


FIGURE 12–8 The early stages of transcription in prokaryotes, showing (a) the components of the process; (b) template binding at the −10 site involving the σ subunit of RNA polymerase and subsequent initiation of RNA synthesis; and (c) chain elongation, after the σ subunit has dissociated from the transcription complex and the enzyme moves along the DNA template.

After a few ribonucleotides have been added to the growing RNA chain, the σ subunit dissociates from the holoenzyme and elongation proceeds under the direction of the core enzyme. In *E. coli,* this process proceeds at the rate of about 50 nucleotides/second at 37°C.

Eventually, the enzyme traverses the entire gene until it encounters a specific nucleotide sequence that acts as a termination signal. The termination sequences, about 40 base pairs in length, are extremely important in prokaryotes because of the close proximity of the end of one gene and the upstream sequences of the adjacent gene. An interesting aspect of termination in bacteria is that the termination sequence alluded to above is actually transcribed into RNA. The unique sequence of nucleotides in this termination region causes the newly formed transcript to fold back on itself, forming what is called a **hairpin secondary structure,** held together by hydrogen bonds. The hairpin is important to termination. In some cases, the termination of synthesis is dependent on the **termination factor, ρ (rho)**—a large hexameric protein that physically interacts with the growing RNA transcript.

At the point of termination, the transcribed RNA molecule is released from the DNA template and the core polymerase enzyme dissociates. The synthesized RNA molecule is precisely complementary to a DNA sequence representing the template strand of a gene. Wherever an A, T, C, or G residue existed, a corresponding U, A, G, or C residue, respectively, is incorporated into the RNA molecule. These RNA molecules ultimately provide the information leading to the synthesis of all proteins present in the cell.

In bacteria, groups of genes whose products are related are often clustered along the chromosome. In many such cases, they are contiguous, and all but the last gene lack the encoded signals for termination. The result is that during transcription, a large mRNA is produced that encodes more than one protein. Since genes in bacteria are sometimes called **cistrons,** the RNA is called a **polycistronic mRNA.** The products of genes transcribed in this fashion are usually all needed by the cell at the same time, so this is an efficient way to transcribe and subsequently translate the needed genetic information. In eukaryotes, **monocistronic mRNAs** are the rule.

12.10 Transcription in Eukaryotes Differs from Prokaryotic Transcription in Several Ways

Much of our knowledge of transcription has been derived from studies of prokaryotes. The general aspects of the mechanics of these processes are mostly similar in eukaryotes, but there are several notable differences:

1. Transcription in eukaryotes occurs within the nucleus under the direction of three separate forms of RNA polymerase. Unlike prokaryotes, in eukaryotes the RNA transcript is not free to associate with ribosomes prior to the completion of transcription. For the mRNA to be translated, it must move out of the nucleus into the cytoplasm.

2. Initiation of transcription of eukaryotic genes requires that the compact chromatin fiber, characterized by nucleosome coiling, must be uncoiled and the DNA made accessible to RNA polymerase and other regulatory proteins. This transition is referred to as **chromatin remodeling,** reflecting the dynamics involved in the conformational change that occurs as the DNA helix is opened (see Chapter 11).

3. Initiation and regulation of transcription entail a more extensive interaction between *cis*-acting DNA sequences and *trans*-acting protein factors involved in stimulating and initiating transcription. Eukaryotic RNA polymerases, for example, rely on *transcription factors* (*TFs*) to scan and bind to DNA. In addition to promoters, other control units, called *enhancers* and *silencers*, may be located in the 5′ regulatory region upstream from the initiation point, but they have also been found within the gene or even in the 3′ downstream region, beyond the coding sequence.

4. Alteration of the primary RNA transcript to produce mature eukaryotic mRNA involves many complex stages referred to generally as "processing." An initial processing step involves the addition of a 5′ cap and a 3′ tail to most transcripts destined to become mRNAs. Other extensive modifications occur to the internal nucleotide sequence of eukaryotic RNA transcripts that eventually serve as mRNAs. The initial (or primary) transcripts are most often much larger than those that are eventually translated. Sometimes called **pre-mRNAs,** they are part of a group of molecules found only in the nucleus—a group referred to collectively as **heterogeneous nuclear RNA (hnRNA).** Such RNA molecules are of variable but large size and are complexed with proteins, forming **heterogeneous nuclear ribonucleoprotein particles (hnRNPs).** Only about 25 percent of hnRNA molecules are converted to mRNA. In those that are converted, substantial amounts of the ribonucleotide sequence are excised, and the remaining segments are spliced back together prior to nuclear export and translation. This phenomenon has given rise to the concepts of **split genes** and **splicing** in eukaryotes.

In the remainder of this chapter, we will look at the basic details of transcription in eukaryotic cells. The process of transcription is highly regulated, determining which DNA sequences are copied into RNA and when and how frequently they are transcribed. We will return to topics directly related to regulation of eukaryotic transcription in Chapter 15.

Initiation of Transcription in Eukaryotes

Eukaryotic RNA polymerase exists in three unique forms, each of which transcribes different types of genes, as indicated in Table 12.6. Each enzyme is larger and more complex than the single prokaryotic polymerase. For example, in yeast, the holoenzyme consists of two large subunits and ten smaller subunits.

TABLE 12.6 RNA Polymerases in Eukaryotes

Form	Product	Location
I	rRNA	Nucleolus
II	mRNA, snRNA	Nucleoplasm
III	5*S* rRNA, tRNA	Nucleoplasm

In regard to the initial template-binding step and promoter regions, most is known about **RNA polymerase II (RNAP II**), which is responsible for the transcription of a wide range of genes in eukaryotes. The activity of RNAP II is dependent on both *cis*-acting elements surrounding the gene itself and a number of *trans*-acting transcription factors that bind to these DNA elements (we will return to the topic of transcription factors below). At least four *cis*-acting DNA elements regulate the initiation of transcription by RNAP II. The first of these elements, the **core-promoter,** determines where RNAP II binds to the DNA and where it begins copying the DNA into RNA. The other three types of regulatory DNA sequences, called **proximal-promoter elements, enhancers,** and **silencers,** influence the efficiency or the rate of transcription initiation by RNAP II. Recall that in prokaryotes, the DNA sequence recognized by RNA polymerase is also called the promoter. In eukaryotes, however, transcriptional initiation is controlled by a larger number of *cis*-acting DNA elements.

In many eukaryotic genes, a *cis*-acting core-promoter element is the **TATA box (**or the **Goldberg-Hogness box).** Located about 30 nucleotide pairs upstream (−30) from the start point of transcription, TATA boxes share a consensus sequence $TATA^{A}/_{T}AAR$, where R indicates any purine nucleotide. The sequence and function of TATA boxes are analogous to those found in the −10 promoter region of prokaryotic genes. However, recall that in prokaryotes, RNA polymerase binds directly to the −10 promoter region. As we will see below, this is not the case in eukaryotes. A wide range of core-promoter and proximal-promoter elements are also found within eukaryotic gene-regulatory regions, and each can have an effect on the efficiency of transcription initiation from the start site. Many of these elements will be discussed in more detail in Chapter 15.

Although eukaryotic promoter elements can determine the site and general efficiency of initiation, other elements, known as enhancers and silencers, have more dramatic effects on eukaryotic gene transcription. As their names suggest, enhancers increase transcription levels and silencers decrease them. The locations of these elements can vary from immediately upstream from a promoter to downstream

from, within, or kilobases away from a gene. Thus they can modulate transcription from a distance. Enhancers and silencers often act to increase or decrease transcription in response to a cell's requirement for a gene product, or at a particular time during development or place within the organism. Each eukaryotic gene has its own unique arrangement of proximal promoter, enhancer, and silencer elements.

Complementing these *cis*-acting regulatory sequences are various *trans*-acting factors that facilitate RNAP II binding and, therefore, the initiation of transcription. These proteins are referred to as **transcription factors.** There are two broad categories of transcription factors: the **general transcription factors** (**GTFs**) that are absolutely required for all RNAP II–mediated transcription, and the **transcriptional activators** and **repressors** that influence the efficiency or the rate of RNAP II transcription initiation. The general transcription factors are essential because RNAP II cannot bind directly to eukaryotic core-promoter sites and initiate transcription without their presence. The general transcription factors involved with human RNAP II binding are well characterized and designated **TFIIA**, **TFIIB**, and so on. One of these, **TFIID**, binds directly to the TATA-box sequence. Once initial binding of TFIID to DNA occurs, the other general transcription factors, along with RNAP II, bind sequentially to TFIID, forming an extensive **pre-initiation complex.**

The specific transcription factors (activators and repressors, above) bind to enhancer and silencer elements and regulate transcription initiation by aiding or preventing the assembly of pre-initiation complexes and the release of RNAP II from pre-initiation into full transcription elongation. They appear to supplant the role of the σ factor seen in the prokaryotic enzyme and are important in eukaryotic gene regulation. We will consider the roles of general and specific transcription factors in eukaryotic gene regulation, as well as the various DNA elements that bind these factors (Chapter 15).

Heterogeneous Nuclear RNA and Its Processing: Caps and Tails

In bacteria the base sequence of DNA is transcribed into an mRNA that is immediately and directly translated into the amino acid sequence as dictated by the genetic code. In contrast, eukaryotic RNA transcripts require significant alteration before they are transported to the cytoplasm and translated. By 1970, accumulating evidence showed that eukaryotic mRNA is transcribed initially as a precursor molecule much larger than that which is translated into protein. This notion was based on the observation by James Darnell and his coworkers of the large **heterogeneous nuclear RNA** (**hnRNA**) in mammalian nuclei that contained nucleotide sequences common to the smaller mRNA molecules present in the cytoplasm. They proposed that the initial transcript of

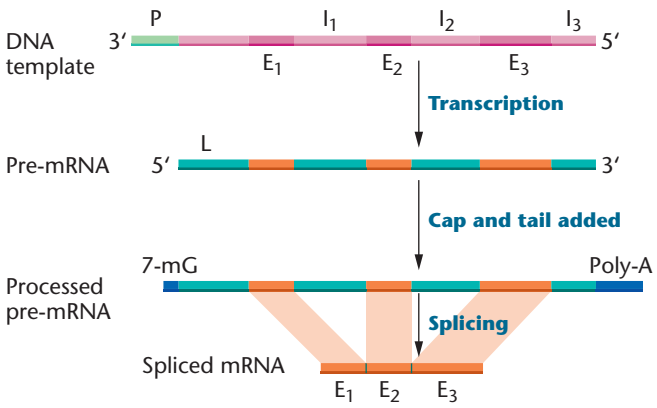


FIGURE 12–9 Posttranscriptional RNA processing in eukaryotes. Transcription produces a pre-mRNA containing a leader sequence (L), several introns (I), and several exons (E), as identified in the DNA template strand. This is processed by the addition of a 5′ 7-mG cap and a 3′ poly-A tail. The introns are then spliced out and the exons joined to create the mature mRNA. While the above figure depicts these steps sequentially, in some eukaryotic transcripts, splicing actually occurs in introns before transcription is complete and the poly-A tail has been added, leading to the concept of *cotranscriptional splicing*.

a gene results in a large RNA molecule that must be processed in the nucleus before it appears in the cytoplasm as a mature mRNA molecule. The various processing steps, discussed in the sections that follow, are summarized in Figure 12–9.

An important **posttranscriptional modification** of eukaryotic RNA transcripts destined to become mRNAs occurs at the 5′-end of these molecules, where a **7-methylguanosine (7-mG)** cap is added. The cap is added even before synthesis of the initial transcript is complete and appears to be important to subsequent processing within the nucleus. The cap also protects the 5′-end of the molecule from nuclease attack. Subsequently, it may be involved in the transport of mature mRNAs across the nuclear membrane into the cytoplasm and in the initiation of translation of the mRNA into protein. The cap is fairly complex and is distinguished by the unique 5′-5′ bonding that connects it to the initial ribonucleotide of the RNA. Some eukaryotes also acquire a methyl group (CH_3) at the 2′-carbon of the ribose sugars of the first two ribonucleotides of the RNA.

Further insights into the processing of RNA transcripts during the maturation of mRNA came from the discovery that both pre-RNAs and mRNAs contain at their 3′-end a stretch of as many as 250 adenylic acid residues. This **poly-A tail** is added after the 3′-end of the initial transcript is cleaved enzymatically at a position some 10 to 35 ribonucleotides from a highly conserved AAUAAA sequence. Poly A has now been found at the 3′-end of almost all mRNAs studied in a variety of eukaryotic organisms. In fact, poly-A tails have also been detected in some prokaryotic mRNAs. The exceptions in eukaryotes seem to be the RNAs that encode the histone proteins.

While the AAUAAA sequence is not found on all eukaryotic transcripts, it appears to be essential to those that have it. If the sequence is changed as a result of a mutation, those transcripts that would normally have it cannot add the poly-A tail. In the absence of this tail, these RNA transcripts are rapidly degraded. Both the 5′ cap and the 3′ poly-A tail are critical if an mRNA transcript is to be transported to the cytoplasm and translated.

ESSENTIAL POINT

The process of creating the initial transcript during transcription is more complex in eukaryotes than in prokaryotes, including the addition of a 5′ 7-mG cap and a 3′ poly-A tail, to the pre-mRNA. ■

12.11 The Coding Regions of Eukaryotic Genes Are Interrupted by Intervening Sequences Called Introns

One of the most exciting findings in molecular genetics occurred in 1977, when Susan Berget, Philip Sharp, and Richard Roberts presented direct evidence that the genes of animal viruses contain *internal* nucleotide sequences that are not expressed in the amino acid sequence of the proteins they encode. These internal DNA sequences are represented in initial RNA transcripts, but they are removed before the mature mRNA is translated (Figure 12–9). Such nucleotide segments are called intervening sequences, and the genes that contain them are split genes. DNA sequences that are not represented in the final mRNA product are also called introns ("int" for intervening), and those retained and expressed are called exons ("ex" for expressed). Splicing involves the removal of the corresponding ribonucleotide sequences representing introns as a result of an excision process and the rejoining of the regions representing exons.

Similar discoveries were soon made in many other eukaryotic genes. Two approaches have been most fruitful for this purpose. The first involves the molecular hybridization of purified, functionally mature mRNAs with DNA containing the genes from which the RNA was originally transcribed. Hybridization between nucleic acids that are not perfectly complementary results in heteroduplexes, in which introns present in the DNA but absent in the mRNA loop out and remain unpaired. Such structures can be visualized with the electron microscope, as shown in **Figure 12–10**. The chicken ovalbumin complex shown in the figure is a heteroduplex with seven loops (A through G), representing seven introns whose sequences are present in the DNA but not in the final mRNA.

The second approach provides more specific information. It involves a direct comparison of nucleotide sequences of DNA with those of mRNA and their correlation with amino acid sequences. Such an approach allows the precise identification of all intervening sequences.

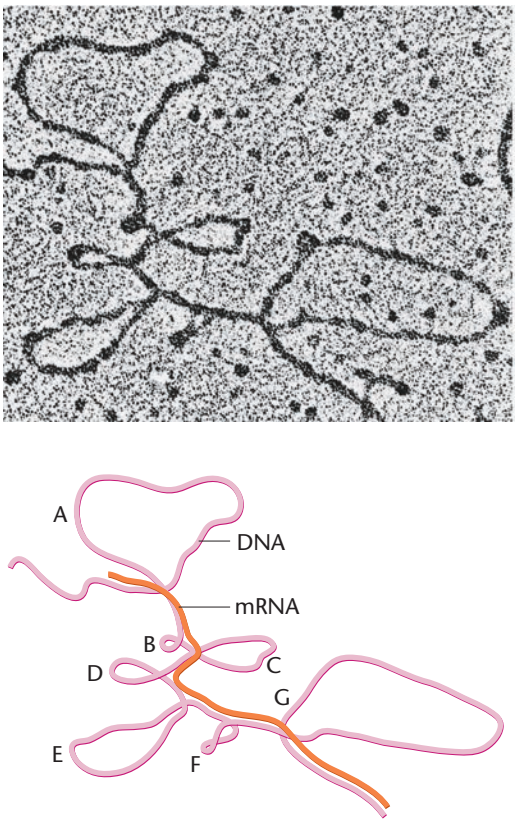


FIGURE 12–10 An electron micrograph and interpretive drawing of the hybrid molecule (heteroduplex) formed between the template DNA strand of the chicken ovalbumin gene and the mature ovalbumin mRNA. Seven DNA introns, labeled A–G, produce unpaired loops.

Thus far, most eukaryotic genes have been shown to contain introns (**Figure 12–11**). One of the first so identified was the β-globin gene in mice and rabbits, studied independently by Philip Leder and Richard Flavell. The mouse gene contains an intron 550 nucleotides long, beginning immediately

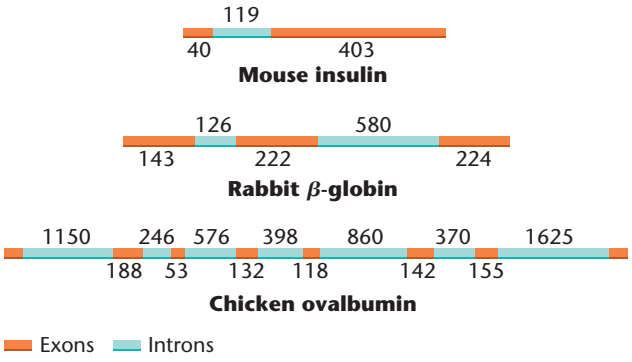


FIGURE 12–11 Intervening sequences in various eukaryotic genes. The numbers indicate the number of nucleotides present in various intron and exon regions.

after the codon specifying the 104th amino acid. In the rabbit, there is an intron of 580 base pairs near the codon for the 110th amino acid. In addition, another intron of about 120 nucleotides exists earlier in both genes. Similar introns have been found in the β-globin gene in all mammals examined.

The ovalbumin gene of chickens has been extensively characterized by Bert O'Malley in the United States and Pierre Chambon in France. As shown in Figure 12–11, the gene contains seven introns. In fact, the majority of the gene's DNA sequence is composed of introns and is thus "silent." The initial RNA transcript is nearly three times the length of the mature mRNA. Compare the ovalbumin gene in Figures 12–10 and 12–11. Can you match the unpaired loops in Figure 12–10 with the order of introns specified in Figure 12–11?

The list of genes containing intervening sequences is long. In fact, few eukaryotic genes seem to lack introns. An extreme example of the number of introns in a single gene is provided by the gene coding for one of the subunits of collagen, the major connective tissue protein in vertebrates. The *pro-α-2(1)* collagen gene contains 50 introns. The precision of cutting and splicing that occurs must be extraordinary if errors are not to be introduced into the mature mRNA. Equally noteworthy is the difference between the size of a typical gene and the size of the final mRNA transcribed from it once introns are removed. As shown in **Table 12.7**, only about 15 percent of the collagen gene consists of exons that finally appear in mRNA. For other proteins, an even more extreme picture emerges. Only about 8 percent of the albumin gene remains to be translated, and in the largest human gene known, dystrophin (which is the protein product absent in Duchenne muscular dystrophy), less than 1 percent of the gene sequence is retained in the mRNA. Two other human genes are also contrasted in Table 12.7.

Although the vast majority of eukaryotic genes examined thus far contain introns, there are several exceptions. Notably, the genes coding for histones and for interferon appear to contain no introns. It is not clear why or how the genes encoding these molecules have been maintained throughout evolution without acquiring the extraneous information characteristic of almost all other genes.

ESSENTIAL POINT

The primary transcript in eukaryotes reflects the presence of intervening sequences, or introns, present in DNA, which must be spliced out to create the mature mRNA. ■

Splicing Mechanisms: Self-Splicing RNAs

The discovery of split genes led to intensive attempts to elucidate the mechanism by which introns of RNA are excised and exons are spliced back together. A great deal of progress has already been made, relying heavily on *in vitro* studies. Interestingly, it appears that somewhat different

TABLE 12.7 Comparing Human Gene Size, mRNA Size, and the Number of Introns

Gene	Gene Size (kb)	mRNA Size (kb)	Number of Introns
Insulin	1.7	0.4	2
Collagen [*pro-α-2(1)*]	38.0	5.0	51
Albumin	25.0	2.1	14
Phenylalanine hydroxylase	90.0	2.4	12
Dystrophin	2000.0	17.0	79

mechanisms exist for different types of RNA, as well as for RNAs produced in mitochondria and chloroplasts.

We might envision the simplest possible mechanism for removing an intron to be as illustrated in Figure 12–9. After an endonucleolytic "cut" is made at each end of an intron, the intron is removed, and the terminal ends of the adjacent exons are ligated by an enzyme (in short, the intron is snipped out, and the exon ends are rejoined). This is apparently what happens to the introns present in transfer RNAs (tRNAs) in bacteria. A specific endonuclease recognizes the intron termini and excises the intervening sequences. Then RNA ligase seals the exon ends to complete each splicing event. However, in the studies of all other RNAs—tRNA in higher eukaryotes and rRNAs and pre-mRNAs in all eukaryotes—precise excision of introns is much more complex and a much more interesting story.

Introns in eukaryotes can be categorized into several groups based on their splicing mechanisms. Group I, represented by introns that are part of the primary transcript of rRNAs, require no additional components for intron excision; the intron itself is the source of the enzymatic activity necessary for removal. This amazing discovery was made in 1982 by Thomas Cech and his colleagues during a study of the ciliate protozoan *Tetrahymena.* RNAs that are capable of catalytic activity are referred to as **ribozymes.** The self-excision process for group I introns serves to illustrate this concept and is shown in **Figure 12–12**. Chemically, two nucleophilic reactions take place—that is, reactions caused by the presence of electron-rich chemical species (in this case, they are *transesterification reactions*). The first is an interaction between guanosine, which acts as a cofactor in the reaction, and the primary transcript [Figure 12–12(a)]. The 3′-OH group of guanosine is transferred to the nucleotide adjacent to the 5′-end of the intron [Figure 12–12(b) and Figure 12–12(c)]. The second reaction involves the interaction of the newly acquired 3′-OH group on the left-hand exon and the phosphate on the 3′-end of the intron [Figure 12–12(c)]. The intron is spliced out and the two exon regions are ligated, leading to the mature RNA [Figure 12–12(d)].

Self-excision of group I introns, as described above, is now known to apply to pre-rRNAs from other protozoans

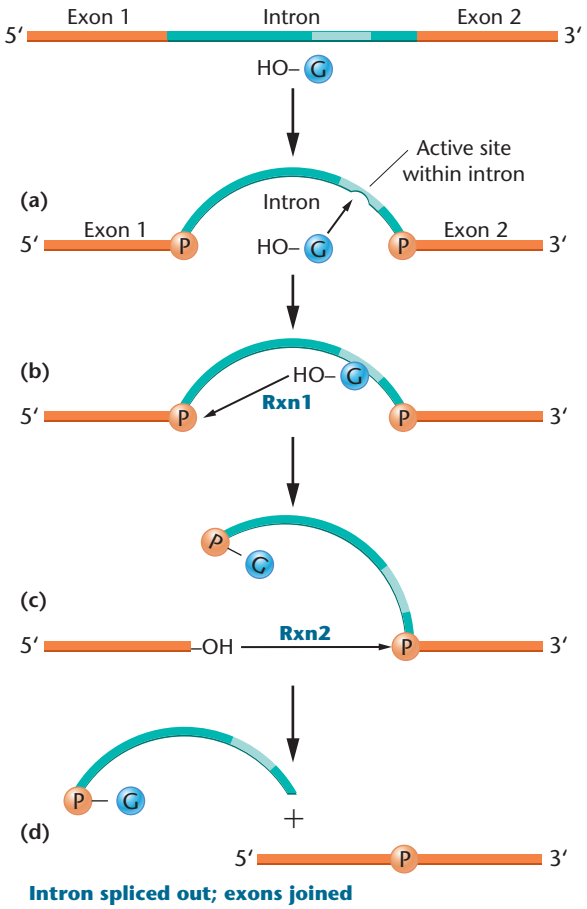


FIGURE 12–12 Splicing mechanism of pre-rRNA involving group I introns that are removed from the initial transcript. The process is one of self-excision involving two transesterification reactions.

besides *Tetrahymena*. Self-excision also seems to govern the removal of introns from the primary mRNA and tRNA transcripts produced in mitochondria and chloroplasts. These are referred to as group II introns. As in group I molecules, splicing here involves two autocatalytic reactions leading to the excision of introns. However, guanosine is not involved as a cofactor with group II introns.

Splicing Mechanisms: The Spliceosome

Introns are a major component of nuclear-derived pre-mRNA transcripts of eukaryotes. Compared to the group I and group II introns discussed above, those in nuclear-derived mRNA can be much larger—up to 20,000 nucleotides—and they are more plentiful. Their removal appears to require a much more complex mechanism. Nevertheless, we now have a good handle on the process.

Interestingly, the splicing reactions are mediated by a huge molecular complex called a **spliceosome,** which has now been identified in extracts of yeast as well as in mammalian cells. This structure is very large, 40*S* in yeast and 60*S* in mammals, being the same size as ribosomal subunits!

One set of essential components of spliceosomes is a unique set of small nuclear RNAs (snRNAs). These RNAs are usually 100 to 200 nucleotides long or less and are complexed with proteins to form small nuclear ribonucleoproteins (snRNPs or snurps). Because they are rich in uridine residues, the snRNAs have been arbitrarily designated U1, U2, . . . , U6.

Figure 12–13 depicts a model illustrating the steps involved in the removal of one intron. Keep in mind that

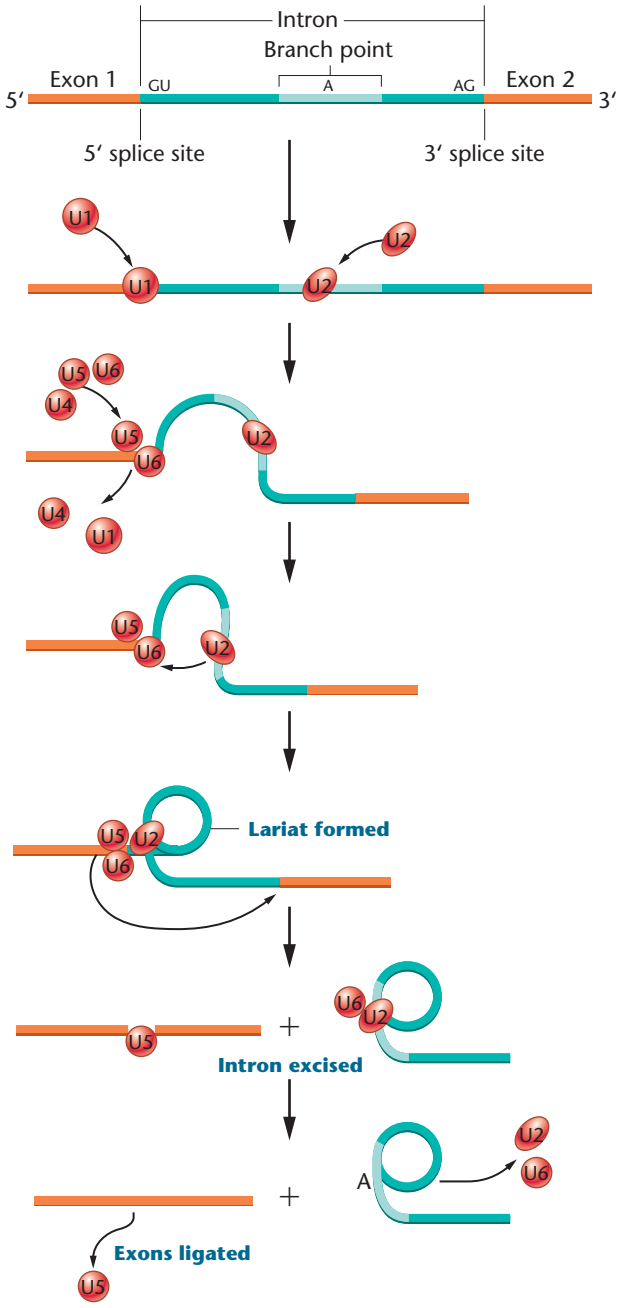


FIGURE 12–13 A model of the splicing mechanism involved during the removal of an intron from a pre-mRNA. Excision is dependent on various snRNAs (U1, U2, . . . , U6) that combine with proteins to form snRNPs., which are part of the spliceosome. The lariat structure in the intermediate stage is characteristic of this mechanism.

while this figure shows separate components, the process involves the huge spliceosome that envelopes the RNA being spliced. The nucleotide sequences near the ends of the intron begin at the 5′-end with a GU dinucleotide sequence, called the *donor sequence,* and terminate at the 3′-end with an AG dinucleotide, called the *acceptor sequence.* These, as well as other consensus sequences shared by introns, attract specific snRNAs of the spliceosome. For example, the snRNA U1 bears a nucleotide sequence that is complementary to the 5′-donor sequence end of the intron. Base pairing resulting from this homology promotes the binding that represents the initial step in the formation of the spliceosome. After the other snRNPs (U2, U4, U5, and U6) are added, splicing commences. As with group I splicing, two transesterification reactions occur. The first involves the interaction of the 3′-OH group from an adenine (A) residue present within the branch point region of the intron. The A residue attacks the 5′-splice site, cutting the RNA chain. In a subsequent step involving several other snRNPs, an intermediate structure is formed and the second reaction ensues, linking the cut 5′-end of the intron to the A. This results in the formation of a characteristic loop structure called a *lariat,* which contains the excised intron. The exons are then ligated and the snRNPs are released.

The processing involved in splicing, which occurs within the nucleus, represents a potential regulatory step in gene expression in eukaryotes. For instance, several cases are known wherein introns present in pre-mRNAs *derived from the same gene* are spliced *in more than one way*, thereby yielding different collections of exons in the mature mRNA. Such alternative splicing yields a group of similar but nonidentical mRNAs that, upon translation, result in a series of related proteins called isoforms. Many examples have been encountered in organisms ranging from viruses to *Drosophila* to humans. Alternative splicing of pre-mRNAs represents a way of producing related proteins from a single gene, increasing the number of gene products that can be derived from an organism's genome. We will return to this topic in Chapter 15 in our discussion of the regulation of gene expression in eukaryotes.

EVOLVING CONCEPT OF THE GENE

The elucidation of the genetic code in the 1960s supported the concept that the gene is composed of a linear series of triplet nucleotides encoding the amino acid sequence of a protein. While this is indeed the case in prokaryotes and viruses, in 1977, it became apparent that in eukaryotes, the gene is divided into coding sequences, called exons, which are interrupted by noncoding sequences, called introns (intervening sequences), which must be spliced out during production of the mature mRNA). ■

12.12 RNA Editing May Modify the Final Transcript

In the late 1980s, still another unexpected form of post-transcriptional RNA processing was discovered in several organisms. In this form, referred to as **RNA editing,** the nucleotide sequence of a pre-mRNA is actually changed prior to translation. As a result, the ribonucleotide sequence of the mature RNA differs from the sequence encoded in the exons of the DNA from which the RNA was transcribed.

Although other variations exist, there are two main types of RNA editing: **insertion/deletion editing,** in which nucleotides are added to or subtracted from the total number of bases; and **substitution editing,** in which the identities of individual nucleotide bases are altered. Substitution editing is used in some nuclear-derived eukaryotic RNAs and is prevalent in mitochondrial and chloroplast RNAs transcribed in plants.

Trypanosoma, a parasite that causes African sleeping sickness, and its relatives use extensive insertion/deletion editing in mitochondrial RNAs. The uridines added to an individual transcript can make up more than 60 percent of the coding sequence, usually forming the initiation codon and bringing the rest of the sequence into the proper reading frame. Insertion/deletion editing in trypanosomes is directed by **gRNA (guide RNA)** templates, which are also transcribed from the mitochondrial genome. These small RNAs share a high degree of complementarity to the edited region of the final mRNAs. They base-pair with the pre-edited mRNAs to direct the editing machinery to make the correct changes.

An excellent example of substitutional editing involves the subunits constituting the *glutamate receptor channels* (*GluR*) in mammalian brain tissue. In this case, adenosine (A) to inosine (I) editing occurs in pre-mRNAs prior to their translation, during which I is read as guanosine (G). A family of three ADAR (*a*denosine *d*eaminase *a*cting on *R*NA) enzymes is believed to be responsible for the editing of various sites within the glutamate channel subunits. The double-stranded RNAs required for editing by the ADAR enzymes are provided by intron/exon pairing of the GluR mRNA transcripts. The editing changes alter the physiological parameters (solute permeability and desensitization response time) of the receptors containing the subunits.

Findings such as these have established that RNA editing provides still another important mechanism of *post-transcriptional modification*, and that this process is not restricted to small or asexually reproducing genomes, such as those in mitochondria. These discoveries have important implications for the regulation of genetic expression.

12.13 Transcription Has Been Visualized by Electron Microscopy

We conclude our coverage of transcription by referring you back to the chapter opening photograph (p. 231), which is a striking visualization of transcription occurring in the oocyte nucleus of *Xenopus laevis,* the clawed frog. Note the central axis that runs horizontally from left to right and from which threads appear to be emanating vertically. This axis, appearing as a thin thread, is the DNA of most of one gene encoding ribosomal RNA (rDNA). Each of the emanating threads, which grows longer the farther to the right it is found, is an rRNA molecule being transcribed. What is apparent is that multiple copies of RNA polymerase have initiated transcription at a point near the left end and that transcription by each of them has proceeded to the right. Simultaneous transcription by many of these polymerases results in the electron micrograph that has captured an image of the entire process.

It is fascinating to visualize the process and to confirm our expectations based on the biochemical analysis of this process.

GENETICS, TECHNOLOGY, AND SOCIETY

Fighting Disease with Antisense Therapeutics

Conventional therapeutic drugs often have toxic side effects, affecting both normal and diseased cells, with diseased or infected cells being only slightly more susceptible than the patient's normal cells. Scientists have long wished for a magic bullet that could seek out and destroy the underlying defects in diseased cells, leaving normal cells alive and healthy. Recently, a group of promising drugs has emerged, collectively described as *antisense* drugs.

Antisense therapies are based on the design and use of short synthetic single-stranded DNA molecules known as *antisense oligonucleotides* (*ASOs*) and have been developed through an understanding of the molecular biology of gene expression. When cells become infected with viruses, are transformed into cancer cells, or suffer from many genetic defects, their underlying disease state is determined by gene mutations or inappropriate gene expression. The mRNAs responsible for these conditions are transcribed from the template strand of DNA and therefore are known as "*sense*" RNAs. If scientists know the sequence of the sense RNA, they can synthesize a strand of DNA that is complementary to any sequence within the sense RNA. This complementary DNA is the antisense oligonucleotide. When introduced into cells, ASOs bind to the specific target sense RNA through Watson–Crick base-pairing. The binding of ASOs to sense RNAs interferes with the mRNA's translation, or causes its degradation.

The antisense approach is exciting because of its potential specificity. Because an antisense molecule has a sequence that specifically binds to a particular sense RNA, it should be possible to inhibit synthesis of the specific protein encoded by the sense RNA. If the encoded protein causes a genetic disease or is necessary for virus reproduction or cancer cell growth—but is not necessary in normal cells—the antisense molecule should have only therapeutic effects.

In ASO-based technologies, scientists design short (about 20 nucleotides long) single-stranded antisense DNA oligonucleotides and then synthesize large amounts of these ASOs *in vitro*. Using various methods, the ASOs can be introduced into cultured cells or whole organisms.

Although the development of effective ASO drugs has been slow and often difficult, two have been approved by the FDA and are currently on the market. One of these drugs, called fomivirsen or Vitravene, targets cytomegalovirus infections in the eye. Another drug, called mipomersen or Kynamro, inhibits the synthesis of apolipoprotein B and is used to lower the levels of cholesterol lipoproteins in patients with familial hypercholesterolemia.

More than 20 ASO antisense drugs are currently in phase II and III clinical trials. These drugs are designed to treat various cancers, asthma, ulcerative colitis, renal failure, and virus infections.

Another new and promising ASO therapeutic approach involves using antisense oligonucleotides to alter mRNA splicing patterns, a therapy called "*antisense-mediated exon skipping*." The most advanced exon-skipping therapy, now in clinical trials, is designed to treat patients with Duchenne muscular dystrophy (DMD). DMD is an X-linked recessive disorder, affecting approximately one in every 3500 newborn males. It results in muscle degeneration, heart disease, and premature death. The disease is caused by mutations in the *dystrophin* gene, which contains 79 exons. These mutations are a variety of deletions, duplications, or other mutations that disrupt the open reading frame of the gene and result in premature translation termination and the absence of functional dystrophin protein.

Many dystrophin gene mutations occur in exon 51. Scientists have designed ASOs that target DNA sequences near the splice junctions of exon 51. These ASOs interfere with normal pre-mRNA splicing, causing exon 50 to be spliced to exon 52. This results in an internal deletion of the mutated exon 51 in the mRNA, but restores the correct reading frame to the internally deleted protein. Although missing some internal amino acid sequences, the new dystrophin protein restores partial function to the muscles. The ASO-mediated exon- skipping approach is also

being applied to treatments for other conditions such as β-thalassemia, spinal muscular atrophy, and some cancers.

Your Turn

Take time, individually or in groups, to answer the following questions. Investigate the references and links to help you discuss some of the technical issues that surround the development and uses of antisense therapies.

1. What are some of the challenges in the use of ASOs as therapeutics? How are scientists addressing these challenges?

A discussion of antisense oligonucleotide drug design is presented in: Watts, J.K. and Corey, D.R. 2012. Silencing disease genes in the laboratory and the clinic. *J. Pathol*. 226: 365-379.

2. What are some limitations in the use of ASOs for exon-skipping therapies in Duchenne muscular dystrophy? What is the status of current clinical trials for these exon-skipping therapies?

A discussion of exon skipping in DMD is presented in: Fairclough, R.J. et al. 2013. Therapy for Duchenne muscular dystrophy: Renewed optimism from genetic approaches. *Nature Rev. Genet*. 14: 373-378.

For information about clinical trials, start your search on PubMed at http://www.ncbi.nlm.nih.gov/pubmed.

3. Mipomersen (Kynamro, Isis Pharmaceuticals) has recently been approved for use to treat familial hypercholesterolemia. How does this ASO work, and why was its use restricted to patients with a homozygous hereditary disease?

To learn about mipomersen, see Bell, D.A. et al. 2012. Mipomersen and other therapies for the treatment of severe familial hypercholesterolemia. *Vascular Health and Risk Management* 8: 651–659.

CASE STUDY | Cystic fibrosis

A two-year-old girl, with poor growth and frequent lung infections, was diagnosed with cystic fibrosis (CF). CF is an autosomal recessive genetic disease caused by mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene. The CFTR gene is 27-exons and 250,000-nucleotides long, and codes for a transmembrane chloride ion transporter. In the girl's case, the CF was caused by two mutations—a frameshift mutation in exon 23, leading to a shortened, non-functioning protein, and a 3-bp deletion of an amino acid (phenylalanine) at position 508 of the protein sequence, resulting in a protein-folding defect. Without the CFTR protein, the chloride ions could not cross the membranes, which caused accumulation of thick mucus in her lungs and inhibited absorption of food.

1. Why would a frameshift mutation result in a shortened protein? Why would it be non-functional?
2. Why was the girl's CF caused by two different mutations?
3. If the parents of the girl give birth to a son, what is the likelihood of the boy having CF?

INSIGHTS AND SOLUTIONS

1. Calculate how many triplet codons would be possible had evolution seized on six bases (three complementary base pairs) rather than four bases within the structure of DNA. Would six bases accommodate a two-letter code, assuming 20 amino acids and start and stop codons?

   Solution: Six things taken three at a time will produce $(6)^3$ or 216 triplet codes. If the code was a doublet, there would be $(6)^2$ or 36 two-letter codes, more than enough to accommodate 20 amino acids and start and stop signals.

2. In a heteropolymer experiment using 1/2C:1/4A:1/4G, how many different triplets will occur in the synthetic RNA molecule? How frequently will the most frequent triplet occur?

   Solution: There will be $(3)^3$ or 27 triplets produced. The most frequent will be CCC, present $(1/2)^3$ or 1/8 of the time.

3. In a regular copolymer experiment, where UUAC is repeated over and over, how many different triplets will occur in the synthetic RNA, and how many amino acids will occur in the polypeptide when this RNA is translated? (Consult Figure 12–7.)

   Solution: The synthetic RNA will repeat four triplets—UUA, CUU, ACU, UAC—over and over. Because both UUA and CUU encode leucine, while ACU and UAC encode threonine and tyrosine, respectively, the polypeptides synthesized under the directions of this RNA would contain three amino acids in the repeating sequence leu-leu-thr-tyr.

4. Actinomycin D inhibits DNA-dependent RNA synthesis. This antibiotic is added to a bacterial culture where a specific protein is being monitored. Compared to a control culture, where no antibiotic is added, translation of the protein declines over a period of 20 minutes, until no further protein is made. Explain these results.

   Solution: The mRNA, which is the basis for translation of the protein, has a lifetime of about 20 minutes. When actinomycin D is added, transcription is inhibited and no new mRNAs are made. Those already present support the translation of the protein for up to 20 minutes.

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we focused on the genetic code and the transcription of genetic information stored in DNA into complementary RNA molecules. Along the way, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions:
   (a) How did we determine the *compositions* of codons encoding specific amino acids?
   (b) How were the specific sequences of triplet codes determined experimentally?
   (c) How were the experimentally derived triplet codon assignments verified in studies using bacteriophage MS2?
   (d) How do we know that mRNA exists and serves as an intermediate between information encoded in DNA and its concomitant gene product?
   (e) How do we know that the initial transcript of a eukaryotic gene contains noncoding sequences that must be removed before accurate translation into proteins can occur?

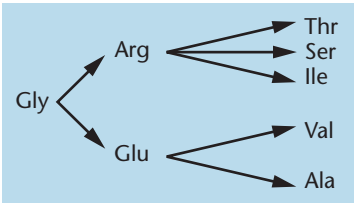
CONCEPT QUESTION

2. Review the Chapter Concepts list on p. 231. These all center on how genetic information is stored in DNA and transferred to RNA prior to translation into proteins. Write a short essay that summarizes the key properties of the genetic code and the process by which RNA is transcribed on a DNA template. ■

3. In studies of frameshift mutations, Crick, Barnett, Brenner, and Watts–Tobin found that either three nucleotide insertions or deletions restored the correct reading frame. (a) Assuming the code is a triplet, what effect would the addition or loss of six nucleotides have on the reading frame? (b) If the code were a sextuplet (consisting of six nucleotides), would the reading frame be restored by the addition or loss of three, six, or nine nucleotides?
4. When the repeating trinucleotide sequence UUCUUCUUC is added to a cell-free translation system, three different polypeptide homopolymers are produced. Why?
5. From the late 1950s to the mid-1960s, numerous experiments using *in vitro* cell-free systems provided information on the nature of the genetic code. Briefly outline significant experiments in the determination of the genetic code.
6. In a coding experiment using repeating copolymers (as shown in Table 12.3), the following data were obtained.

Copolymer	Codons Produced	Amino Acids in Polypeptide
AG	AGA, GAG	arg, glu
AAG	AGA, AAG, GAA	lys, arg, glu

   AGG is known to code for arginine. Taking into account the wobble hypothesis, assign each of the four remaining different triplet codes to its correct amino acid.
7. "Breaking the genetic code" has been referred to as one of the most significant scientific achievements in modern times. Describe (in outline or brief statement form) the procedures used to break the code.
8. When the amino acid sequences of insulin isolated from different organisms were determined, some differences were noted. For example, alanine was substituted for threonine, serine was substituted for glycine, and valine was substituted for isoleucine at corresponding positions in the protein. List the single-base changes that could occur in triplets to produce these amino acid changes.
9. In studies of the amino acid sequence of wild-type and mutant forms of tryptophan synthetase in *E. coli,* the following changes have been observed:



   Determine a set of triplet codes in which only a single-nucleotide change produces each amino acid change.
10. Why doesn't polynucleotide phosphorylase (Ochoa's enzyme) synthesize RNA *in vivo*?
11. Refer to Table 12.1. Can you hypothesize why a mixture of (Poly U) + (Poly A) would not stimulate incorporation of ^{14}C-phenylalanine into protein?
12. Predict the amino acid sequence produced during translation of the short theoretical mRNA sequences below. (Note that the second sequence was formed from the first by a deletion of only one nucleotide.) What type of mutation gave rise to sequence 2?

    Sequence 1: 5′-AUGCCGGAUUAUAGUUGA-3′

    Sequence 2: 5′-AUGCCGGAUUAAGUUGA-3′

13. A short RNA molecule was isolated that demonstrated a hyperchromic shift indicating secondary structure (see p. 191 in Chapter 9). Its sequence was determined to be

    5′-AGGCGCCGACUCUACU-3′

    (a) Propose a two-dimensional model for this molecule.
    (b) What DNA sequence would give rise to this RNA molecule through transcription?
    (c) If the molecule were a tRNA fragment containing a CGA anticodon, what would the corresponding codon be?
    (d) If the molecule were an internal part of a message, what amino acid sequence would result from it following translation? (Refer to the code chart in Figure 12–7.)
14. An alanine residue exists at position 180 of a certain plant protein. If the codon specifying alanine is GCU, how many single-base substitutions will result in an amino acid substitution at position 180, and what are they?
15. Shown here is a theoretical viral mRNA sequence

    5′-AUGCAUACCUAUGAGACCCUUGGA-3′

    (a) Assuming that it could arise from overlapping genes, how many different polypeptide sequences can be produced? Using the chart in Figure 12–7, what are the sequences?
    (b) A base-substitution mutation that altered the sequence in part (a) eliminated the synthesis of all but one polypeptide. The altered sequence is shown below. Use Figure 12–7 to determine why it was altered.

    5′-AUGCAUACCUAUGUGACCCUUGGA-3′

16. A novel protein discovered in a certain plant has many leucine-rich regions, fewer alanine-rich regions, and even fewer tyrosine residues. Correlate the number of codons for these three amino acids with this information.
17. Suppose that in the use of polynucleotide phosphorylase, nucleotides A and C are added in a ratio of 1A:5C. What is the probability that an AAA sequence will occur?
18. Describe the structure of RNA polymerase in bacteria. What is the core enzyme? What is the role of the σ subunit?
19. Illustrating the importance of triphosphate and monophosphate molecules, explain the process of RNA biosynthesis by RNA polymerase.
20. Sydney Brenner argued that the code was nonoverlapping because he considered that coding restrictions would occur if it were overlapping. A second major argument against an overlapping code involved the effect of a single nucleotide change. In an overlapping code, how many adjacent amino acids would be affected by a point mutation? In a nonoverlapping code, how many amino acid(s) would be affected?
21. One form of posttranscriptional modification of most eukaryotic RNA transcripts is the addition of a poly-A sequence at the 3′-end. The absence of a poly-A sequence leads to rapid degradation of the transcript. Poly-A sequences of various lengths are also added to many prokaryotic RNA transcripts where, instead of promoting stability, they enhance degradation. In both cases, RNA secondary structures, stabilizing proteins, or degrading enzymes interact with poly-A sequences. Considering the activities of RNAs, what might be the general functions of 3′-polyadenylation?
22. In a mixed copolymer experiment, messages were created with either 4/5C:1/5A or 4/5A:1/5C. These messages yielded proteins with the amino acid compositions shown in the following table. Using these data, predict the most specific *coding composition* for each amino acid.

4/5C:1/5A		4/5A:1/5C	
Proline	63.0%	Proline	3.5%
Histidine	13.0%	Histidine	3.0%
Threonine	16.0%	Threonine	16.6%
Glutamine	3.0%	Glutamine	13.0%
Asparagine	3.0%	Asparagine	13.0%
Lysine	0.5%	Lysine	50.0%
	98.5%		99.1%

23. Shown in this problem are the amino acid sequences of the wild type and three mutant forms of a short protein.
(a) Using Figure 12–7, predict the type of mutation that created each altered protein.
(b) Determine the specific ribonucleotide change that led to the synthesis of each mutant protein.
(c) The wild-type RNA consists of nine triplets. What is the role of the ninth triplet?
(d) For the first eight wild-type triplets, which, if any, can you determine specifically from an analysis of the mutant proteins? In each case, explain why or why not.
(e) Another mutation (mutant 4) is isolated. Its amino acid sequence is unchanged, but mutant cells produce abnormally low amounts of the wild-type proteins. As specifically as you can, predict where this mutation exists in the gene.

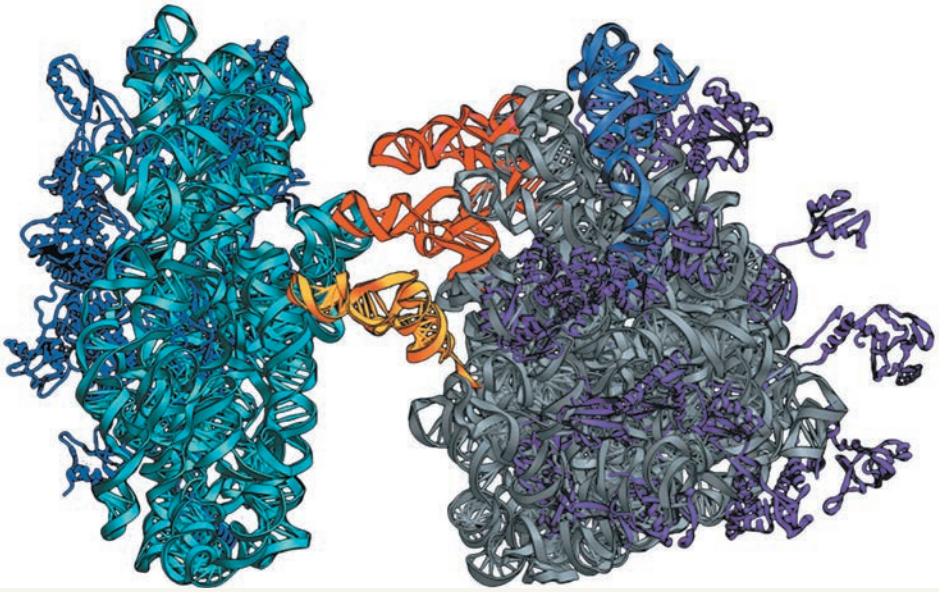
Wild type:	met-trp-tyr-arg-gly-ser-pro-thr
Mutant 1:	met-trp
Mutant 2:	met-trp-his-arg-gly-ser-pro-thr
Mutant 3:	met-cys-ile-val-val-val-gln-his

24. Alternative splicing is a common mechanism for eukaryotes to expand their repertoire of gene functions. Studies by Xu and colleagues (2002. *Nuc. Acids Res.* 30: 3754–3766) indicate that approximately 50 percent of human genes use alternative splicing, and approximately 15 percent of disease-causing mutations involve aberrant alternative splicing. Different tissues show remarkably different frequencies of alternative splicing, with the brain accounting for approximately 18 percent of such events.
(a) Define alternative splicing and speculate on the evolutionary strategy alternative splicing offers to organisms.
(b) Why might some tissues engage in more alternative splicing than others?
25. The genetic code is degenerate. Amino acids are encoded by either 1, 2, 3, 4, or 6 triplet codons. (See Figure 12–7.) An interesting question is whether the number of triplet codes for a given amino acid is in any way correlated with the frequency with which that amino acid appears in proteins. That is, is the genetic code optimized for its intended use? Some approximations of the frequency of appearance of nine amino acids in proteins in *E. coli* are

Amino Acid	Percentage
Met	2
Cys	2
Gln	5
Pro	5
Arg	5
Ile	6
Glu	7
Ala	8
Leu	10

(a) Determine how many triplets encode each amino acid.
(b) Devise a way to graphically compare the two sets of information (data).
(c) Analyze your data to determine what, if any, correlations can be drawn between the relative frequency of amino acids making up proteins and the number of codons for each. Write a paragraph that states your specific and general conclusions.
(d) How would you proceed with your analysis if you wanted to pursue this problem further?

13 Translation and Proteins



Crystal structure of a *Thermus thermophilus* 70*S* ribosome containing three bound transfer RNAs.

CHAPTER CONCEPTS

- The ribonucleotide sequence of messenger RNA (mRNA) reflects genetic information stored in DNA that makes up genes and corresponds to the amino acid sequences in proteins encoded by those genes.
- The process of translation decodes the information in mRNA, leading to the synthesis of polypeptide chains.
- Translation involves the interactions of mRNA, tRNA, ribosomes, and a variety of translation factors essential to the initiation, elongation, and termination of the polypeptide chain.
- Proteins, the final product of most genes, achieve a three-dimensional conformation that is based on the primary amino acid sequences of the polypeptide chains making up each protein.
- The function of any protein is closely tied to its three-dimensional structure, which can be disrupted by mutation.

In Chapter 12, we established that a genetic code stores information in the form of triplet nucleotides in DNA and that this information is initially expressed through the process of transcription into a messenger RNA that is complementary to one strand of the DNA helix. However, in most instances, the final product of gene expression is a polypeptide chain consisting of a linear series of amino acids whose sequence has been prescribed by the genetic code. In this chapter, we will examine how the information present in mRNA is utilized to create polypeptides, which then fold into protein molecules. We will also review the evidence confirming that proteins are the end products of gene expression, and we will briefly discuss the various levels of protein structure, diversity, and function. This information extends our understanding of gene expression and provides an important foundation for interpreting how the mutations that arise in DNA can result in the diverse phenotypic effects observed in organisms.

13.1 Translation of mRNA Depends on Ribosomes and Transfer RNAs

Translation of mRNA is the biological polymerization of amino acids into polypeptide chains. This process, alluded to in our discussion of the genetic code in Chapter 12, occurs only in association with ribosomes, which serve as nonspecific workbenches. The central question in translation is how triplet ribonucleotides of mRNA direct specific amino acids into their correct position in the polypeptide. This question was answered once **transfer RNA (tRNA)** was discovered. This class of molecules adapts specific triplet codons in mRNA to their correct amino acids. The *adaptor hypothesis* for the role of tRNA was postulated by Francis Crick in 1957.

In association with a ribosome, mRNA presents a triplet codon that calls for a specific amino acid. A specific tRNA molecule contains within its nucleotide sequence three consecutive ribonucleotides complementary to the codon, called the **anticodon,** which can base-pair with the codon. Another region of this tRNA is covalently bonded to its corresponding amino acid.

Hydrogen bonding of tRNAs to mRNA holds the amino acids in proximity so that a peptide bond can be formed. This process occurs over and over as mRNA runs through the ribosome and amino acids are polymerized into a polypeptide. Before we discuss the actual process of translation, let's first consider the structures of the ribosome and tRNA.

Ribosomal Structure

Because of its essential role in the expression of genetic information, the **ribosome** has been extensively analyzed. One bacterial cell contains about 60,000 ribosomes, and a eukaryotic cell contains many times more. Electron microscopy reveals that the bacterial ribosome is about 40 nm at its largest dimension and consists of two subunits, one large and one small. Both subunits consist of one or more molecules of rRNA and an array of **ribosomal proteins.** When the two subunits are associated with each other in a single ribosome, the structure is sometimes called a **monosome.**

The main differences between prokaryotic and eukaryotic ribosomes are summarized in **Figure 13–1**. The subunit and rRNA components are most easily isolated and characterized on the basis of their sedimentation behavior in sucrose gradients (their rate of migration, or *Svedberg coefficient S,* which is a reflection of their density, mass, and shape). In prokaryotes, the monosome is a 70*S* particle; in eukaryotes, it is approximately 80*S*. Sedimentation coefficients, which reflect the variable rate of migration of different-sized particles and molecules, are not additive. For example, the prokaryotic 70*S* monosome consists of a 50*S* and a 30*S* subunit, and the eukaryotic 80*S* monosome consists of a 60*S* and a 40*S* subunit.

The larger subunit in prokaryotes consists of a 23*S* rRNA molecule, a 5*S* rRNA molecule, and 31 ribosomal proteins. In the eukaryotic equivalent, a 28*S* rRNA molecule is accompanied by a 5.8*S* and 5*S* rRNA molecule and

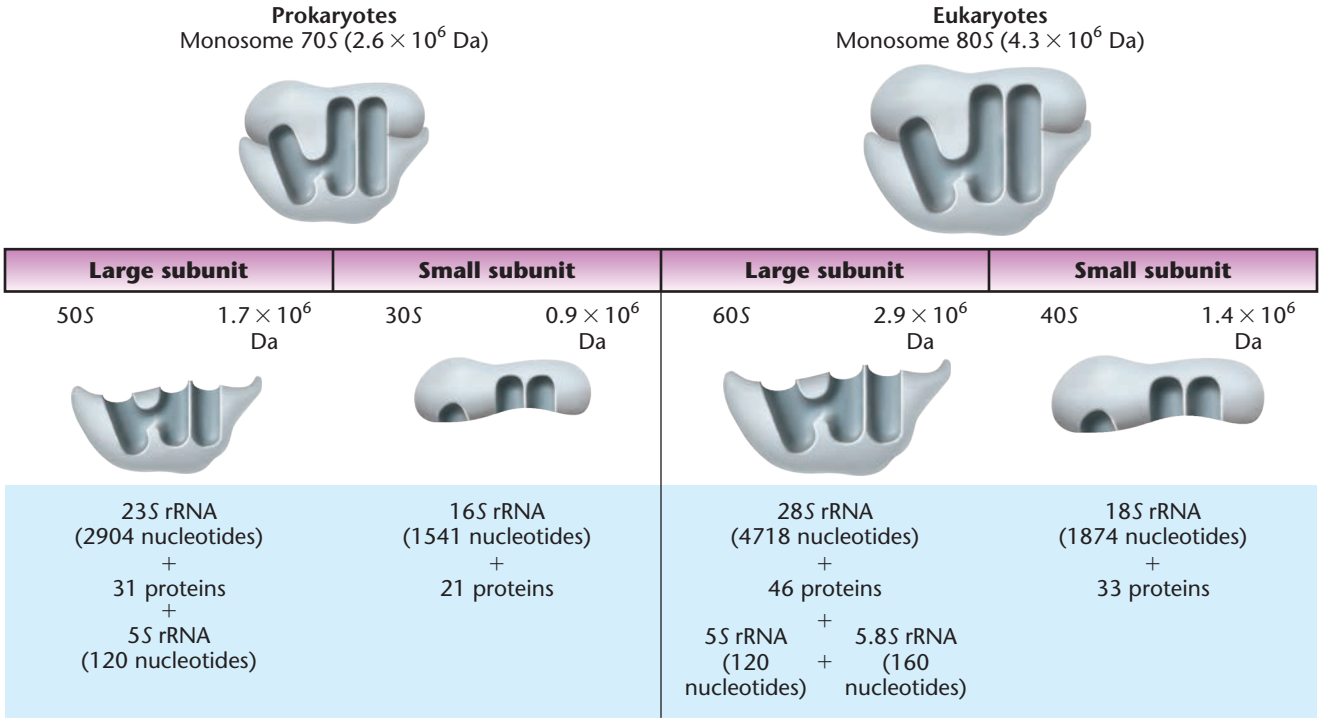


FIGURE 13–1 A comparison of the components of prokaryotic and eukaryotic ribosomes.

46 proteins. The smaller prokaryotic subunits consist of a 16*S* rRNA component and 21 proteins. In the eukaryotic equivalent, an 18*S* rRNA component and 33 proteins are found. The approximate molecular weights (MWs) and the number of nucleotides of these components are also shown in Figure 13–1.

It is now clear that the RNA molecules perform the all-important catalytic functions associated with translation. The many proteins, whose functions were long a mystery, are thought to promote the binding of the various molecules involved in translation and, in general, to fine-tune the process. This conclusion is based on the observation that some of the catalytic functions in ribosomes still occur in experiments involving "ribosomal protein-depleted" ribosomes.

Molecular hybridization studies have established the degree of redundancy of the genes coding for the rRNA components. The *E. coli* genome contains seven copies of a single sequence that encodes all three components—23*S,* 16*S,* and 5*S*. The initial transcript of each set of these genes produces a 30*S* RNA molecule that is enzymatically cleaved into these smaller components. Coupling of the genetic information encoding these three rRNA components ensures that after multiple transcription events, equal quantities of all three will be present as ribosomes are assembled.

In eukaryotes, many more copies of a sequence encoding the 28*S,* 18*S,* and 5.8*S* components are present. In *Drosophila,* approximately 120 copies per haploid genome are each transcribed into a molecule of about 34*S*. This molecule is then processed into the 28*S,* 18*S,* and 5.8*S* rRNA species. These species are homologous to the three rRNA components of *E. coli.* In *Xenopus laevis,* over 500 copies of the 34*S* component are present per haploid genome. In mammalian cells, the initial transcript is even larger at 45*S*.

The rRNA genes, called **rDNA,** are part of the moderately repetitive DNA fraction and are present in clusters at various chromosomal sites. Each cluster in eukaryotes consists of tandem repeats, with each unit separated by a noncoding spacer DNA sequence. In humans, these gene clusters have been localized near the ends of chromosomes 13, 14, 15, 21, and 22. The unique 5*S* rRNA component of eukaryotes is not part of this larger transcript. Instead, genes coding for the 5*S* ribosomal component are distinct and located separately. In humans, a gene cluster encoding 5*S* rRNA has been located on chromosome 1.

Despite their detailed knowledge of the structure and genetic origin of the ribosomal components, a complete understanding of the function of these components has eluded geneticists. This is not surprising; the ribosome is the largest and perhaps the most intricate of all cellular structures. For example, the bacterial monosome has a combined molecular weight of 2.6 million Da!

ESSENTIAL POINT

Translation is the synthesis of polypeptide chains under the direction of mRNA in association with ribosomes. ■

tRNA Structure

Because of their small size and stability in the cell, transfer RNAs (tRNAs) have been investigated extensively and are the best-characterized RNA molecules. They are composed of only 75 to 90 nucleotides, displaying a nearly identical structure in bacteria and eukaryotes. In both types of organisms, tRNAs are transcribed as larger precursors, which are cleaved into mature 4*S* tRNA molecules. In *E. coli,* for example, $tRNA^{Tyr}$ (the superscript identifies the specific tRNA and the cognate amino acid that binds to it) is composed of 77 nucleotides, yet its precursor contains 126 nucleotides.

In 1965, Robert Holley and his colleagues reported the complete sequence of $tRNA^{Ala}$ isolated from yeast. Of great interest was their finding that a number of nucleotides are unique to tRNA, containing a so-called modified base. Two of these nucleotides, inosinic acid and pseudouridylic acid, are illustrated in **Figure 13–2**. These modified structures are created *after* transcription of tRNA, illustrating the more general concept of **posttranscriptional modification.** While it is still not clear why such modified bases are created, it is believed that their presence enhances hydrogen bonding efficiency during translation.

Holley's sequence analysis led him to propose the two-dimensional **cloverleaf model of tRNA.** It was known that tRNA demonstrates a secondary structure due to base pairing. Holley discovered that he could arrange the linear model in such a way that several stretches of base pairing would result. This arrangement created a series of paired stems and unpaired loops resembling the shape of a cloverleaf. Loops consistently contained modified bases that did not generally form base pairs. Holley's model is shown in **Figure 13–3**.

The triplets GCU, GCC, and GCA specify alanine; therefore, Holley looked for an anticodon sequence complementary to one of these codons in his $tRNA^{Ala}$ molecule. He

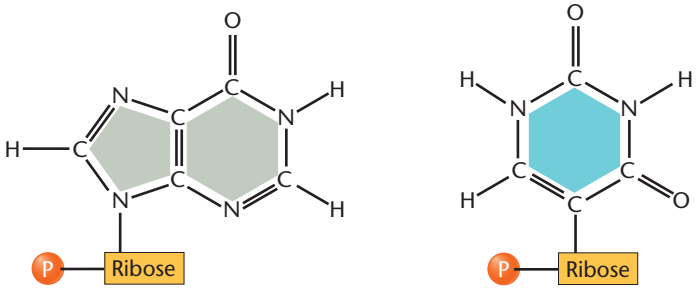


FIGURE 13–2 Ribonucleotides containing two unusual nitrogenous bases found in transfer RNA.

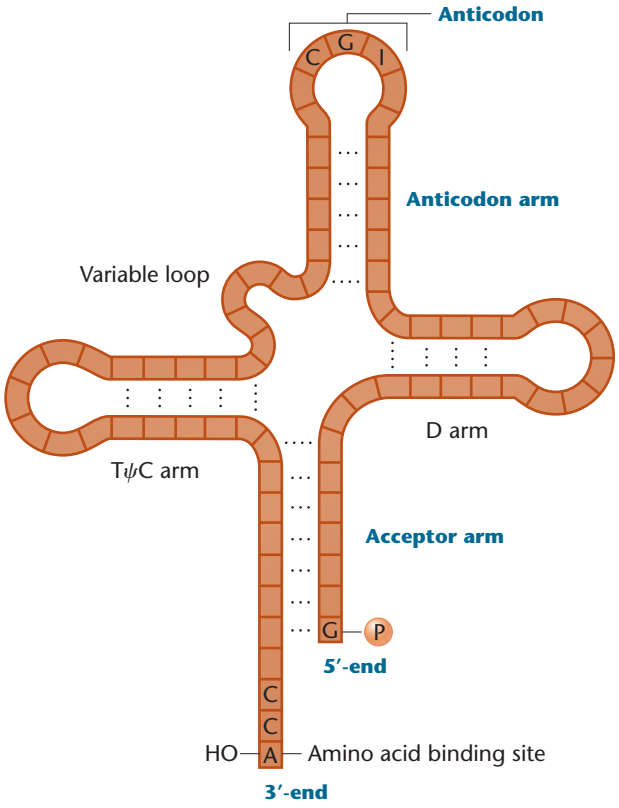


FIGURE 13–3 Holley's two-dimensional cloverleaf model of transfer RNA. Blocks represent nitrogenous bases.

found it in the form of CGI (the 3′ to 5′ direction) in one loop of the cloverleaf. The nitrogenous base I (inosinic acid) can form hydrogen bonds with U, C, or A, the third members of the alanine triplets. Thus, the **anticodon loop** was established.

Studies of other tRNA species reveal many constant features. At the 3′-end, all tRNAs contain the sequence (. . . pCpCpA-3′). This is the end of the molecule where the amino acid is covalently joined to the terminal adenosine residue. All tRNAs contain the nucleotide (5′-Gp . . .) at the other end of the molecule. In addition, the lengths of various stems and loops are very similar. Each tRNA that has been examined also contains an anticodon complementary to the known amino acid codon for which it is specific, and all anticodon loops are present in the same position of the cloverleaf.

Because the cloverleaf model was predicted strictly on the basis of nucleotide sequence, there was great interest in the X-ray crystallographic examination of tRNA, which reveals a three-dimensional structure. By 1974, Alexander Rich and his colleagues in the United States, and Jon Roberts, Brian Clark, Aaron Klug, and their colleagues in England had succeeded in crystallizing tRNA and performing X-ray crystallography at a resolution of 3 Å. At this resolution, the pattern formed by individual nucleotides is discernible.

As a result of these studies, a complete three-dimensional model of tRNA was proposed, as shown in **Figure 13–4**. At one end of the molecule is the anticodon loop and stem, and at the other end is the 3′-acceptor region where the amino acid is bound. Geneticists speculate that the shapes of the intervening loops may be recognized by the specific enzymes responsible for adding amino acids to tRNAs—a subject to which we now turn our attention.

Charging tRNA

Before translation can proceed, the tRNA molecules must be chemically linked to their respective amino acids. This activation process, called **charging,** occurs under the direction of enzymes called **aminoacyl tRNA synthetases.** There are 20 different amino acids, so there must be at least 20 different tRNA molecules and as many different enzymes. In theory, because there are 61 triplets that encode amino acids, there could be 61 specific tRNAs and enzymes. However, because of the ability of the third member of a triplet code to "wobble," it is now thought that there are only 31 different tRNAs. It is also believed that there are only 20 synthetases, one for each amino acid, regardless of the greater number of corresponding tRNAs.

The charging process is outlined in **Figure 13–5**. In the initial step, the amino acid is converted to an activated form, reacting with ATP to create an **aminoacyl-adenylic acid.** A covalent linkage is formed between the 5′-phosphate group of ATP and the carboxyl end of the amino acid. This molecule remains associated with the synthetase enzyme, forming a complex that then reacts

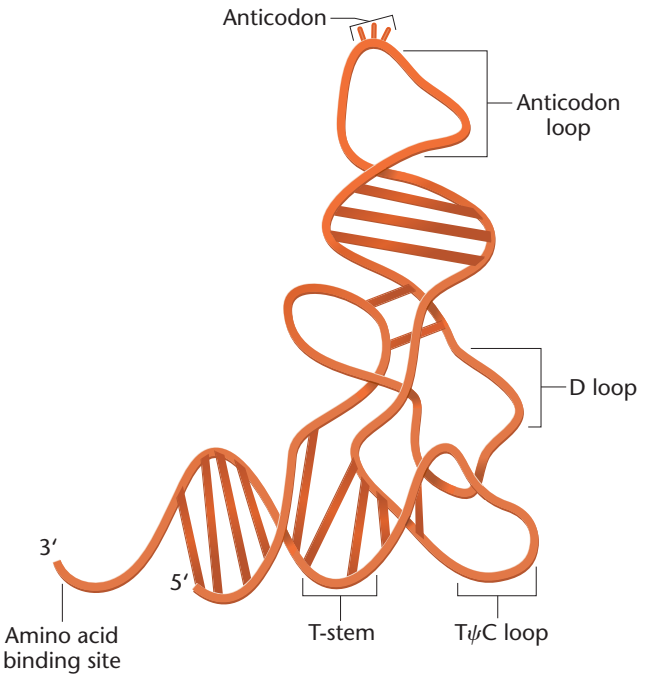


FIGURE 13–4 A three-dimensional model of transfer RNA.

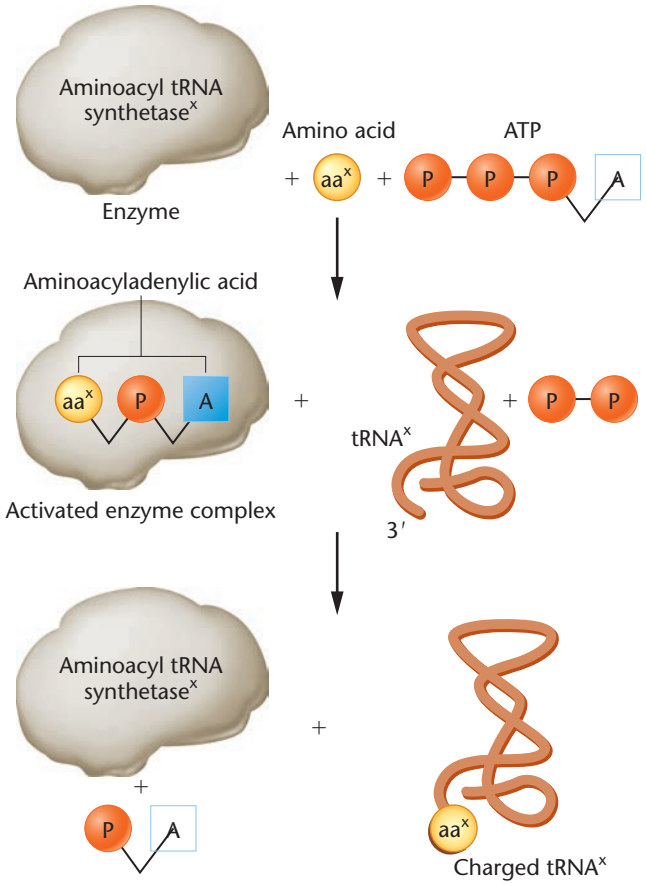


FIGURE 13–5 Steps involved in charging tRNA. The superscript x denotes that only the corresponding specific tRNA and specific aminoacyl tRNA synthetase enzyme are involved in the charging process for each amino acid.

with a specific tRNA molecule. During this next step, the amino acid is transferred to the appropriate tRNA and bonded covalently to the adenine residue at the 3′-end. The charged tRNA may now participate directly in protein synthesis. Aminoacyl tRNA synthetases are highly specific enzymes because they recognize only one amino acid and the subset of corresponding tRNAs called **isoaccepting tRNAs.** Accurate charging is crucial if fidelity of translation is to be maintained.

ESSENTIAL POINT

Translation depends on tRNA molecules that serve as adaptors between triplet codons in mRNA and the corresponding amino acids. ■

NOW SOLVE THIS

13–1 In 1962, F. Chapeville and others reported an experiment in which they isolated radioactive ^{14}C-cysteinyl-tRNA^{Cys} (charged tRNA^{Cys} + cysteine). They then removed the sulfur group from the cysteine, creating alanyl-tRNA^{Cys} (charged tRNA^{Cys} + alanine). When alanyl-tRNA^{Cys} was added to a synthetic mRNA calling for cysteine, but not alanine, a polypeptide chain was synthesized containing alanine. What can you conclude from this experiment?

■ **HINT:** *This problem is concerned with establishing whether tRNA or the amino acid added to the tRNA during charging is responsible for attracting the charged tRNA to mRNA during translation. The key to its solution is the observation that in this experiment, when the triplet codon in mRNA calls for cysteine, alanine is inserted during translation, even though it is the "incorrect" amino acid.*

13.2 Translation of mRNA Can Be Divided into Three Steps

Like transcription, the process of translation can be best described by breaking it into discrete phases. We will consider three phases, each with its own illustration, but keep in mind that translation is a dynamic, continuous process. You should correlate the following discussion with the step-by-step characterization in the figures. Many of the protein factors and their roles in translation are summarized in Table 13.1.

TABLE 13.1 Various Protein Factors Involved during Translation in *E. coli*

Process	Factor	Role
Initiation of translation	IF1	Stabilizes 30*S* subunit
	IF2	Binds fmet-tRNA to 30*S*-mRNA complex; binds to GTP
	IF3	Binds 30*S* subunit to mRNA
Elongation of polypeptide	EF-Tu	Binds GTP; mediates aminoacyl-tRNA entry to the A site of ribosome
	EF-Ts	Generates active EF-Tu
	EF-G	Stimulates translocation; GTP-dependent
Termination of translation and release of polypeptide	RF1	Catalyzes release of the polypeptide chain from tRNA and dissociation of the translocation complex; specific for UAA and UAG termination codons
	RF2	Behaves like RF1; specific for UGA and UAA codons
	RF3	Stimulates RF1 and RF2

Initiation

Initiation of translation is depicted in **Figure 13–6**. Recall that the ribosome serves as a nonspecific workbench for the translation process. Most ribosomes, when they are not involved in translation, are dissociated into their large and small subunits. Initiation of translation in *E. coli* involves the small ribosome subunit, an mRNA molecule, a specific charged tRNA, GTP, Mg^{2+}, and at least three proteinaceous **initiation factors (IFs)** that enhance the binding affinity of the various translational components. In prokaryotes, the initiation codon of mRNA (AUG) calls for the modified amino acid **formylmethionine (fmet).**

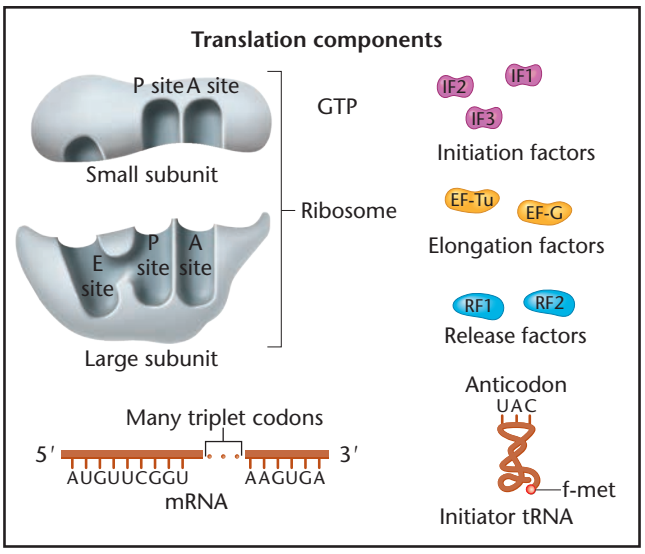
The small ribosomal subunit binds to IF1, and this complex then binds to mRNA (Step 1). In bacteria, this binding involves a sequence of up to six ribonucleotides (AGGAGG, not shown), which *precedes* the initial AUG start codon of mRNA. This sequence (containing only purines and called the **Shine–Dalgarno sequence**) base-pairs with a region of the 16*S* rRNA of the small ribosomal subunit, facilitating initiation.

While IF1 primarily blocks the A site from being bound to a tRNA and IF3 serves to inhibit the small subunit from associating with the large subunit, IF2 plays a more direct role in initiation. Essentially a GTPase, IF2 interacts with the mRNA and the charged tRNA, stabilizing them in the P site (Step 2). This step "sets" the reading frame so that all subsequent groups of three ribonucleotides are translated accurately. The aggregate, upon release of IF3, then combines with the large ribosomal subunit to create the 70*S* initiation complex. In this process, a molecule of GTP linked to IF2 is hydrolyzed, providing the required energy, and IF1 and 2 are subsequently released (Step 3).

Elongation

The second phase of translation, elongation, is depicted in **Figure 13–7**. As per our discussion above, the initiation complex is now poised for the insertion into the A site of the second aminoacyl tRNA bearing the amino acid corresponding to the second triplet sequence on the mRNA. Charged tRNAs are transported into the complex by one of the elongation factors, EF-Tu (Step 1). Like IF2 during initiation, EF-Tu is a GTPase and is bound by a GTP, the hydrolysis of which provides energy for the process.

The next step is for the terminal amino acid in the P site (methionine in this case) to be linked to the amino acid now present on the tRNA in the A site by the formation of a peptide bond. Such lengthening of the growing polypeptide chain by one amino acid is called *elongation*. Just prior to this, the covalent bond between the tRNA occupying the P site and its cognate amino acid is hydrolyzed (broken). The newly formed dipeptide remains attached to the end of the tRNA still residing in the A site



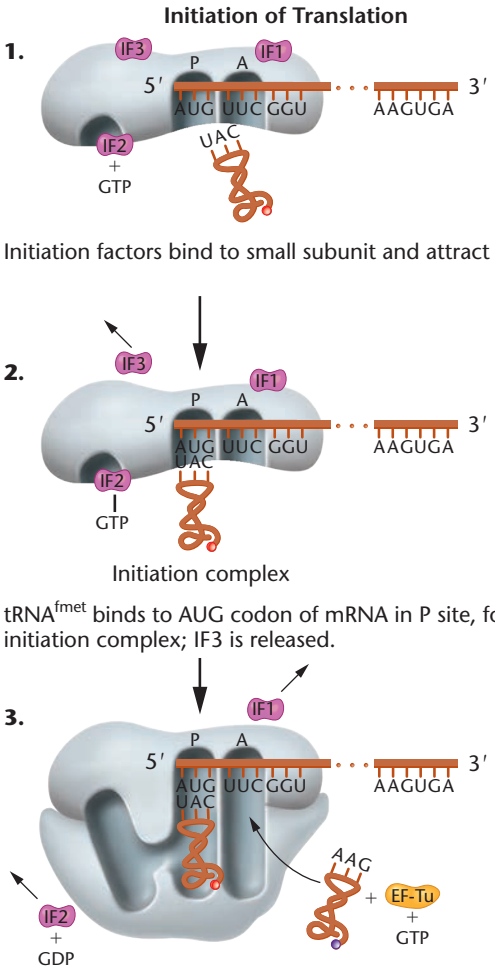


FIGURE 13–6 Initiation of translation. The components are depicted at the top of the figure.

(Step 2). These reactions were initially believed to be catalyzed by an enzyme called **peptidyl transferase,** embedded in the large subunit of the ribosome. However, it is now clear that this catalytic activity is a function of the

23*S* rRNA of the large subunit. In such a case, as we saw with splicing of pre-mRNAs (see Chapter 12), we refer to the complex as a **ribozyme,** recognizing the catalytic role that RNA plays in the process.

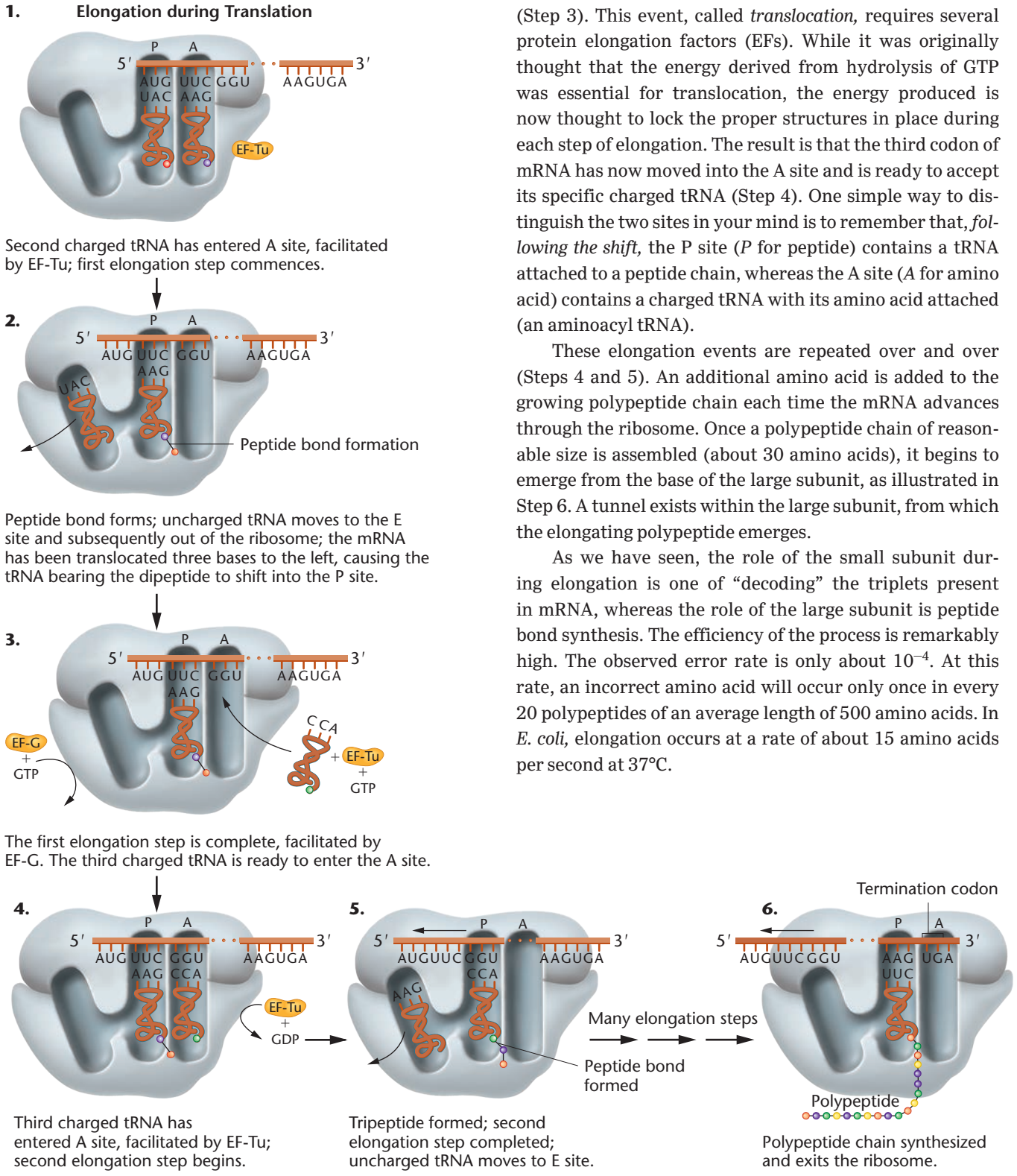


FIGURE 13–7 Elongation of the growing polypeptide chain during translation.

Before elongation can be repeated, the tRNA attached to the P site, which is now uncharged, must be released from the large subunit. The uncharged tRNA moves briefly into a third site on the ribosome called the **E (exit) site.** The entire **mRNA–tRNA–aa$_2$–aa$_1$** complex then shifts in the direction of the P site by a distance of three nucleotides (Step 3). This event, called *translocation,* requires several protein elongation factors (EFs). While it was originally thought that the energy derived from hydrolysis of GTP was essential for translocation, the energy produced is now thought to lock the proper structures in place during each step of elongation. The result is that the third codon of mRNA has now moved into the A site and is ready to accept its specific charged tRNA (Step 4). One simple way to distinguish the two sites in your mind is to remember that, *following the shift,* the P site (*P* for peptide) contains a tRNA attached to a peptide chain, whereas the A site (*A* for amino acid) contains a charged tRNA with its amino acid attached (an aminoacyl tRNA).

These elongation events are repeated over and over (Steps 4 and 5). An additional amino acid is added to the growing polypeptide chain each time the mRNA advances through the ribosome. Once a polypeptide chain of reasonable size is assembled (about 30 amino acids), it begins to emerge from the base of the large subunit, as illustrated in Step 6. A tunnel exists within the large subunit, from which the elongating polypeptide emerges.

As we have seen, the role of the small subunit during elongation is one of "decoding" the triplets present in mRNA, whereas the role of the large subunit is peptide bond synthesis. The efficiency of the process is remarkably high. The observed error rate is only about 10^{-4}. At this rate, an incorrect amino acid will occur only once in every 20 polypeptides of an average length of 500 amino acids. In *E. coli,* elongation occurs at a rate of about 15 amino acids per second at 37°C.

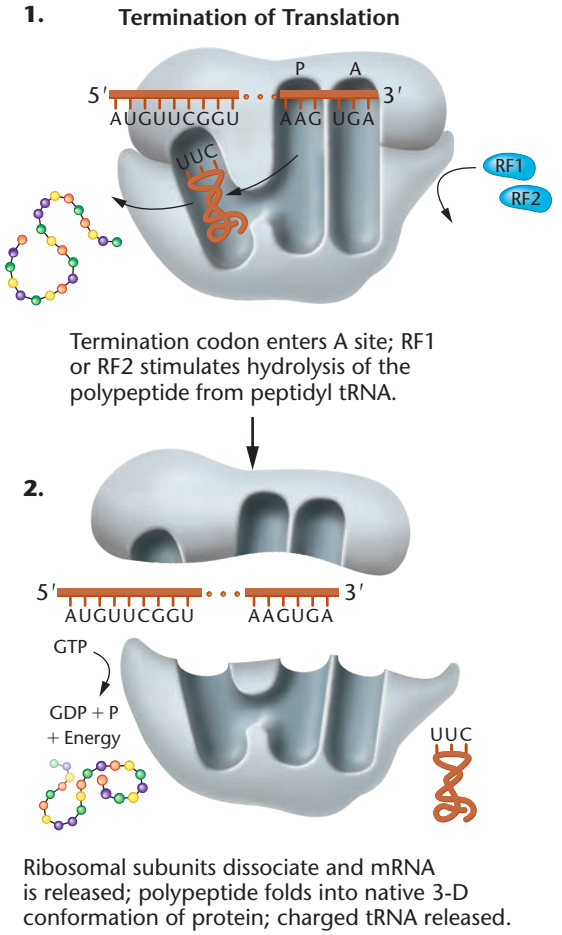


FIGURE 13–8 Termination of the process of translation.

Termination

Termination, the third phase of translation, is depicted in Figure 13–8. The process is signaled by one or more of three triplet codes in the A site: UAG, UAA, or UGA. These codons do not specify an amino acid, nor do they call for a tRNA in the A site. They are called **stop codons, termination codons,** or **nonsense codons.** Often, several such consecutive codons are part of an mRNA. The finished polypeptide is therefore still attached to the terminal tRNA at the P site, and the A site is empty. The termination codon signals the action of **GTP-dependent release factors,** which cleave the polypeptide chain from the terminal tRNA, releasing it from the translation complex (Step 1). Then, the tRNA is released from the ribosome, which then dissociates into its subunits (Step 2). If a termination codon should appear in the middle of an mRNA molecule as a result of mutation, the same process occurs, and the polypeptide chain is prematurely terminated.

Polyribosomes

As elongation proceeds and the initial portion of mRNA has passed through the ribosome, this mRNA is free to associate with another small subunit to form a second initiation complex. This process can be repeated several times with a single mRNA and results in what are called **polyribosomes,** or just **polysomes.**

Polyribosomes can be isolated and analyzed following a gentle lysis of cells. The photos in Figure 13–9 show these complexes as seen under an electron microscope. In Figure 13–9(a), you can see the thin lines of mRNA between the individual ribosomes. The micrograph in Figure 13–9(b) is even more remarkable, for it shows the polypeptide chains emerging from the ribosomes during translation. The formation of polysome complexes represents an efficient use of the components available for protein synthesis during a particular unit of time. Using the analogy of a song recorded on a tape and a tape recorder, in polysome complexes one

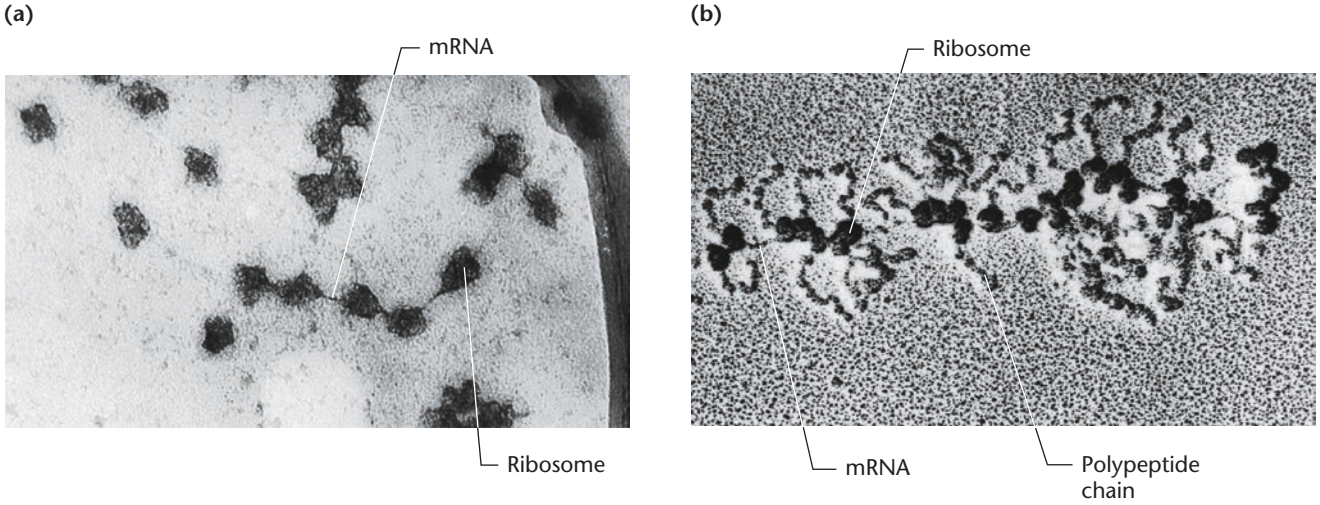


FIGURE 13–9 Polyribosomes as seen under the electron microscope. Those in (a) were derived from rabbit reticulocytes engaged in the translation of hemoglobin mRNA. The polyribosomes in (b) were taken from the giant salivary gland cells of the midgefly, *Chironomus thummi.* Note that the nascent polypeptide chains are apparent as they emerge from each ribosome. Their length increases as translation proceeds from left (5′) to right (3′) along the mRNA.

tape (mRNA) would be played simultaneously by several recorders (the ribosomes), but at any given moment, each recorder would be playing a different part of the song (the polypeptide being synthesized in each ribosome).

ESSENTIAL POINT

Translation, like transcription, is subdivided into the stages of initiation, elongation, and termination and relies on base-pairing affinities between complementary nucleotides. ■

13.3 High-Resolution Studies Have Revealed Many Details about the Functional Prokaryotic Ribosome

Our knowledge of the process of translation and the structure of the ribosome is based primarily on biochemical and genetic observations, in addition to the visualization of ribosomes under the electron microscope. To confirm and refine this information, the next step is to examine the ribosome at even higher levels of resolution. For example, X-ray diffraction analysis of ribosome crystals is one way to achieve this. However, because of its tremendous size and the complexity of molecular interactions occurring in the functional ribosome, it was extremely difficult to obtain the crystals necessary to perform X-ray diffraction studies. Nevertheless, great strides have been made over the past decade. First, the individual ribosomal subunits were crystallized and examined in several laboratories, most prominently that of Venkatraman Ramakrishnan. Then, the crystal structure of the intact 70*S* ribosome, complete with associated mRNA and tRNAs, was examined by Harry Noller and colleagues. In essence, the entire translational complex was seen at the atomic level. Both Ramakrishnan and Noller derived the ribosomes from the bacterium *Thermus thermophilus.*

Many noteworthy observations have come from these investigations. For example, the shape of the ribosome changes during different functional states, attesting to the dynamic nature of the process of translation. A great deal has also been learned about the location of the RNA components of the subunits. About one-third of the 16*S* RNA is responsible for producing a flat projection, referred to as the *platform,* within the smaller 30*S* subunit, and it modulates movement of the mRNA–tRNA complex during translocation. One of the models based on Noller's findings is shown in the opening photograph of this chapter (p. 254).

Crystallographic analysis also supports the concept that RNA is the real "player" in the ribosome during translation. The interface between the two subunits, considered to be the location in the ribosome where polymerization of amino acids occurs, is composed almost exclusively of RNA. In contrast, the numerous ribosomal proteins are found mostly on the periphery of the ribosome. These observations confirm what has been predicted on genetic grounds—the catalytic steps that join amino acids during translation occur under the direction of RNA, not proteins.

Another interesting finding involves the actual location of the various sites predicted to house tRNAs during translation. All three sites (A, P, and E) have been identified in X-ray diffraction studies, and in each case, the RNA of the ribosome makes direct contact with the various loops and domains of the tRNA molecule. This observation helps us understand why the distinctive three-dimensional conformation that is characteristic of all tRNA molecules has been preserved throughout evolution.

Still another noteworthy observation is that the intervals between the A, P, and E sites are at least 20 Å, and perhaps as much as 50 Å, wide, thus defining the atomic distance that the tRNA molecules must shift during each translocation event. This is considered a fairly large distance relative to the size of the tRNAs themselves. Further analysis has led to the identification of molecular (RNA–protein) bridges existing between the three sites and apparently involved in the translocation events. These observations provide us with a much more complete picture of the dynamic changes that must occur within the ribosome during translation. A final observation takes us back almost 50 years, to when Francis Crick proposed the *wobble hypothesis,* as introduced in Chapter 12. The Ramakrishnan group has identified the precise location along the 16*S* rRNA of the 30*S* subunit involved in the decoding step that connects mRNA to the proper tRNA. At this location, two particular nucleotides of the 16*S* rRNA actually flip out and probe the codon:anticodon region, and are believed to check for accuracy of base pairing during this interaction. According to the wobble hypothesis, the stringency of this step is high for the first two base pairs but less so for the third (or wobble) base pair.

As our knowledge of the translation process in prokaryotes has continued to grow, a remarkable study was reported in 2010 by Niels Fischer and colleagues. Using a unique high-resolution approach—the technique of *time-resolved single particle cryo-electron microscopy* (*cryo-EM*)—the 70*S E. coli* ribosome was captured and examined while in the process of translation at a resolution of 5.5 Å. In this work, over two million images were obtained and computationally analyzed, establishing a temporal snapshot of the trajectories of tRNA during the process of translocation. This research team examined how tRNA is translocated during elongation of the polypeptide chain. They demonstrated that the trajectories are coupled with dynamic conformational changes in the components of the ribosome. Surprisingly, the work has revealed that during translation, the ribosome behaves as a complex molecular machine *powered by Brownian movement driven by thermal*

energy. That is, the energetic requirements for achieving the various conformational changes essential to translocation are inherent to the ribosome itself.

Numerous questions about ribosome structure and function still remain. In particular, the precise role of the many ribosomal proteins is yet to be clarified. Nevertheless, the models that are emerging from the above research provide us with a much better understanding of the mechanism of translation.

13.4 Translation Is More Complex in Eukaryotes

The general features of the model we just discussed were initially derived from investigations of the translation process in bacteria. As we have seen (Figure 13–1), one main difference between prokaryotes and eukaryotes is that in eukaryotes, translation occurs on larger ribosomes whose rRNA and protein components are more complex. Interestingly, prokaryotic and eukaryotic rRNAs do share what is called a *core sequence,* but in eukaryotes, they are lengthened by the addition of *expansion sequences* (*ES*), which presumably impart added functionality. Another significant distinction is that whereas transcription and translation are coupled in prokaryotes, in eukaryotes these two processes are separated both spatially and temporally. In eukaryotic cells, transcription occurs in the nucleus and translation in the cytoplasm. This separation provides multiple opportunities for regulation of genetic expression in eukaryotic cells.

A number of aspects of the initiation of translation vary in eukaryotes. Three differences center on the mRNA that is being translated. First, the 5′-end of mRNA is capped with a **7-methylguanosine** (**7-mG**) residue at maturation (see Chapter 12). The presence of the cap, absent in prokaryotes, is essential for efficient initiation of translation. A second difference is that many mRNAs contain a purine (A or G) three bases upstream from the AUG initiator codon, which is followed by a G (A/GNNAUGG). Named after its discoverer, Marilyn Kozak, its presence in eukaryotes is considered to increase the efficiency of translation by interacting with the initator tRNA. This **Kozak sequence** is considered analogous to the *Shine–Dalgarno* sequence found in the upstream region of prokaryotic mRNAs. Above, N depicts any base.

Third, eukaryotic mRNAs require the posttranscriptional addition of a **poly-A tail** on their 3′-end; that is, they are *polyadenylated.* In the absence of poly A, these potential messages are rapidly degraded in the cytoplasm. Interestingly, histone mRNAs serve as an exception and are not polyadenylated. Still another difference related to initiation of translation is that in eukaryotes the amino acid formylmethionine is not required as it is in prokaryotes. However, the AUG triplet, which encodes methionine, is essential to the formation of the translational complex, and a unique transfer RNA ($tRNA_i^{Met}$) is used during initiation.

Still other differences are noteworthy. Eukaryotic mRNAs are much longer lived than are their prokaryotic counterparts. Most exist for hours rather than minutes prior to degradation by nucleases in the cell; thus they remain available much longer to orchestrate protein synthesis. And, during translation, protein factors similar to those in prokaryotes guide the initiation, elongation, and termination of translation in eukaryotes. Many of these eukaryotic factors are clearly homologous to their counterparts in prokaryotes. However, a greater number of factors are usually required during each step, and some are more complex than in prokaryotes. Finally, recall that in eukaryotes, many, but not all, of the cell's ribosomes are found in association with the membranes that make up the endoplasmic reticulum (forming the rough ER). Such membranes are absent from the cytoplasm of prokaryotic cells. This association in eukaryotes facilitates the secretion of newly synthesized proteins from the ribosomes directly into the channels of the endoplasmic reticulum. Recent studies using electron microscopy have established how this occurs. A *tunnel* in the large subunit of the ribosome begins near the point where the two subunits interface and exits near the back of the large subunit. The location of the tunnel within the large subunit is the basis for the belief that it provides the conduit for the movement of the newly synthesized polypeptide chain out of the ribosome. In studies in yeast, newly synthesized polypeptides enter the ER through a membrane channel formed by a specific protein, Sec61. This channel is perfectly aligned with the exit point of the ribosomal tunnel. In prokaryotes, the polypeptides are released by the ribosome directly into the cytoplasm.

13.5 The Initial Insight That Proteins Are Important in Heredity Was Provided by the Study of Inborn Errors of Metabolism

Let's consider how we know that proteins are the end products of genetic expression. The first insight into the role of proteins in genetic processes was provided by observations made by Sir Archibald Garrod and William Bateson early in the twentieth century. Garrod was born into an English family of medical scientists. His father was a physician with a strong interest in the chemical basis of rheumatoid arthritis, and his eldest brother was a leading zoologist in London. It is

not surprising, then, that as a practicing physician, Garrod became interested in several human disorders that seemed to be inherited. Although he also studied albinism and cystinuria, we shall describe his investigation of the disorder **alkaptonuria.** Individuals afflicted with this disorder have an important metabolic pathway blocked. As a result, they cannot metabolize the alkapton 2,5-dihydroxyphenylacetic acid, also known as homogentisic acid. Homogentisic acid accumulates in cells and tissues and is excreted in the urine. The molecule's oxidation products are black and easily detectable in the diapers of newborns. The products tend to accumulate in cartilaginous areas, causing the ears and nose to darken. The deposition of homogentisic acid in joints leads to a benign arthritic condition. This rare disease is not serious, but it persists throughout an individual's life.

Garrod studied alkaptonuria by looking for patterns of inheritance of this benign trait. Eventually he concluded that it was genetic in nature. Of 32 known cases, he ascertained that 19 were confined to seven families, with one family having four affected siblings. In several instances, the parents were unaffected but known to be related as first cousins, and therefore *consanguineous,* a term describing individuals descended from a common recent ancestor. Parents who are so related have a higher probability than unrelated parents of producing offspring that express recessive traits because such parents are both more likely to be heterozygous for some of the same recessive traits. Garrod concluded that this inherited condition was the result of an alternative mode of metabolism, thus implying that hereditary information controls chemical reactions in the body. While *genes* and *enzymes* were not familiar terms during Garrod's time, he used the corresponding concepts of *unit factors* and *ferments.* Garrod published his initial observations in 1902.

Only a few geneticists, including Bateson, were familiar with or referred to Garrod's work. Garrod's ideas fit nicely with Bateson's belief that inherited conditions are caused by the lack of some critical substance. In 1909, Bateson published *Mendel's Principles of Heredity,* in which he linked Garrod's ferments with heredity. However, for almost 30 years, most geneticists failed to see the relationship between genes and enzymes. Garrod and Bateson, like Mendel, were ahead of their time.

13.6 Studies of *Neurospora* Led to the One-Gene:One-Enzyme Hypothesis

In two separate investigations beginning in 1933, George Beadle provided the first convincing experimental evidence that genes are directly responsible for the synthesis of enzymes. The first investigation, conducted in collaboration with Boris Ephrussi, involved *Drosophila* eye pigments. Together, they confirmed that mutant genes that alter the eye color of fruit flies could be linked to biochemical errors that, in all likelihood, involved the loss of enzyme function. Encouraged by these findings, Beadle then joined with Edward Tatum to investigate nutritional mutations in the pink bread mold *Neurospora crassa.* This investigation led to the **one-gene:one-enzyme hypothesis.**

Analysis of *Neurospora* Mutants by Beadle and Tatum

In the early 1940s, Beadle and Tatum chose to work with *Neurospora* because much was known about its biochemistry and because mutations could be induced and isolated with relative ease. By inducing mutations, they produced strains that had genetic blocks of reactions essential to the growth of the organism.

Beadle and Tatum knew that this mold could manufacture nearly everything necessary for normal development. For example, using rudimentary carbon and nitrogen sources, this organism can synthesize nine water-soluble vitamins, 20 amino acids, numerous carotenoid pigments, and all essential purines and pyrimidines. Beadle and Tatum irradiated asexual conidia (spores) with X rays to increase the frequency of mutations and allowed them to be grown on "complete" medium containing all the necessary growth factors (e.g., vitamins and amino acids). Under such growth conditions, a mutant strain unable to grow on minimal medium was able to grow by virtue of supplements present in the enriched complete medium. All the cultures were then transferred to minimal medium. If growth occurred on the minimal medium, the organisms were able to synthesize all the necessary growth factors themselves, and the researchers concluded that the culture did not contain a nutritional mutation. If no growth occurred on minimal medium, they concluded that the culture contained a nutritional mutation, and the only task remaining was to determine its type. These results are shown in **Figure 13–10(a)**.

Many thousands of individual spores from this procedure were isolated and grown on complete medium. In subsequent tests on minimal medium, many cultures failed to grow, indicating that a nutritional mutation had been induced. To identify the mutant type, the mutant strains were then tested on a series of different minimal media [**Figure 13–10(b)**], each containing groups of supplements, and subsequently on media containing single vitamins, purines, pyrimidines, or amino acids [**Figure 13–10(c)**] until one specific supplement that permitted growth was found. Beadle and Tatum reasoned that *the supplement that restored growth would be the molecule that the mutant strain could not synthesize.*

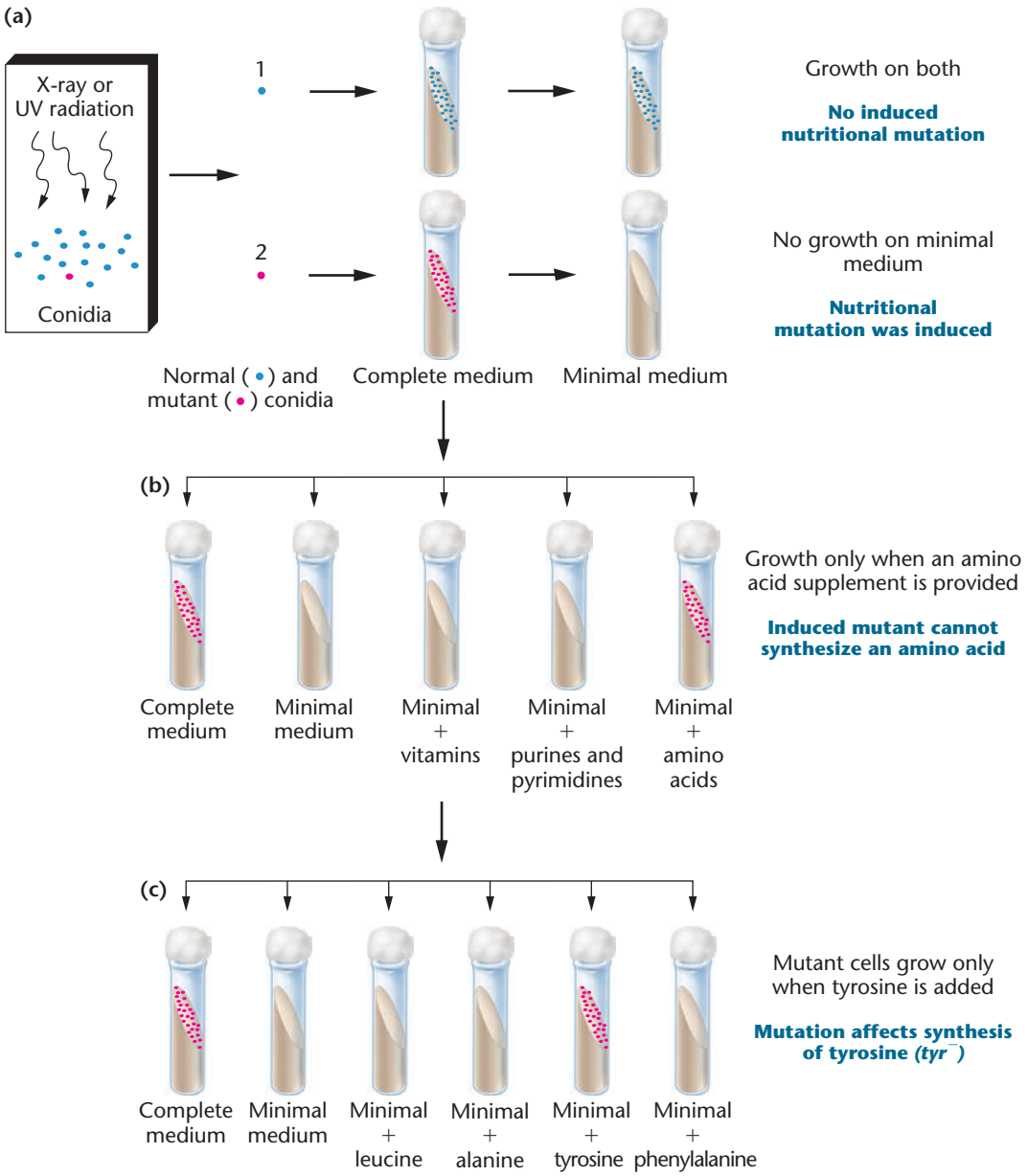


FIGURE 13–10 Induction, isolation, and characterization of a nutritional auxotrophic mutation in *Neurospora*. (a) Most conidia are not affected, but one conidium (shown in red) contains a mutation. In (b) and (c), the precise nature of the mutation is established and found to involve the biosynthesis of tyrosine.

The first mutant strain they isolated required vitamin B_6 (pyridoxine) in the medium, and the second required vitamin B_1 (thiamine). Using the same procedure, Beadle and Tatum eventually isolated and studied hundreds of mutants deficient in the ability to synthesize other vitamins, amino acids, or other substances.

The findings derived from testing over 80,000 spores convinced Beadle and Tatum that genetics and biochemistry have much in common. It seemed likely that each nutritional mutation caused the loss of the enzymatic activity that facilitated an essential reaction in wild-type organisms. It also appeared that a mutation could be found for nearly any enzymatically controlled reaction. Beadle and Tatum had thus provided sound experimental evidence for the hypothesis that *one gene specifies one enzyme,* an idea alluded to over 30 years earlier by Garrod and Bateson. With modifications, this concept was to become another major principle of genetics.

ESSENTIAL POINT

Beadle and Tatum's work with nutritional mutations in *Neurospora* led them to propose that one gene encodes one enzyme. ■

13.7 Studies of Human Hemoglobin Established That One Gene Encodes One Polypeptide

The one-gene:one-enzyme hypothesis that was developed in the early 1940s was not immediately accepted by all geneticists. This is not surprising because it was not yet clear how mutant enzymes could cause variation in many phenotypic traits. For example, *Drosophila* mutants demonstrate altered eye size, wing shape, wing-vein pattern, and so on. Plants exhibit mutant varieties of seed texture, height, and fruit size. How an inactive mutant enzyme could result in such phenotypes puzzled many geneticists.

Two factors soon modified the one-gene:one-enzyme hypothesis. First, although *nearly all enzymes are proteins, not all proteins are enzymes.* As the study of biochemical genetics progressed, it became clear that all proteins are specified by the information stored in genes, leading to the more accurate phraseology, **one-gene:one-protein hypothesis.** Second, proteins often show a substructure consisting of two or more polypeptide chains. This is the basis of the quaternary protein structure, which we will discuss later in this chapter.

Because each distinct polypeptide chain is encoded by a separate gene, a more accurate statement of Beadle and Tatum's basic tenet is **one-gene:one-polypeptide chain hypothesis.** These modifications of the original hypothesis became apparent during the analysis of hemoglobin structure in individuals afflicted with sickle-cell anemia.

Sickle-Cell Anemia

The first direct evidence that genes specify proteins other than enzymes came from work on mutant hemoglobin molecules found in humans afflicted with the disorder **sickle-cell anemia.** Affected individuals have erythrocytes that, under low oxygen tension, become elongated and curved because of the polymerization of hemoglobin. The sickle shape of these erythrocytes is in contrast to the biconcave disc shape characteristic in unaffected individuals (**Figure 13–11**). Those with the disease suffer attacks when red blood cells aggregate in the venous side of capillary systems, where oxygen tension is very low. As a result, a variety of tissues are deprived of oxygen and suffer severe damage. When this occurs, an individual is said to experience a *sickle-cell crisis.* If left untreated, a crisis can be fatal. The kidneys, muscles, joints, brain, gastrointestinal tract, and lungs can be affected.

In addition to suffering crises, these individuals are anemic because their erythrocytes are destroyed more rapidly than are normal red blood cells. Compensatory physiological mechanisms include increased red blood cell production by bone marrow, along with accentuated heart action. These mechanisms lead to abnormal bone size and shape, as well as dilation of the heart.

FIGURE 13–11 A comparison of an erythrocyte from a healthy individual (left) and from an individual afflicted with sickle-cell anemia (right).

In 1947, James Neel and E. A. Beet demonstrated that the disease is inherited as a Mendelian trait. Pedigree analysis revealed three genotypes and phenotypes controlled by a single pair of alleles, Hb^A and Hb^S. Unaffected and affected individuals result from the homozygous genotypes Hb^AHb^A and Hb^SHb^S, respectively. The red blood cells of heterozygotes, who exhibit the **sickle-cell trait** but not the disease, undergo much less sickling because over half of their hemoglobin is normal. Although they are largely unaffected, heterozygotes are "carriers" of the defective gene, which is transmitted on average to 50 percent of their offspring.

In the same year, Linus Pauling and his coworkers provided the first insight into the molecular basis of the disease. They showed that hemoglobins isolated from diseased and normal individuals differ in their rates of electrophoretic migration. In this technique, charged molecules migrate in an electric field. If the net charge of two molecules is different, their rates of migration will be different. On this basis, Pauling and his colleagues concluded that a chemical difference exists between normal **(HbA)** and sickle-cell **(HbS)** hemoglobin.

Pauling's findings suggested two possibilities. It was known that hemoglobin consists of four nonproteinaceous, iron-containing *heme groups* and a *globin portion* that contains four polypeptide chains. The alteration in net charge in HbS had to be due, theoretically, to a chemical change in one of these components. Pauling established that the globin portions were identical, and then around 1957, Vernon Ingram demonstrated that the chemical change occurs in the primary structure of the globin portion of the hemoglobin molecule. Ingram showed that HbS differs in amino acid composition compared to HbA. Human adult hemoglobin contains two identical

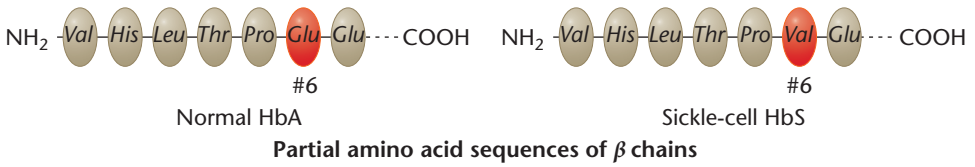


FIGURE 13–12 A comparison of the amino acid sequence of the β chain found in HbA and HbS.

α chains of 141 amino acids and two identical β chains of 146 amino acids in its quaternary structure. Analysis revealed just a single amino acid change: valine was substituted for glutamic acid at the sixth position of the β chain (**Figure 13–12**).

The significance of this discovery has been multifaceted. It clearly establishes that a single gene provides the genetic information for a single polypeptide chain. Studies of HbS also demonstrate that a mutation can affect the phenotype by directing a single amino acid substitution. Also, by providing the explanation for sickle-cell anemia, the concept of *inherited molecular disease* was firmly established. Finally, this work has led to a thorough study of human hemoglobins, which has provided valuable genetic insights.

In the United States, sickle-cell anemia is found almost exclusively in the African-American population. It affects about 1 in every 625 African-American infants. Currently, about 50,000 to 75,000 individuals are afflicted. In 1 of about every 145 African-American married couples, both partners are heterozygous carriers. In these cases, each of their children has a 25 percent chance of having the disease.

ESSENTIAL POINT

Pauling and Ingram's investigations of hemoglobin from patients with sickle-cell anemia led to the modification of the one-gene:one-enzyme hypothesis to indicate that one gene encodes one polypeptide chain. ■

EVOLVING CONCEPT OF THE GENE

In the 1940s, a time when the molecular nature of the gene had yet to be defined, groundbreaking work of Beadle and Tatum provided the first experimental evidence concerning the product of genes, their "one-gene:one-enzyme" hypothesis. This idea received further support and was later modified to indicate that one gene specifies one polypeptide chain. ■

13.8 Variation in Protein Structure Is the Basis of Biological Diversity

Having established that the genetic information is stored in DNA and influences cellular activities through the proteins it encodes, we turn now to a brief discussion of protein structure. How can these molecules play such a critical role in determining the complexity of cellular activities? As we shall see, the fundamental aspects of the structure of proteins provide the basis for incredible complexity and diversity. At the outset, we should differentiate between **polypeptides** and **proteins.** Both are molecules composed of amino acids. They differ, however, in their state of assembly and functional capacity. Polypeptides are the precursors of proteins. As it is assembled on the ribosome during translation, the molecule is called a *polypeptide.* When released from the ribosome following translation, a polypeptide folds up and assumes a higher order of structure. When this occurs, a three-dimensional conformation emerges. In many cases, several polypeptides interact to produce this conformation. When the final conformation is achieved, the molecule is now fully functional and is appropriately called a *protein.* Its three-dimensional conformation is essential to the function of the molecule.

The polypeptide chains of proteins, like nucleic acids, are linear nonbranched polymers. There are 20 commonly occurring amino acids that serve as the subunits (the building blocks) of proteins. Each amino acid has a **carboxyl group,** an **amino group,** and an **R (radical) group** (a side chain) bound covalently to a **central carbon (C) atom.** The R group gives each amino acid its chemical identity exhibiting a variety of configurations that can be divided into four main classes: *nonpolar* (hydrophobic), *polar* (hydrophilic), *positively charged,* and *negatively charged.* **Figure 13–13** shows the chemical structure of an amino acid and one example from each of these categories. Because polypeptides are often long polymers and because each position may be occupied by any 1 of the 20 amino acids with their unique chemical properties, enormous variation in chemical conformation and activity is possible. For example, if an average polypeptide is composed of 200 amino acids (molecular weight of about 20,000 Da), 20^{200} different molecules, each with a unique sequence, can be created using the 20 different building blocks.

Around 1900, German chemist Emil Fischer determined the manner in which the amino acids are bonded together. He showed that the amino group of one amino acid reacts with the carboxyl group of another amino acid during a dehydration reaction, releasing a molecule of H_2O. The resulting covalent bond is a **peptide bond.** Two

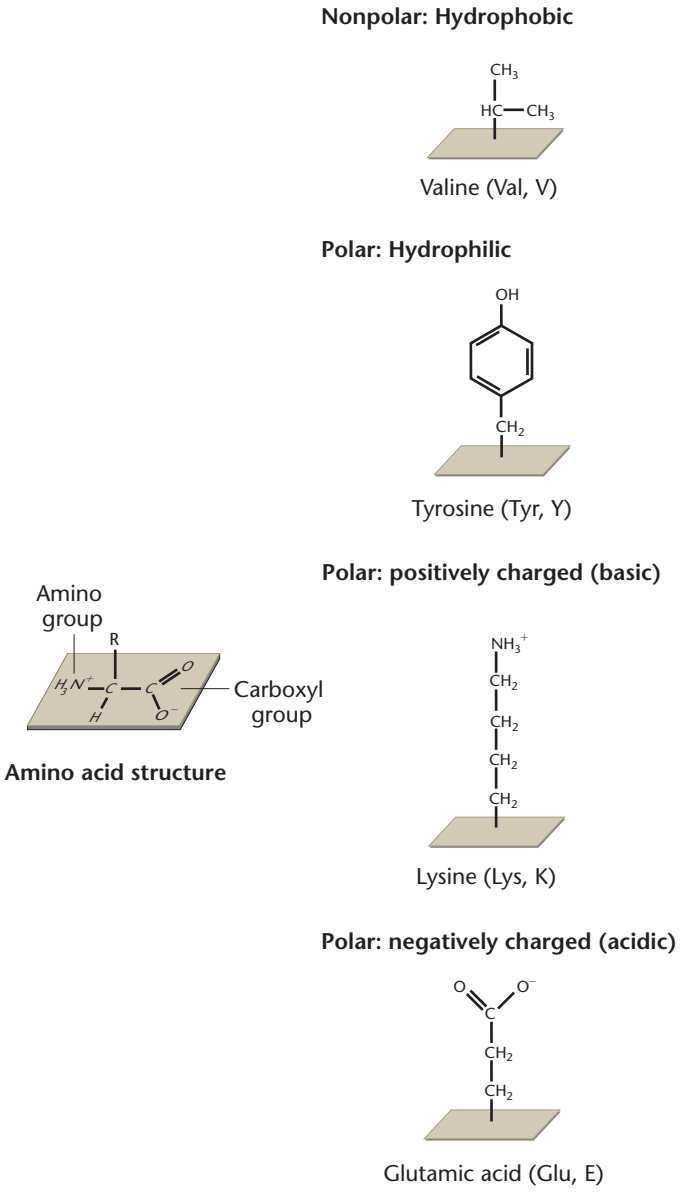


FIGURE 13–13 Chemical structure of an amino acid as well as an example of the R group characterizing each of the four categories of amino acids. Each amino acid has two abbreviations, often based on the first three letters of its name; for example, valine is designated either Val or V.

amino acids linked together constitute a **dipeptide,** three a **tripeptide,** and so on. Once 10 or more amino acids are linked by peptide bonds, the chain is referred to as a polypeptide. Generally, no matter how long a polypeptide is, it will contain a free amino group at one end (the N-terminus) and a free carboxyl group at the other end (the C-terminus).

Four levels of protein structure are recognized: primary, secondary, tertiary, and quaternary. The sequence of amino acids in the linear backbone of the polypeptide constitutes its **primary structure.** It is specified by the sequence of deoxyribonucleotides in DNA via an mRNA intermediate. The primary structure of a polypeptide helps determine the specific characteristics of the higher orders of organization as a protein is formed.

Secondary structures are certain regular or repeating configurations in space assumed by amino acids lying close to one another in the polypeptide chain. In 1951, Linus Pauling, Herman Branson, and Robert Corey predicted, on theoretical grounds, an ***α*-helix** as one type of secondary structure. The α-helix model [**Figure 13–14(a)**] has since been confirmed by X-ray crystallographic studies. The helix is composed of a spiral chain of amino acids stabilized by hydrogen bonds.

The side chains (the R groups) of amino acids extend outward from the helix, and each amino acid residue occupies a vertical distance of 1.5 Å in the helix. There are 3.6 residues per turn. While left-handed helices are theoretically possible, all proteins seen with an a helix are right-handed.

Also in 1951, Pauling and Corey proposed a second structure, the ***β*-pleated sheet.** In this model, a single-polypeptide chain folds back on itself, or several chains run in either parallel or antiparallel fashion next to one another. Each such structure is stabilized by hydrogen bonds formed between atoms on adjacent chains [**Figure 13–14(b)**]. A zigzagging plane is formed in space with adjacent amino acids 3.5 Å apart. As a general rule, most proteins demonstrate a mixture of α-helix and β-pleated-sheet structures.

While the secondary structure describes the arrangement of amino acids within certain areas of a polypeptide chain, the **tertiary structure** defines the three-dimensional conformation of the entire chain in space. Each protein twists and turns and loops around itself in a very particular fashion, characteristic of the specific protein. A model of the three-dimensional tertiary structure of the respiratory pigment myoglobin is shown in **Figure 13–15**.

The three-dimensional conformation achieved by any protein is a product of the *primary structure* of the polypeptide. As the polypeptide is folded, the most thermodynamically stable conformation is created. This level of organization is essential because the specific function of any protein is directly related to its tertiary structure.

The concept of **quaternary structure** applies to those proteins composed of more than one polypeptide chain and indicates the position of the various chains in relation to one another. Hemoglobin, a protein consisting of four polypeptide chains, has been studied in great detail. Most enzymes, including DNA and RNA polymerase, demonstrate quaternary structure.

ESSENTIAL POINT

Proteins, the end products of genes, demonstrate four levels of structural organization that together describe their three-dimensional conformation, which is the basis of each molecule's function. ■

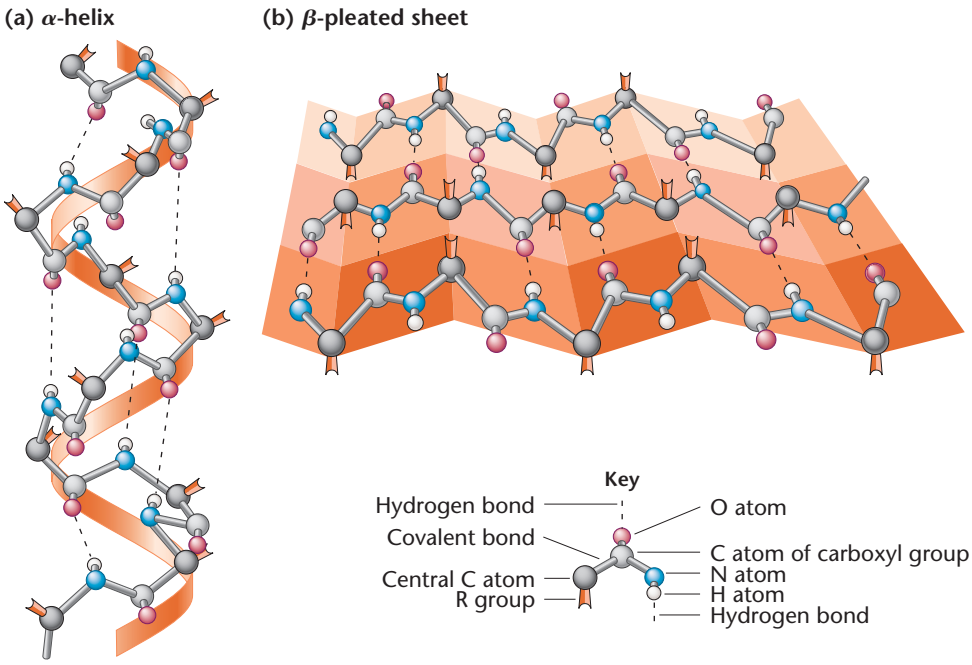


FIGURE 13–14 (a) The right-handed α-helix, which represents one form of secondary structure of a polypeptide chain. (b) The β-pleated sheet, an alternative form of secondary structure of polypeptide chains. To maintain clarity, not all atoms are shown.

NOW SOLVE THIS

13–2 HbS results from the substitution of valine for glutamic acid at the number 6 position in the β chain of human hemoglobin. HbC is the result of a change at the same position in the β chain, but in this case lysine replaces glutamic acid. Return to the genetic code table (Figure 12–7) and determine whether single-nucleotide changes can account for these mutations. Then view Figure 13–13 and examine the R groups in the amino acids glutamic acid, valine, and lysine. Describe the chemical differences between the three amino acids. Predict how the changes might alter the structure of the molecule and lead to altered hemoglobin function.

■ **HINT:** *This problem asks you to consider the potential impact of several amino acid substitutions that result from mutations in one of the genes encoding one of the chains making up human hemoglobin. The key to its solution is to consider and compare the structure of the three amino acids (glutamic acid, lysine, and valine) and their net charge (see Figure 13–13).*

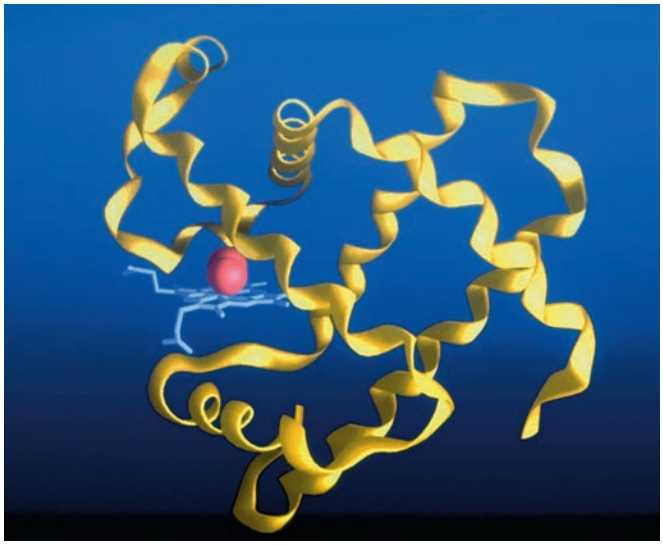


FIGURE 13–15 The tertiary level of protein structure in a respiratory pigment, myoglobin. The bound oxygen atom is shown in red.

Protein Folding and Misfolding

It was long thought that **protein folding** was a spontaneous process whereby a linear molecule exiting the ribosome achieved a three-dimensional, thermodynamically stable conformation based solely on the combined chemical properties inherent in the amino acid sequence. This indeed is the case for many proteins. However, numerous studies have shown that for other proteins, correct folding is dependent on members of a family of molecules called **chaperones.** Chaperones are themselves proteins (sometimes called *molecular chaperones* or *chaperonins*) that function by mediating the folding process by excluding the formation of alternative, incorrect patterns. While they may initially interact with the protein in question, like enzymes, they do not become part of the final product. Initially discovered in *Drosophila,* in which they are called **heat-shock proteins,** chaperones are ubiquitous, having now been discovered in all organisms. They are even present in mitochondria and chloroplasts.

In eukaryotic cells, chaperones are particularly important when translation occurs on membrane-bound ribosomes, where the newly translated polypeptide is extruded into the lumen of the endoplasmic reticulum. Even in their presence, misfolding may still occur, and one more system of "quality

control" exists. As misfolded proteins are transported out of the endoplasmic reticulum to the cytoplasm, they are "tagged" by another class of small proteins called **ubiquitins.** The protein–ubiquitin complex moves to a cellular structure called the **proteasome,** within which the ubiquitin is released and the misfolded proteins are degraded by proteases.

Protein folding is a critically important process, not only because misfolded proteins may be nonfunctional, but also because improperly folded proteins can accumulate and be detrimental to cells and the organisms that contain them. For example, a group of transmissible brain disorders in mammals—**scrapie** in sheep, **bovine spongioform encephalopathy** (**mad cow disease**) in cattle, and **Creutzfeldt–Jakob disease** in humans—are caused by the presence in the brain of **prions,** which are aggregates of a misfolded protein. The misfolded protein (called PrP^{Sc}) is an altered version of a normal cellular protein (called PrP^{C}) synthesized in neurons and found in the brains of all adult animals. The difference between PrP^{C} and PrP^{Sc} lies in their secondary protein structures. Normal, noninfectious PrP^{c} folds into an α-helix, whereas infectious PrP^{Sc} folds into a β-pleated sheet. When an abnormal PrP^{Sc} molecule contacts a PrP^{C} molecule, the normal protein refolds into the abnormal conformation. The process continues as a chain reaction, with potentially devastating results—the formation of prion particles that eventually destroy the brain. Hence, this group of disorders can be considered diseases of secondary protein structure.

Currently, many laboratories are studying protein folding and misfolding, particularly as related to genetics. Numerous inherited human disorders are caused by misfolded proteins that form abnormal aggregates. Sickle-cell anemia, discussed earlier in this chapter, is a case in point, where the β chains of hemoglobin are altered as the result of a single amino acid change, causing the molecules to aggregate within erythrocytes, with devastating results. An autosomal dominant inherited form of Creutzfeldt–Jakob disease is known in which the mutation alters the PrP amino acid sequence, leading to prion formation. Various progressive neurodegenerative diseases such as **Huntington disease, Alzheimer disease,** and **Parkinson disease** are linked to the formation of abnormal protein aggregates in the brain. Huntington disease is inherited as an autosomal dominant trait, whereas less clearly defined genetic components are associated with Alzheimer and Parkinson diseases.

13.9 Proteins Function in Many Diverse Roles

The essence of life on Earth rests at the level of diverse cellular function. One can argue that DNA and RNA simply serve as vehicles to store and express genetic information. However, proteins are at the heart of cellular function. And it is the capability of cells to assume diverse structures and functions that distinguishes most eukaryotes from less evolutionarily advanced organisms such as bacteria. Therefore, an introductory understanding of protein function is critical to a complete view of genetic processes.

Proteins are the most abundant macromolecules found in cells. As the end products of genes, they play many diverse roles. For example, the respiratory pigments **hemoglobin** and **myoglobin,** discussed earlier in the chapter, transport oxygen, which is essential for cellular metabolism. **Collagen** and **keratin** are structural proteins associated with the skin, connective tissue, and hair of organisms. **Actin** and **myosin** are contractile proteins, found in abundance in muscle tissue, while **tubulin** is the basis of the function of microtubules in mitotic and meiotic spindles. Still other examples are the **immunoglobulins,** which function in the immune system of vertebrates; **transport proteins,** involved in the movement of molecules across membranes; some of the hormones and their receptors, which regulate various types of chemical activity; **histones,** which bind to DNA in eukaryotic organisms; and **transcription factors** that regulate gene expression.

Nevertheless, the most diverse and extensive group of proteins (in terms of function) are the enzymes, to which we have referred throughout this chapter. Enzymes specialize in catalyzing chemical reactions within living cells. Like all catalysts, they increase the rate at which a chemical reaction reaches equilibrium, but they do not alter the end-point of the chemical equilibrium. Their remarkable, highly specific catalytic properties largely determine the metabolic capacity of any cell type and provide the underlying basis of what we refer to as biochemistry. The specific functions of many enzymes involved in the genetic and cellular processes of cells are described throughout the text.

ESSENTIAL POINT

Of the myriad functions performed by proteins, the most influential role belongs to enzymes, which serve as highly specific biological catalysts that play a central role in the production of all classes of molecules in living systems. ■

Protein Domains Impart Function

We conclude this chapter by briefly discussing the important finding that regions made up of specific amino acid sequences are associated with specific functions in protein molecules. Such sequences, usually between 50 and 300 amino acids, constitute **protein domains** and represent modular portions of the protein that fold into stable, unique conformations independently of the rest of the molecule. Different domains impart different functional capabilities. Some proteins contain only a single domain, while others contain two or more.

The significance of domains resides in the tertiary structures of proteins. Each domain can contain a mixture of secondary structures, including α-helices and β-pleated sheets. The unique conformation of a given domain imparts a specific function to the protein. For example, a domain may serve as the catalytic site of an enzyme, or it may impart an ability to bind to a specific ligand. Thus, discussions of proteins may mention *catalytic domains, DNA-binding domains,* and so on. In short, a protein must be seen as being composed of a series of structural and functional modules. Obviously, the presence of multiple domains in a single protein increases the versatility of each molecule and adds to its functional complexity.

CASE STUDY | Crippled ribosomes

Diamond Blackfan anemia (DBA) is a rare, dominantly inherited syndrome characterized by bone marrow failure, birth defects, and a significant predisposition to cancer. Those affected with DBA usually develop anemia in the first year of life, have abnormal numbers of cell types in their bone marrow, and have an increased risk of developing leukemia and bone cancer. At the molecular level, DBA is caused by a mutation in any of 11 genes that encode ribosomal proteins. The common feature of all these mutations is the disruption of ribosome formation, ultimately affecting the stability or function of ribosomes.

Many questions about this disorder remain to be answered.

1. Given the central importance of ribosomes in maintaining life, how is it possible that individuals carrying mutations in ribosomal protein genes survive?
2. Why might some cells in the body, such as those in bone marrow, be more susceptible to ribosomal protein mutations than other cell types?
3. DBA exhibits variable penetrance with significant differences in the clinical symptoms. How does this provide a way of studying the molecular events that cause this disorder?

INSIGHTS AND SOLUTIONS

1. As an extension of Beadle and Tatum's work with *Neurospora,* it is possible to study multiple mutations whose impact is on the same biochemical pathway. The growth responses in the following chart were obtained using four mutant strains of *Neurospora* and the chemically related compounds A, B, C, and D. None of the mutants grow on minimal medium. Draw all possible conclusions from these data.

	Growth Supplement			
Mutation	A	B	C	D
1	–	–	–	–
2	+	+	–	+
3	+	+	–	–
4	–	+	–	–

Solution: Nothing can be concluded about mutation *1* except that it lacks some essential growth factor, perhaps even unrelated to the biochemical pathway represented by mutations *2, 3,* and *4*. Nor can anything be concluded about compound C. If it is involved in the pathway, it is a product that was synthesized prior to compounds A, B, and D.

We now analyze these three compounds and the control of their synthesis by the enzymes encoded by mutations *2, 3,* and *4*. Because product B allows growth in all three cases, it may be considered the "end product"—it bypasses the block in all three instances. Using similar reasoning, product A precedes B in the pathway because it bypasses the block in two of the three steps, and product D precedes B yielding a partial solution

$$\text{C(?)} \longrightarrow \text{D} \longrightarrow \text{A} \longrightarrow \text{B}$$

Now let's determine which mutations control which steps. Since mutation *2* can be alleviated by products D, B, and A, it must control a step prior to all three products, perhaps the direct conversion to D (although we cannot be certain). Mutation *3* is alleviated by B and A, so its effect must precede them in the pathway. Thus, we assign it as controlling the conversion of D to A. Likewise, we can assign mutation *4* to the conversion of A to B, leading to a more complete solution

$$\text{C(?)} \xrightarrow{2(?)} \text{D} \xrightarrow{3} \text{A} \xrightarrow{4} \text{B}$$

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we focused on the translation of mRNA into proteins as well as on protein structure and function. Along the way, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions:

(a) What experimentally derived information led to Holley's proposal of the two-dimensional cloverleaf model of tRNA?
(b) What experimental information verifies that certain codons in mRNA specify chain termination during translation?
(c) How do we know, based on studies of *Neurospora* nutritional mutations, that one gene specifies one enzyme?
(d) On what basis have we concluded that proteins are the end products of genetic expression?

CHAPTER CONCEPTS

2. Review the Chapter Concepts list on p. 254. These all relate to the translation of genetic information stored in mRNA into proteins and how chemical information in proteins impart function to those molecules. Write a brief essay that discusses the role of ribosomes in the process of translation as it relates to these concepts. ■

3. List and describe the role of all molecular constituents present in a functional polyribosome.
4. Contrast the roles of tRNA and mRNA during translation, and list all enzymes that participate in the translation processes.
5. tRNA adapts specific triplet codons in mRNA to their correct amino acids. Do you agree with this statement? Justify your answer.
6. Knowing that the base sequence of any given messenger RNA is responsible for precisely ordering the amino acids in a protein, present two mechanisms by which intrinsic properties of mRNA may regulate the "net output" of a given gene.
7. Summarize the steps involved in charging tRNAs with their appropriate amino acids.
8. Based on the cloverleaf model and the three-dimensional structure of tRNA, mention the different regions present in a tRNA molecule.
9. Explain why the one-gene:one-enzyme hypothesis is no longer considered to be totally accurate.
10. Hemoglobin is a tetramer consisting of two α and two β chains. What level of protein structure is described in the above statement?
11. Using sickle-cell anemia as a basis, describe what is meant by a genetic or inherited molecular disease. What are the similarities and dissimilarities between this type of a disorder and a disease caused by an invading microorganism?
12. Explain the characteristic of sickle-cell hemoglobin that makes it different from normal hemoglobin.
13. Assume that an mRNA molecule that has 12 triplet codons, excluding the start and stop codons, occupies space in a ribosome that is 21 nm in diameter. If the entire primary protein sequence can be accommodated in that ribosome, predict the length of each nucleotide.
14. Review the concept of colinearity in Section 12.5 (p. 239) and consider the following question: Certain mutations called *amber* in bacteria and viruses result in premature termination of polypeptide chains during translation. Many *amber* mutations have been detected at different points along the gene that codes for a head protein in phage T4. How might this system be further investigated to demonstrate and support the concept of colinearity?
15. Explain the importance of primary and tertiary structures in the functioning of a protein.
16. List and describe the function of as many nonenzymatic proteins as you can that are unique to eukaryotes.
17. How does an enzyme function? Why are enzymes essential for living organisms?
18. Shown in the following table are several amino acid substitutions in the α and β chains of human hemoglobin. Use the genetic code table in Figure 12–7 to determine how many of them can occur as a result of a single nucleotide change.

Hb Type	Normal Amino Acid	Substituted Amino Acid
HbJ Toronto	Ala	Asp (α-5)
HbJ Oxford	Gly	Asp (α-15)
Hb Mexico	Gln	Glu (α-54)
Hb Bethesda	Tyr	His (β-145)
Hb Sydney	Val	Ala (β-67)
HbM Saskatoon	His	Tyr (β-63)

19. Three independently assorting genes are known to control the biochemical pathway below that provides the basis for flower color in a hypothetical plant

$$\text{colorless} \xrightarrow{A-} \text{yellow} \xrightarrow{B-} \text{green} \xrightarrow{C-} \text{speckled}$$

Homozygous recessive mutations, which disrupt enzyme function controlling each step, are known. Determine the phenotypic results in the F_1 and F_2 generations resulting from the P_1 crosses involving true-breeding plants given here.

(a) speckled	(*AABBCC*)	×	yellow	(*AAbbCC*)
(b) yellow	(*AAbbCC*)	×	green	(*AABBcc*)
(c) colorless	(*aaBBCC*)	×	green	(*AABBcc*)

20. How would the results in cross (a) of Problem 19 vary if genes *A* and *B* were linked with no crossing over between them? How would the results of cross (a) vary if genes *A* and *B* were linked and 20 map units apart?
21. A series of mutations in the bacterium *Salmonella typhimurium* results in the requirement of either tryptophan or some related molecule in order for growth to occur. From the data shown here, suggest a biosynthetic pathway for tryptophan.

	Growth Supplement				
Mutation	Minimal Medium	Anthranilic Acid	Indole Glycerol Phosphate	Indole	Tryptophan
trp-8	–	+	+	+	+
trp-2	–	–	+	+	+
trp-3	–	–	–	+	+
trp-1	–	–	–	–	+

22. The emergence of antibiotic-resistant strains of *Enterococci* and transfer of resistant genes to other bacterial pathogens have highlighted the need for new generations of antibiotics to combat serious infections. To grasp the range of potential sites for the action of existing antibiotics, sketch the components of the translation machinery (e.g., see Step 3 of Figure 13–6), and using a series of numbered pointers, indicate the specific location for the action of the antibiotics shown in the following table.

Antibiotic	Action
1. Streptomycin	Binds to 30*S* ribosomal subunit
2. Chloramphenicol	Inhibits the peptidyl transferase function of 70*S* ribosome
3. Tetracycline	Inhibits binding of charged tRNA to ribosome
4. Erythromycin	Binds to free 50*S* particle and prevents formation of 70*S* ribosome
5. Kasugamycin	Inhibits binding of $tRNA^{fmet}$
6. Thiostrepton	Prevents translocation by inhibiting EF-G

14 Gene Mutation, DNA Repair, and Transposition

Pigment mutations within an ear of corn, caused by transposition of the *Ds* element.

CHAPTER CONCEPTS

- Mutations comprise any change in the base-pair sequence of DNA.
- Mutations are a source of genetic variation and provide the raw material for natural selection. They are also the source of genetic damage that contributes to cell death, genetic diseases, and cancer.
- Mutations have a wide range of effects on organisms depending on the type of base-pair alteration, the location of the mutation within the chromosome, and the function of the affected gene product.
- Mutations can occur spontaneously as a result of natural biological and chemical processes, or they can be induced by external factors, such as chemicals or radiation.
- Single-gene mutations cause a wide variety of human diseases.
- Organisms rely on a number of DNA repair mechanisms to detect and correct mutations. These mechanisms range from proofreading and correction of replication errors to base excision and homologous recombination repair.
- Mutations in genes whose products control DNA repair lead to genome hypermutability, human DNA repair diseases, and cancers.
- Transposable elements may move into and out of chromosomes, causing chromosome breaks and inducing mutations both within coding regions and in gene-regulatory regions.

The ability of DNA molecules to store, replicate, transmit, and decode information is the basis of genetic function. But equally important are the changes that occur to DNA sequences. Without the variation that arises from changes in DNA sequences, there would be no phenotypic variability, no adaptation to environmental changes, and no evolution. Gene mutations are the source of new alleles and are the origin of genetic variation within populations. On the downside, they are also the source of genetic changes that can lead to cell death, genetic diseases, and cancer.

Mutations also provide the basis for genetic analysis. The phenotypic variations resulting from mutations allow geneticists to identify and study the genes responsible for the modified trait. In genetic investigations, mutations act as identifying "markers" for genes so that they can be followed during their transmission from parents to offspring. Without phenotypic variability, classical genetic analysis would be impossible. For example, if all pea plants displayed a uniform phenotype, Mendel would have had no foundation for his research.

As discussed earlier in the text (see Chapter 6), we examined mutations in large regions of chromosomes—chromosomal mutations. In contrast, the mutations we will now explore are those occurring primarily in the base-pair sequence of DNA within individual genes—**gene mutations**. We will also describe how the cell defends itself from such mutations using various mechanisms of DNA repair.

14.1 Gene Mutations Are Classified in Various Ways

A mutation can be defined as an alteration in DNA sequence. Any base-pair change in any part of a DNA molecule can be considered a mutation. A mutation may comprise a single base-pair substitution, a deletion or insertion of one or more base pairs, or a major alteration in the structure of a chromosome.

Mutations may occur within regions of a gene that code for protein or within noncoding regions of a gene such as introns and regulatory sequences. Mutations may or may not bring about a detectable change in phenotype. The extent to which a mutation changes the characteristics of an organism depends on which type of cell suffers the mutation and the degree to which the mutation alters the function of a gene product or a gene-regulatory region.

Mutations can occur in somatic cells or within germ cells. Those that occur in germ cells are heritable and are the basis for the transmission of genetic diversity and evolution, as well as genetic diseases. Those that occur in somatic cells are not transmitted to the next generation but may lead to altered cellular function or tumors.

Because of the wide range of types and effects of mutations, geneticists classify mutations according to several different schemes. These organizational schemes are not mutually exclusive. In this section, we outline some of the ways in which gene mutations are classified.

Classification Based on Type of Molecular Change

Geneticists often classify gene mutations in terms of the nucleotide changes that constitute the mutation. A change of one base pair to another in a DNA molecule is known as a **point mutation,** or **base substitution** (Figure 14–1). A change of one nucleotide of a triplet within a protein-coding portion of a gene may result in the creation of a new triplet that codes for a different amino acid in the protein product. If this occurs, the mutation is known as a **missense mutation.** A second possible outcome is that the triplet will be changed into a stop codon, resulting in the termination of translation of the protein. This is known as a **nonsense mutation.** If the point mutation alters a codon but does not result in a change in the amino acid at that position in the protein (due to degeneracy of the genetic code), it can be considered a **silent mutation.**

You will often see two other terms used to describe base substitutions. If a pyrimidine replaces a pyrimidine or a purine replaces a purine, a **transition** has occurred. If a purine replaces a pyrimidine, or vice versa, a **transversion** has occurred.

Another type of change is the insertion or deletion of one or more nucleotides at any point within the gene. As illustrated in Figure 14–1, the loss or addition of a single nucleotide causes all of the subsequent three-letter codons to be changed. These are called **frameshift mutations** because the frame of triplet reading during translation is altered. A frameshift mutation will occur when any number of bases are added or deleted, except multiples of three, which would reestablish the initial frame of reading. It is possible that one of the many altered triplets will be UAA, UAG, or UGA, the translation termination codons. When one of these triplets is encountered during translation, polypeptide synthesis is terminated at that point. Obviously, the results of frameshift mutations can be very severe, especially if they occur early in the coding sequence.

Classification Based on Phenotypic Effects

Depending on their type and location, mutations can have a wide range of phenotypic effects, from none to severe. As discussed earlier in the text (see Chapter 4), a **loss-of-function mutation** is one that reduces or eliminates the function of the gene product. Any type of mutation, from a point mutation to deletion of the entire gene, may lead to a loss of function. Mutations that result in complete loss of function are known as **null mutations.**

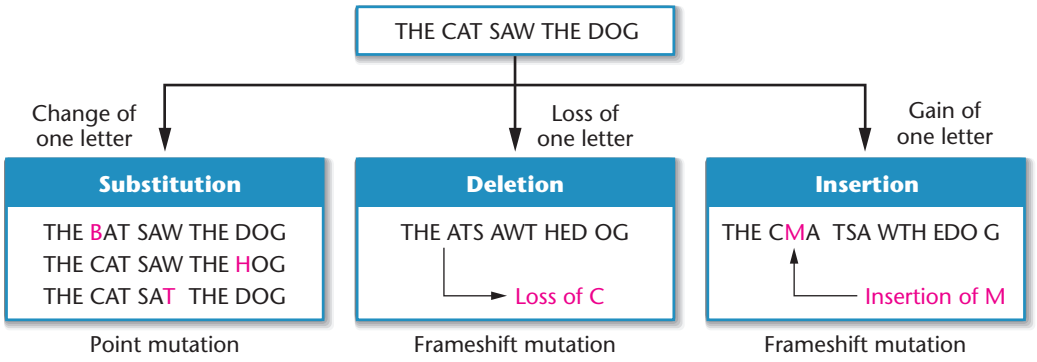


FIGURE 14–1 Analogy showing the effects of substitution, deletion, and insertion of one letter in a sentence composed of three-letter words to demonstrate point and frameshift mutations.

Most loss-of-function mutations are recessive. A **recessive mutation** results in a wild-type phenotype when present in a diploid organism and the other allele is wild type. In this case, the presence of less than 100 percent of the gene product is sufficient to bring about the wild-type phenotype.

Some loss-of-function mutations can be dominant. A **dominant mutation** results in a mutant phenotype in a diploid organism, even when the wild-type allele is also present. Dominant mutations can have two different types of effects. **Haploinsufficiency** occurs when the single wild-type copy of the gene does not produce enough gene product to bring about a wild-type phenotype. In humans, Marfan syndrome is an example of a disorder caused by haploinsufficiency—in this case as a result of a loss-of-function mutation in one copy of the *FBN1* gene. In contrast, a **dominant gain-of-function mutation** results in a gene product with enhanced, negative, or new functions. This may be due to a change in the amino acid sequence of the protein that confers a new activity, or it may result from a mutation in a regulatory region of the gene, leading to expression of the gene at higher levels or at abnormal times or places. A **dominant negative mutation** may directly interfere with the function of the product of the wild-type allele. Often this occurs when the mutant nonfunctional gene product binds to the wild-type gene product, inactivating it.

The most easily observed mutations are those affecting a morphological trait. These mutations are known as **visible mutations** and are recognized by their ability to alter a normal or wild-type visible phenotype. For example, all of Mendel's pea characteristics and many genetic variations encountered in *Drosophila* fit this designation, since they cause obvious changes to the morphology of the organism.

Some mutations give rise to nutritional or biochemical effects. In bacteria and fungi, a typical **nutritional mutation** results in a loss of ability to synthesize an amino acid or vitamin. In humans, sickle-cell anemia and hemophilia are examples of diseases resulting from **biochemical mutations.** Although such mutations do not always affect morphological characters, they affect the function of proteins that can impinge on the well-being and survival of the affected individual.

Still another category consists of mutations that affect the behavior patterns of an organism. The primary effect of **behavioral mutations** is often difficult to analyze. For example, the mating behavior of a fruit fly may be impaired if it cannot beat its wings. However, the defect may be in the flight muscles, the nerves leading to them, or the brain, where the nerve impulses that initiate wing movements originate.

Another group of mutations—**regulatory mutations**—affect the regulation of gene expression. A mutation in a regulatory gene or a gene control region can disrupt normal regulatory processes and inappropriately activate or inactivate expression of a gene. For example, as we will see with the *lac* operon discussed later in the text (see Chapter 15), a regulatory gene produces a product that controls the transcription of the entire *lac* operon. Mutations within this regulatory gene can lead to the production of a regulatory protein with abnormal effects on the *lac* operon. Our knowledge of genetic regulation has been dependent on the study of such regulatory mutations. Regulatory mutations may also occur in regions such as splice junctions, promoters, or other regulatory regions of a gene that affect many aspects of gene regulation including transcription initiation, mRNA splicing, and mRNA stability.

It is also possible that a mutation may adversely affect a gene product that is essential to the survival of the organism. In this case, it is referred to as a **lethal mutation.** Various inherited human biochemical disorders are examples of lethal mutations. For example, Tay–Sachs disease and Huntington disease are caused by mutations that result in lethality, but at different points in the life cycle of humans.

Another interesting class of mutations exerts effects on the organism in ways that depend on the environment in which the organism finds itself. Such mutations are called **conditional mutations** because they are present in the genome of an organism but can be detected only under certain conditions. Among the best examples of conditional mutations are **temperature-sensitive mutations.** At a "permissive" temperature, the mutant gene product functions normally, but it loses its function at a different, "restrictive" temperature. Therefore, when the organism is shifted from the permissive to the restrictive temperature, the effect of the mutation becomes apparent. The temperature-sensitive coat color variations in Siamese cats and Himalayan rabbits, discussed earlier in the text (see Chapter 4), are striking examples of the effects of conditional mutations.

A **neutral mutation** is a mutation that can occur either in a protein-coding region or in any part of the genome, and its effect on the genetic fitness of the organism is negligible. For example, a neutral mutation within a gene may change a lysine codon (AAA) to an arginine codon (AGA). The two amino acids are chemically similar; therefore, this change may be insignificant to the function of the protein. Because eukaryotic genomes consist mainly of noncoding regions, the vast majority of mutations are likely to occur in the large portions of the genome that do not contain genes. These may be considered neutral mutations, if they do not affect gene products or gene expression.

Classification Based on Location of Mutation

Mutations may be classified according to the cell type or chromosomal locations in which they occur. **Somatic mutations** are those occurring in any cell in the body

except germ cells. **Autosomal mutations** are mutations within genes located on the autosomes, whereas **X-linked** and **Y-linked mutations** are those within genes located on the X or Y chromosome, respectively.

Mutations arising in somatic cells are not transmitted to future generations. When a recessive autosomal mutation occurs in a somatic cell of a diploid organism, it is unlikely to result in a detectable phenotype. The expression of most such mutations is likely to be masked by expression of the wild-type allele within that cell. Somatic mutations will have a greater impact if they are dominant or, in males, if they are X-linked, since such mutations are most likely to be immediately expressed. Similarly, the impact of dominant or X-linked somatic mutations will be more noticeable if they occur early in development, when a small number of undifferentiated cells replicate to give rise to several differentiated tissues or organs. Dominant mutations that occur in cells of adult tissues are often masked by the activity of thousands upon thousands of nonmutant cells in the same tissue that perform the nonmutant function.

Mutations in germ cells are of greater significance because they may be transmitted to offspring as gametes. They have the potential of being expressed in all cells of an offspring. Inherited dominant autosomal mutations will be expressed phenotypically in the first generation. X-linked recessive mutations arising in the gametes of a **homogametic** female may be expressed in hemizygous male offspring. This will occur provided that the male offspring receives the affected X chromosome. Because of heterozygosity, the occurrence of an autosomal recessive mutation in the gametes of either males or females (even one resulting in a lethal allele) may go unnoticed for many generations, until the resultant allele has become widespread in the population. Usually, the new allele will become evident only when a chance mating brings two copies of it together into the homozygous condition.

Spontaneous and Induced Mutations

Mutations can be classified as either spontaneous or induced, although these two categories overlap to some degree. **Spontaneous mutations** are changes in the nucleotide sequence of genes that appear to occur naturally. No specific agents are associated with their occurrence, and they are generally assumed to be accidental. Many of these mutations arise as a result of normal biological or chemical processes in the organism that alter the structure of nitrogenous bases. Often, spontaneous mutations occur during the enzymatic process of DNA replication, as we discuss later in this chapter.

In contrast to spontaneous mutations, mutations that result from the influence of extraneous factors are considered to be **induced mutations.** Induced mutations may be the result of either natural or artificial agents. For example, radiation from cosmic and mineral sources and ultraviolet radiation from the sun are energy sources to which most organisms are exposed and, as such, may be factors that cause induced mutations.

The earliest demonstration of the artificial induction of mutations occurred in 1927, when Hermann J. Muller reported that X rays could cause mutations in *Drosophila.* In 1928, Lewis J. Stadler reported that X rays had the same effect on barley. In addition to various forms of radiation, numerous natural and synthetic chemical agents are also mutagenic.

Several generalizations can be made regarding spontaneous mutation rates in organisms. The **mutation rate** is defined as the likelihood that a gene will undergo a mutation in a single generation or in forming a single gamete. First, the rate of spontaneous mutation is exceedingly low for all organisms. Second, the rate varies between different organisms. Third, even within the same species, the spontaneous mutation rate varies from gene to gene.

Viral and bacterial genes undergo spontaneous mutation at an average of about 1 in 100 million (10^{-8}) replications or cell divisions. Maize and *Drosophila* demonstrate rates several orders of magnitude higher. The genes studied in these groups average between 1 in 1,000,000 (10^{-6}) and 1 in 100,000 (10^{-5}) mutations per gamete formed. Some mouse genes are another order of magnitude higher in their spontaneous mutation rate, 1 in 100,000 to 1 in 10,000 (10^{-5} to 10^{-4}). It is not clear why such large variations occur in mutation rates. The variation between genes in a given organism may be due to inherent differences in mutability in different regions of the genome. Some DNA sequences appear to be highly susceptible to mutation and are known as **mutation hot spots.** The variation between organisms may, in part, reflect the relative efficiencies of their DNA proofreading and repair systems. We will discuss these systems later in the chapter.

ESSENTIAL POINT

Mutations can be spontaneous or induced, somatic or germ-line, autosomal or X-linked. They can have many different effects on gene function, depending on the type of nucleotide changes that comprise the mutation. Phenotypic effects can range from neutral or silent to loss of function or gain of function to lethality. ■

NOW SOLVE THIS

14–1 If one spontaneous mutation occurs within a human egg cell genome, and this mutation changes an A to a T, what is the most likely effect of this mutation on the phenotype of an offspring that develops from this mutated egg?

■ **HINT:** *This problem asks you to predict the effects of a single base-pair mutation on phenotype. The key to its solution involves an understanding of the organization of the human genome as well as the effects of mutations on coding and noncoding regions of genes, and the effects of mutations on development.*

14.2 Spontaneous Mutations Arise from Replication Errors and Base Modifications

In this section, we will outline some of the processes that lead to spontaneous mutations. It is useful to keep in mind, however, that many of the DNA changes that occur during spontaneous mutagenesis also occur, at a higher rate, during induced mutagenesis.

DNA Replication Errors and Slippage

As we learned earlier in the text (see Chapter 10), the process of DNA replication is imperfect. Occasionally, DNA polymerases insert incorrect nucleotides during replication of a strand of DNA. Although DNA polymerases can correct most of these replication errors using their inherent 3′ to 5′ exonuclease proofreading capacity, misincorporated nucleotides may persist after replication. If these errors are not detected and corrected by DNA repair mechanisms, they may lead to mutations. Replication errors due to mispairing predominantly lead to point mutations. The fact that bases can take several forms, known as **tautomers,** increases the chance of mispairing during DNA replication, as we explain next.

In addition to mispairing and point mutations, DNA replication can lead to the introduction of small insertions or deletions. These mutations can occur when one strand of the DNA template loops out and becomes displaced during replication, or when DNA polymerase slips or stutters during replication. If a loop occurs in the template strand during replication, DNA polymerase may miss the looped-out nucleotides, and a small deletion in the new strand will be introduced. If DNA polymerase repeatedly introduces nucleotides that are not present in the template strand, an insertion of one or more nucleotides will occur, creating an unpaired loop on the newly synthesized strand. Insertions and deletions may lead to frameshift mutations, or amino acid insertions or deletions in the gene product.

Replication slippage can occur anywhere in the DNA but seems distinctly more common in regions containing tandemly repeated sequences. Repeat sequences are hot spots for DNA mutation and in some cases contribute to hereditary diseases, such as fragile-X syndrome and Huntington disease. The hypermutability of repeat sequences in noncoding regions of the genome is the basis for current methods of forensic DNA analysis.

Tautomeric Shifts

Purines and pyrimidines can exist in tautomeric forms—that is, in alternate chemical forms that differ by only a single proton shift in the molecule. The biologically important tautomers are the keto–enol forms of thymine and guanine and the amino–imino forms of cytosine and adenine. These shifts change the bonding structure of the molecule, allowing hydrogen bonding with noncomplementary bases. Hence, **tautomeric shifts** may lead to permanent base-pair changes and mutations. **Figure 14–2** compares

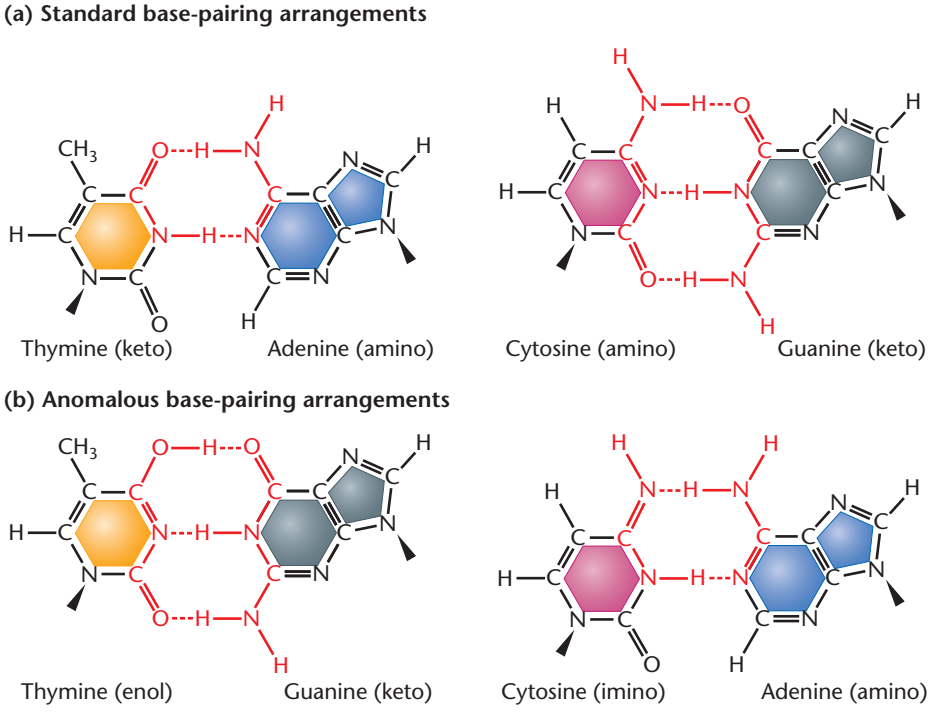


FIGURE 14–2 Standard base-pairing relationships (a) compared with examples of the anomalous base-pairing that occurs as a result of tautomeric shifts (b). The long triangle indicates the point at which the base bonds to the pentose sugar.

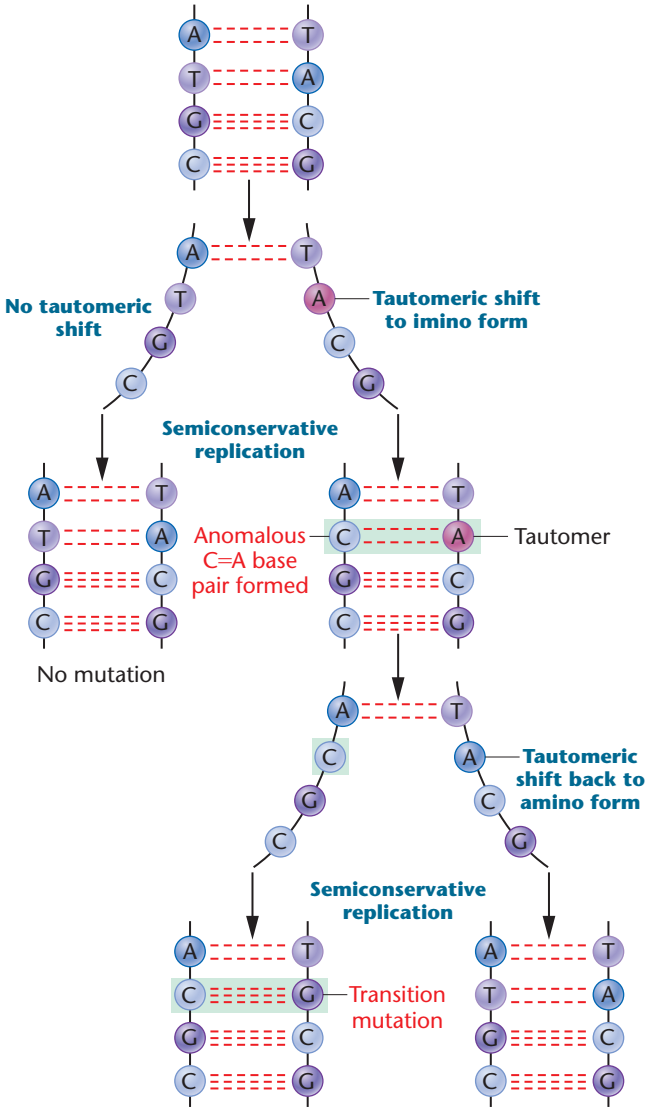


FIGURE 14–3 Formation of an A═T to G≡C transition mutation as a result of a tautomeric shift in adenine.

normal base-pairing relationships with rare unorthodox pairings. Anomalous T≡G and C═A pairs, among others, may be formed.

A mutation occurs during DNA replication when a transiently formed tautomer in the template strand pairs with a noncomplementary base. In the next round of replication, the "mismatched" members of the base pair are separated, and each becomes the template for its normal complementary base. The end result is a point mutation (**Figure 14–3**).

Depurination and Deamination

Some of the most common causes of spontaneous mutations are two forms of DNA base damage: depurination and deamination. **Depurination** is the loss of one of the nitrogenous bases in an intact double-helical DNA molecule. Most frequently, the base is a purine—either guanine or adenine. These bases may be lost if the glycosidic bond linking the 1′-C of the deoxyribose and the number 9 position of the purine ring is broken, leaving an **apurinic site** on one strand of the DNA. Geneticists estimate that thousands of such spontaneous lesions are formed daily in the DNA of mammalian cells in culture. If apurinic sites are not repaired, there will be no base at that position to act as a template during DNA replication. As a result, DNA polymerase may introduce a nucleotide at random at that site.

In **deamination,** an amino group in cytosine or adenine is converted to a keto group (**Figure 14–4**). In these cases, cytosine is converted to uracil, and adenine is changed to hypoxanthine. The major effect of these changes is an alteration in the base-pairing specificities of these two bases during DNA replication. For example, cytosine normally pairs with guanine. Following its conversion to uracil, which pairs with adenine, the original G≡C pair is converted to an A═U pair and then, in the next replication, is converted to an A═T pair. When adenine is deaminated, the original A═T pair is converted to a G≡C pair because hypoxanthine pairs naturally with cytosine. Deamination may occur spontaneously or as a result of treatment with chemical mutagens such as nitrous acid (HNO_2).

Oxidative Damage

DNA may also suffer damage from the by-products of normal cellular processes. These by-products include

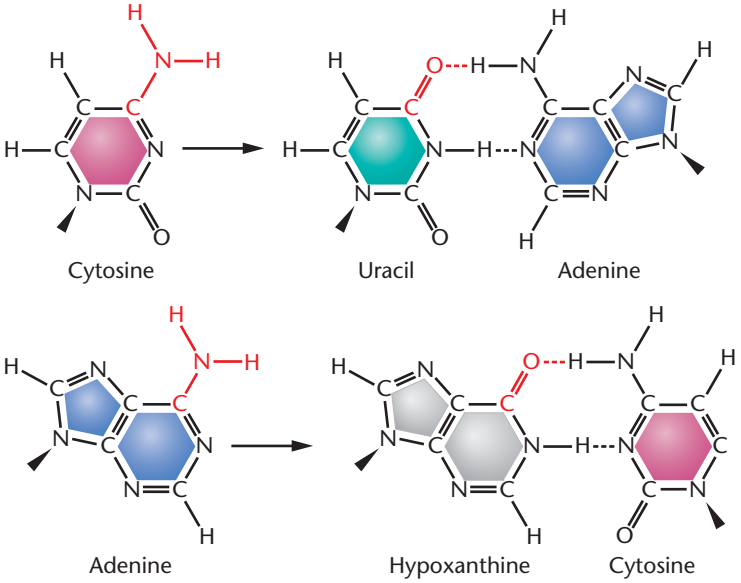


FIGURE 14–4 Deamination of cytosine and adenine, leading to new base pairing and mutation. Cytosine is converted to uracil, which base-pairs with adenine. Adenine is converted to hypoxanthine, which base-pairs with cytosine.

reactive oxygen species (electrophilic oxidants) that are generated during normal aerobic respiration. For example, superoxides (O_2^-), hydroxyl radicals (·OH), and hydrogen peroxide (H_2O_2) are created during cellular metabolism and are constant threats to the integrity of DNA. Such **reactive oxidants,** also generated by exposure to high-energy radiation, can produce more than 100 different types of chemical modifications in DNA, including modifications to bases, loss of bases, and single-stranded breaks.

ESSENTIAL POINT

Spontaneous mutations occur in many ways, ranging from errors during DNA replication to changes in DNA base pairing caused by tautomeric shifts, depurinations, deaminations, and reactive oxidant damage. ■

NOW SOLVE THIS

14–2 One of the most famous cases of an X-linked recessive mutation in humans is that of hemophilia found in the descendants of Britain's Queen Victoria. The pedigree of the royal family indicates that Victoria was heterozygous for the trait; however, her father was not affected, and there is no evidence that her mother was a carrier. What are some possible explanations of how the mutation arose? What types of mutations could lead to the disease?

■ **HINT:** *This problem asks you to determine the sources of new mutations. The key to its solution is to consider the ways in which mutations occur, the types of cells in which they can occur, and how they are inherited.*

14.3 Induced Mutations Arise from DNA Damage Caused by Chemicals and Radiation

Induced mutations are those that increase the rate of mutation above the spontaneous background. All cells on Earth are exposed to a plethora of agents called **mutagens,** which have the potential to damage DNA and cause induced mutations. Some of these agents, such as some fungal toxins, cosmic rays, and ultraviolet light, are natural components of our environment. Others, including some industrial pollutants, medical X rays, and chemicals within tobacco smoke, can be considered as unnatural or human-made additions to our modern world. On the positive side, geneticists harness some mutagens for use in analyzing genes and gene functions. The mechanisms by which some of these natural and unnatural agents lead to mutations are outlined in this section.

Base Analogs

One category of mutagenic chemicals is **base analogs,** compounds that can substitute for purines or pyrimidines during nucleic acid biosynthesis. For example, **5-bromouracil** (**5-BU**), a derivative of uracil, behaves as a thymine analog but is halogenated at the number 5 position of the pyrimidine ring. If 5-BU is chemically linked to deoxyribose, the nucleoside analog **bromodeoxyuridine** (**BrdU**) is formed. **Figure 14–5** compares the structure of this analog with that of thymine. The presence of the bromine atom in place of the methyl group increases the probability that a tautomeric shift will occur. If 5-BU is incorporated into DNA in place of thymine and a tautomeric shift to the enol form occurs, 5-BU base-pairs with guanine. After one round of replication, an A═T to G≡C transition results. Furthermore, the presence of 5-BU within DNA increases the sensitivity of the molecule to ultraviolet (UV) light, which itself is mutagenic.

There are other base analogs that are mutagenic. For example, **2-amino purine** (**2-AP**) can act as an analog of adenine. In addition to its base-pairing affinity with thymine, 2-AP can also base-pair with cytosine, leading to possible transitions from A═T to G≡C following replication.

Alkylating, Intercalating, and Adduct-Forming Agents

A number of naturally occurring and human-made chemicals alter the structure of DNA and cause mutations. The sulfur-containing mustard gases, discovered during World War I, were some of the first chemical mutagens identified in chemical warfare studies. Mustard gases are **alkylating agents**—that is, they donate an alkyl group, such as CH_3 or CH_3CH_2, to amino or keto groups in nucleotides. Ethylmethane sulfonate (EMS), for example, alkylates the keto groups in the number 6 position of guanine and in the number 4 position of thymine. As with base analogs, base-pairing affinities are altered, and transition mutations result. For example, 6-ethylguanine acts as an analog of adenine and pairs with thymine (**Figure 14–6**).

Intercalating agents are chemicals that have dimensions and shapes that allow them to wedge between the base pairs of DNA. When bound between base pairs, intercalating agents cause base pairs to distort and DNA strands to unwind. These changes in DNA structure affect many functions including transcription, replication, and repair. Deletions and insertions occur during DNA replication and repair, leading to frameshift mutations.

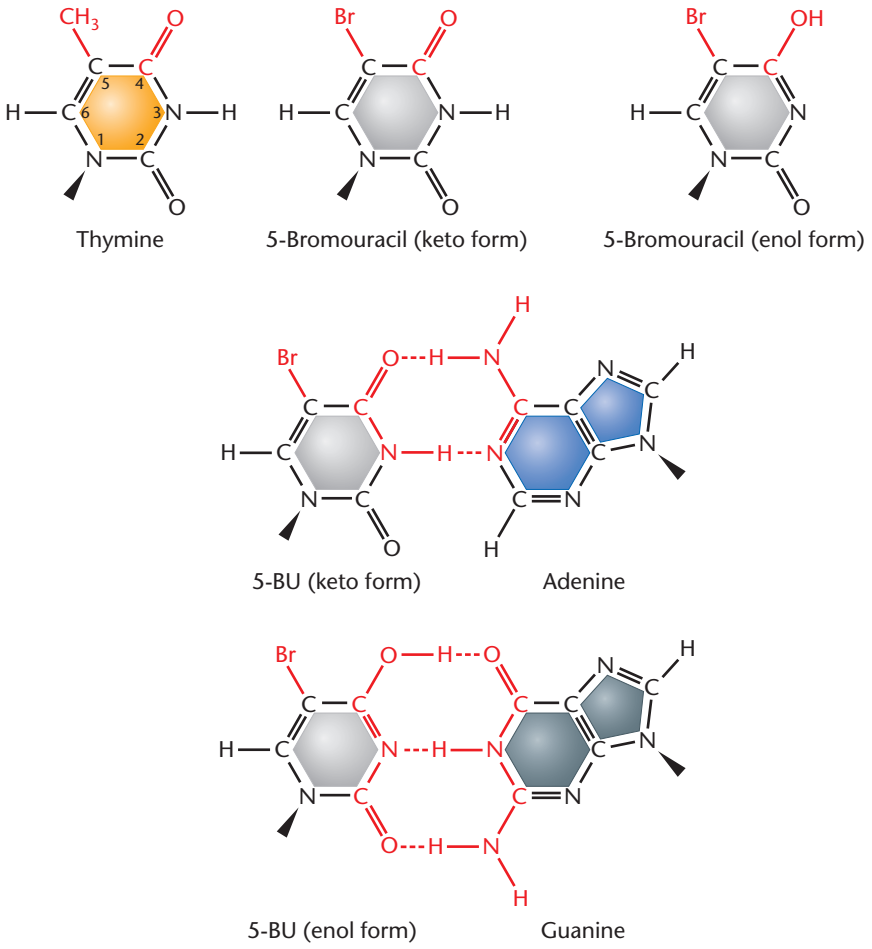


FIGURE 14–5 Similarity of the chemical structure of 5-bromouracil (5-BU) and thymine. In the common keto form, 5-BU base-pairs normally with adenine, behaving as a thymine analog. In the rare enol form, it pairs anomalously with guanine.

Another group of chemicals that cause mutations are known as adduct-forming agents. A DNA adduct is a substance that covalently binds to DNA, altering its conformation and interfering with replication and repair. Two examples of adduct-forming substances are acetaldehyde (a component of cigarette smoke) and heterocyclic amines (HCAs). HCAs are cancer-causing chemicals that are created during the cooking of meats such as beef, chicken, and fish. HCAs are formed at high temperatures from amino acids and creatine. Many HCAs covalently bind to guanine bases. At least 17 different HCAs have been linked to the development of cancers, such as those of the stomach, colon, and breast.

Ultraviolet Light

All electromagnetic radiation consists of energetic waves that we define by their different wavelengths (**Figure 14–7**). The full range of wavelengths is referred to as the **electromagnetic spectrum,** and the energy of any radiation in the spectrum varies inversely with its wavelength. Waves in the range of visible light and longer are benign when they interact with most organic molecules. However, waves of shorter length than visible light, being inherently more energetic, have the potential to disrupt organic molecules. As we know, purines and pyrimidines absorb **ultraviolet (UV) radiation** most intensely at a wavelength of about 260 nm. Although Earth's ozone layer absorbs the most dangerous types of UV radiation, sufficient UV radiation can induce thousands of DNA lesions per hour in any cell exposed to this radiation. One major effect of UV radiation on DNA is the creation of **pyrimidine dimers**—chemical species consisting of two identical pyrimidines—particularly ones consisting of two thymine residues (**Figure 14–8**). The dimers distort the DNA conformation and inhibit normal replication. As a result, errors can be introduced in the base sequence of DNA during replication. When UV-induced dimerization is extensive, it is responsible (at least in part) for the killing effects of UV radiation on cells.

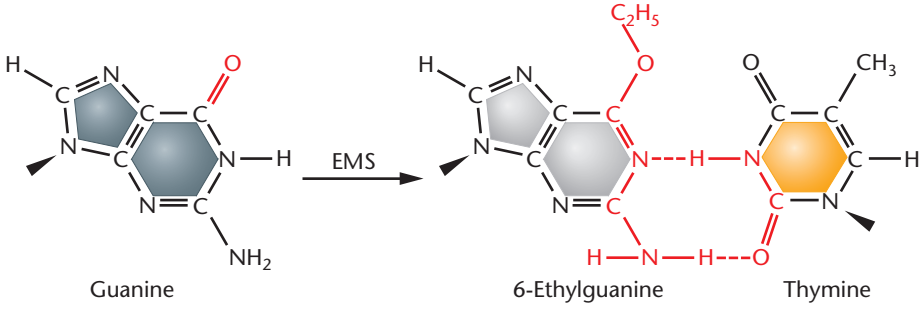


FIGURE 14–6 Conversion of guanine to 6-ethylguanine by the alkylating agent ethylmethane sulfonate (EMS). The 6-ethylguanine base-pairs with thymine.

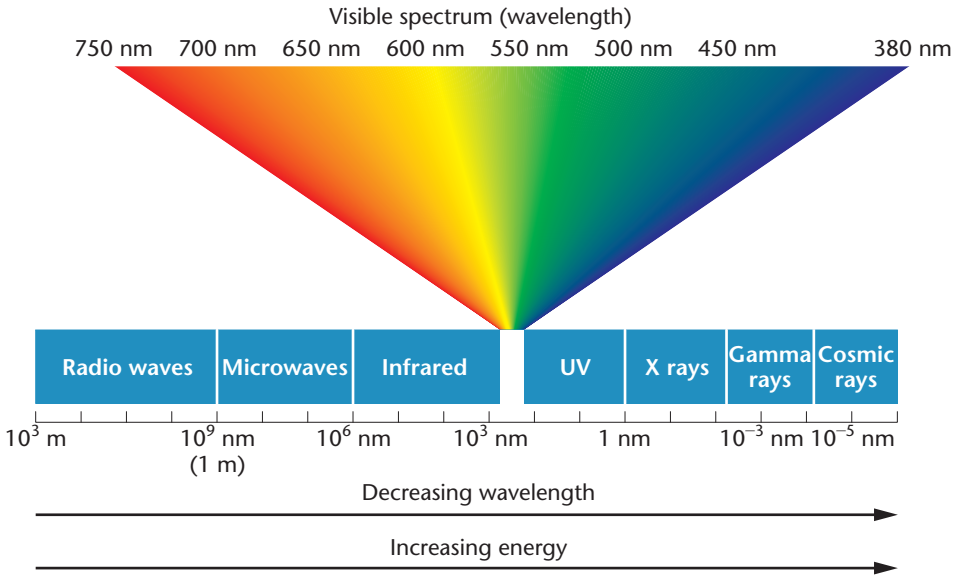


FIGURE 14–7 The regions of the electromagnetic spectrum and their associated wavelengths.

Ionizing Radiation

As noted above, the energy of radiation varies inversely with wavelength. Therefore, **X rays, gamma rays,** and **cosmic rays** are more energetic than UV radiation (Figure 14–7). As a result, they penetrate deeply into tissues, causing ionization of the molecules encountered along the way. Hence, this type of radiation is called **ionizing radiation.**

As ionizing radiation penetrates cells, stable molecules and atoms are transformed into **free radicals**—chemical species containing one or more unpaired electrons. Free radicals can directly or indirectly affect the genetic material, altering purines and pyrimidines in DNA, breaking phosphodiester bonds, disrupting the integrity of chromosomes, and producing a variety of chromosomal aberrations, such as deletions, translocations, and chromosomal fragmentation.

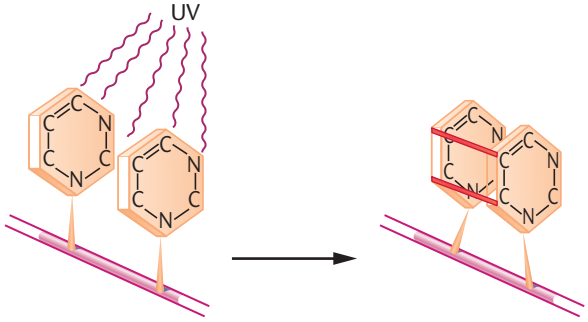


FIGURE 14–8 Induction of a thymine dimer by UV radiation, leading to distortion of the DNA. The covalent crosslinks occur between the atoms of the pyrimidine ring.

Research has shown that the relationship between ionizing radiation dose and mutation rate is linear. For each doubling of the dose, twice as many mutations are induced.

ESSENTIAL POINT

Mutations can be induced by many types of chemicals and radiation. These agents can damage both DNA bases and the sugar-phosphate backbones of DNA molecules. ■

NOW SOLVE THIS

14–3 The cancer drug melphalan is an alkylating agent of the mustard gas family. It acts in two ways: by causing alkylation of guanine bases and by cross linking DNA strands together. Describe two ways in which melphalan might kill cancer cells. What are two ways in which cancer cells could repair the DNA-damaging effects of melphalan?

■ **HINT:** *This problem asks you to consider the effect of the alkylation of guanine on base pairing during DNA replication. The key to its solution is to consider the effects of mutations on cellular processes that allow cells to grow and divide. In Section 14.6, you will learn about the ways in which cells repair the types of mutations introduced by alkylating agents.*

14.4 Single-Gene Mutations Cause a Wide Range of Human Diseases

Although most human genetic diseases are **polygenic**—that is, caused by variations in several genes—even a single base-pair change in one of the approximately 20,000 human genes can lead to a serious inherited disorder. These **monogenic** diseases can be caused by many different types

TABLE 14.1 Examples of Human Disorders Caused by Single-Gene Mutations

Type of DNA Mutation	Disorder	Molecular Change
Missense	Achondroplasia	Glycine to arginine at position 380 of *FGFR2* gene
Nonsense	Marfan syndrome	Tyrosine to STOP codon at position 2113 of *fibrillin-1* gene
Insertion	Familial hypercholesterolemia	Various short insertions throughout the *LDLR* gene
Deletion	Cystic fibrosis	Three-base-pair deletion of phenylalanine codon at position 508 of *CFTR* gene
Trinucleotide repeat expansions	Huntington disease	More than 40 repeats of (CAG) sequence in coding region of *Huntingtin* gene

of single-gene mutations. Table 14.1 lists some examples of the types of single-gene mutations that can lead to serious genetic diseases. A comprehensive database of human genes, mutations, and disorders is available in the Online Mendelian Inheritance in Man (OMIM) database, which is described in the "Exploring Genomics" feature earlier in the text (see Chapter 3). As of 2015, the OMIM database has catalogued more than 4400 human phenotypes for which the molecular basis is known.

Geneticists estimate that approximately 30 percent of mutations that cause human diseases are single base-pair changes that create nonsense mutations. These mutations not only code for a prematurely terminated protein product, but also trigger rapid decay of the mRNA. Many more mutations are missense mutations that alter the amino acid sequence of a protein and frameshift mutations that alter the protein sequence and create internal nonsense codons. Other common disease-associated mutations affect the sequences of gene promoters, mRNA splicing signals, and other noncoding sequences that affect transcription, processing, and stability of mRNA or protein. One recent study showed that about 15 percent of all point mutations that cause human genetic diseases result in abnormal mRNA splicing. Approximately 85 percent of these splicing mutations alter the sequence of 5′ and 3′ splice signals. The remainder create new splice sites within the gene. Splicing defects often result in degradation of the abnormal mRNA or creation of abnormal protein products.

Another type of single-gene mutation is caused by expansions of **trinucleotide repeat sequences**—specific short DNA sequences repeated many times. Normal individuals have a low number of repetitions of these sequences; however, individuals with over 20 different human disorders appear to have abnormally large numbers of repeat sequences—in some cases, over 200—within and surrounding specific genes.

Examples of diseases associated with these trinucleotide repeat expansions are fragile-X syndrome (discussed in Chapter 6), myotonic dystrophy, and Huntington disease (discussed in Chapter 4). When trinucleotide repeats such as $(CAG)_n$ occur within a coding region, they can be translated into long tracks of glutamine. These glutamine tracks may cause the proteins to aggregate abnormally. When the repeats occur outside coding regions, but within the mRNA, it is thought that the mRNAs may act as "toxic" RNAs that bind to important regulatory proteins, sequestering them away from their normal functions in the cell. Another possible consequence of long trinucleotide repeats is that the regions of DNA containing the repeats may become abnormally methylated, leading to silencing of gene transcription.

The mechanisms by which the repeated sequences expand from generation to generation are of great interest. It is thought that expansion may result from errors during either DNA replication or DNA damage repair. Whatever the cause may be, the presence of these short and unstable repeat sequences seems to be prevalent in humans and in many other organisms.

14.5 Organisms Use DNA Repair Systems to Detect and Correct Mutations

Living systems have evolved a variety of elaborate repair systems that counteract both spontaneous and induced DNA damage. These **DNA repair** systems are absolutely essential to the maintenance of the genetic integrity of organisms and, as such, to the survival of organisms on Earth. The balance between mutation and repair results in the observed mutation rates of individual genes and organisms. In addition, DNA repair systems correct the genetic damage that would otherwise result in human genetic diseases and cancer. The link between defective DNA repair and cancer susceptibility is described in detail later in the text (see Chapter 16).

We now embark on a review of some systems of DNA repair, with the emphasis on the major approaches that organisms use to counteract genetic damage.

Proofreading and Mismatch Repair

Some of the most common types of mutations arise during DNA replication when an incorrect nucleotide is inserted by DNA polymerase. The major DNA synthesizing enzyme in bacteria (**DNA polymerase III**) makes an error approximately once every 100,000 insertions, leading to an error rate of 10^{-5}. Fortunately, DNA polymerase proofreads each step, catching 99 percent of those errors. If an incorrect nucleotide is inserted during polymerization, the enzyme can recognize the error and "reverse" its direction. It then behaves as a 3′ to 5′ exonuclease, cutting out the incorrect nucleotide and replacing it with the correct one. This improves the efficiency of replication 100-fold, creating only 1 mismatch in every 10^7 insertions, for a final error rate of 10^{-7}.

To deal with errors such as base–base mismatches, small insertions, and deletions that remain after proofreading, another mechanism, called **mismatch repair,** may be activated. During mismatch repair, the mismatches are detected, the incorrect nucleotide is removed, and the correct nucleotide is inserted in its place. But how does the repair system recognize which nucleotide is correct (on the template strand) and which nucleotide is incorrect (on the newly synthesized strand)? If the mismatch is recognized but no such discrimination occurs, the excision will be random, and the strand bearing the correct base will be clipped out 50 percent of the time. Hence, strand discrimination is a critical step.

The process of strand discrimination has been elucidated in some bacteria, including *E. coli,* and is based on **DNA methylation.** These bacteria contain an enzyme, **adenine methylase,** which recognizes the DNA sequence

5′——GATC——3′
3′——CTAG——5′

as a substrate, adding a methyl group to each of the adenine residues during DNA replication.

Following replication, the newly synthesized DNA strand remains temporarily unmethylated, as the adenine methylase lags behind the DNA polymerase. Prior to methylation, the repair enzyme recognizes the mismatch and binds to the unmethylated (newly synthesized) DNA strand. An **endonuclease** enzyme creates a nick in the backbone of the unmethylated DNA strand, either 5′ or 3′ to the mismatch. An **exonuclease** unwinds and degrades the nicked DNA strand, until the region of the mismatch is reached. Finally, DNA polymerase fills in the gap created by the exonuclease, using the correct DNA strand as a template. DNA ligase then seals the gap.

A series of *E. coli* gene products, MutH, MutL, and MutS, as well as exonucleases, DNA polymerase III and ligase, are involved in mismatch repair. Mutations in the *MutH*, *MutL,* and *MutS* genes result in bacterial strains deficient in mismatch repair. While the preceding mechanism occurs in *E. coli,* similar mechanisms involving homologous proteins exist in yeast and in mammals.

In humans, mutations in genes that code for DNA mismatch repair proteins (such as the *hMSH2* and *hMLH1,* which are the human equivalents of the *MutS* and *MutL* genes of *E. coli*) are associated with the hereditary nonpolyposis colon cancer. Mismatch repair defects are commonly found in other cancers, such as leukemias, lymphomas, and tumors of the ovary, prostate, and endometrium. Cells from these cancers show genome-wide increases in the rate of spontaneous mutation. The link between defective mismatch repair and cancer is supported by experiments with mice. Mice that are engineered to have deficiencies in mismatch repair genes accumulate large numbers of mutations and are cancer-prone.

Postreplication Repair and the SOS Repair System

Another DNA repair system, called **postreplication repair,** responds *after* damaged DNA has escaped repair and has failed to be completely replicated. As illustrated in **Figure 14–9**, when DNA bearing a lesion of some sort (such as a pyrimidine dimer) is being replicated, DNA polymerase may stall at the lesion and then skip over it, leaving an unreplicated gap on the newly synthesized strand. To correct the gap, the RecA protein directs a recombinational exchange with the corresponding region on the undamaged parental strand of the same polarity (the "donor" strand). When the undamaged segment of the donor strand DNA replaces the gapped segment, a gap is created on the donor strand. The gap can be filled by repair synthesis as replication proceeds. Because a recombinational event is involved in this type of DNA repair, it is considered to be a form of homologous recombination repair.

Still another repair pathway, the *E. coli* **SOS repair system,** also responds to damaged DNA, but in a different way. In the presence of a large number of unrepaired DNA mismatches and gaps, bacteria can induce the expression of about 20 genes (including *lexA, recA,* and *uvr*) whose products allow DNA replication to occur even in the presence of these lesions. This type of repair is a last resort to minimize DNA damage, hence its name. During SOS repair, DNA synthesis becomes error-prone, inserting random and possibly incorrect nucleotides in places that would normally stall DNA replication. As a result, SOS repair itself becomes mutagenic—although it may allow the cell to survive DNA damage that would otherwise kill it.

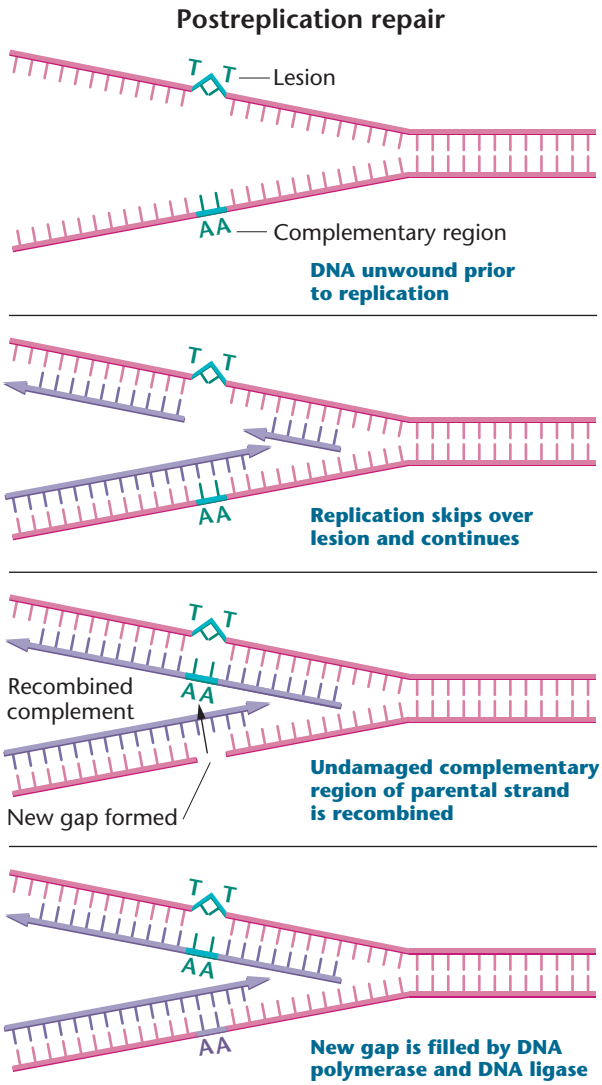


FIGURE 14–9 Postreplication repair occurs if DNA replication has skipped over a lesion such as a thymine dimer. Through the process of recombination, the correct complementary sequence is recruited from the parental strand and inserted into the gap opposite the lesion. The new gap is filled by DNA polymerase and DNA ligase.

Photoreactivation Repair: Reversal of UV Damage

As illustrated in Figure 14–8, UV light is mutagenic as a result of the creation of pyrimidine dimers. UV-induced damage to *E. coli* DNA can be partially reversed if, following irradiation, the cells are exposed briefly to light in the blue range of the visible spectrum. The process is dependent on the activity of a protein called **photoreactivation enzyme** (**PRE**). The enzyme's mode of action is to cleave the bonds between thymine dimers, thus directly reversing the effect of UV radiation on DNA (**Figure 14–10**). Although the enzyme will associate with a thymine dimer in the dark, it must absorb a photon of light to cleave the dimer. In spite of its ability to reduce the number of UV-induced mutations, **photoreactivation repair** is not absolutely essential in *E. coli*; we know this because a mutation creating a null allele in the gene coding for PRE is not lethal. Nonetheless, the enzyme is detectable in many organisms, including bacteria, fungi, plants, and some vertebrates—though not in humans. Humans and other organisms that lack photoreactivation repair must rely on other repair mechanisms to reverse the effects of UV radiation.

Base and Nucleotide Excision Repair

A number of light-independent DNA repair systems exist in all prokaryotes and eukaryotes. The basic mechanisms involved in these types of repair—collectively referred to as **excision repair** or cut-and-paste mechanisms—consist of the following three steps.

1. The distortion or error present on one of the two strands of the DNA helix is recognized and enzymatically clipped out by an endonuclease. Excisions in the phosphodiester backbone usually include a number of nucleotides adjacent to the error as well, leaving a gap on one strand of the helix.
2. A DNA polymerase fills in the gap by inserting nucleotides complementary to those on the intact strand, which

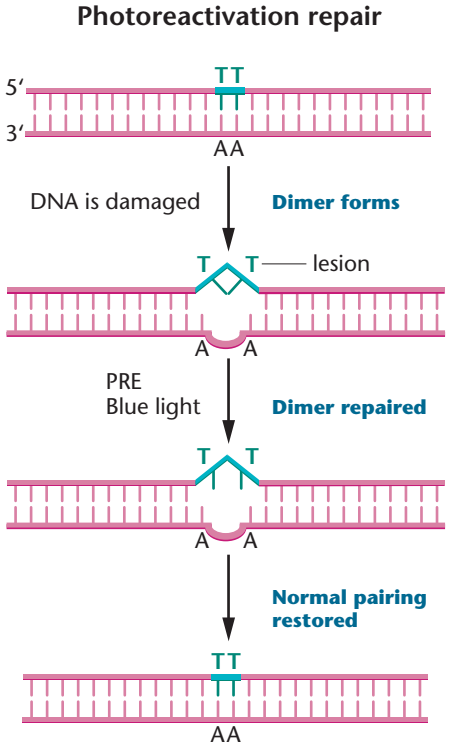


FIGURE 14–10 Damaged DNA repaired by photoreactivation repair. The bond creating the thymine dimer is cleaved by the photoreactivation enzyme (PRE), which must be activated by blue light in the visible spectrum.

it uses as a replicative template. The enzyme adds these nucleotides to the free 3′-OH end of the clipped DNA. In *E. coli,* this step is usually performed by DNA polymerase I.

3. DNA ligase seals the final "nick" that remains at the 3′-OH end of the last nucleotide inserted, closing the gap.

There are two types of excision repair: base excision repair and nucleotide excision repair. **Base excision repair** (**BER**) corrects DNA that contains a damaged DNA base. The first step in the BER pathway in *E. coli* involves the recognition of the altered base by an enzyme called **DNA glycosylase.** There are a number of DNA glycosylases, each of which recognizes a specific base (**Figure 14–11**). For example, the enzyme uracil DNA glycosylase recognizes the presence of uracil in DNA. DNA glycosylases first cut the glycosidic bond between the base and the sugar, creating an **apyrimidinic** or **apurinic site.** The sugar with the missing base is then recognized by an enzyme called **AP endonuclease.** The AP endonuclease makes a cut in the phosphodiester backbone at the apyrimidinic or apurinic site. Endonucleases then remove the deoxyribose sugar, and the gap is filled by DNA polymerase and DNA ligase.

Although much has been learned about the mechanisms of BER in *E. coli,* BER systems have also been detected in eukaryotes from yeast to humans. Experimental evidence shows that both mouse and human cells that are defective in BER activity are hypersensitive to the killing effects of gamma rays and oxidizing agents.

Nucleotide excision repair (**NER**) pathways repair "bulky" lesions in DNA that alter or distort the double helix. These lesions include the UV-induced pyrimidine dimers and DNA adducts discussed previously.

The NER pathway (**Figure 14–12**) was first discovered in *E. coli* by Paul Howard-Flanders and coworkers, who isolated several independent mutants that are sensitive to UV radiation. One group of genes was designated *uvr* (ultraviolet repair) and included the *uvrA, uvrB,* and *uvrC* mutations. In the NER pathway, the *uvr* gene products are involved in recognizing and clipping out lesions in the DNA. Usually, a specific number of nucleotides is clipped out around both sides of the lesion. In *E. coli,* usually a total of 13 nucleotides is removed, including the lesion. The repair is then completed by DNA polymerase I and DNA ligase, in a manner similar to that occurring in BER. The undamaged strand opposite the lesion is used as a template for the replication, resulting in repair.

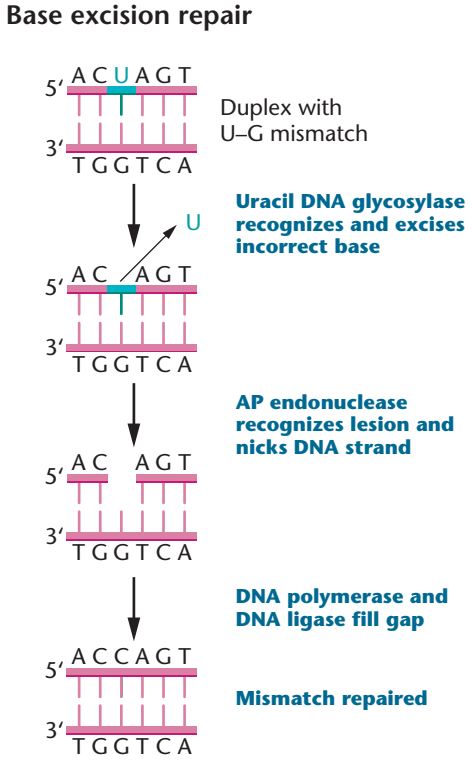


FIGURE 14–11 Base excision repair (BER) accomplished by uracil DNA glycosylase, AP endonuclease, DNA polymerase, and DNA ligase. Uracil is recognized as a noncomplementary base, excised, and replaced with the complementary base (C).

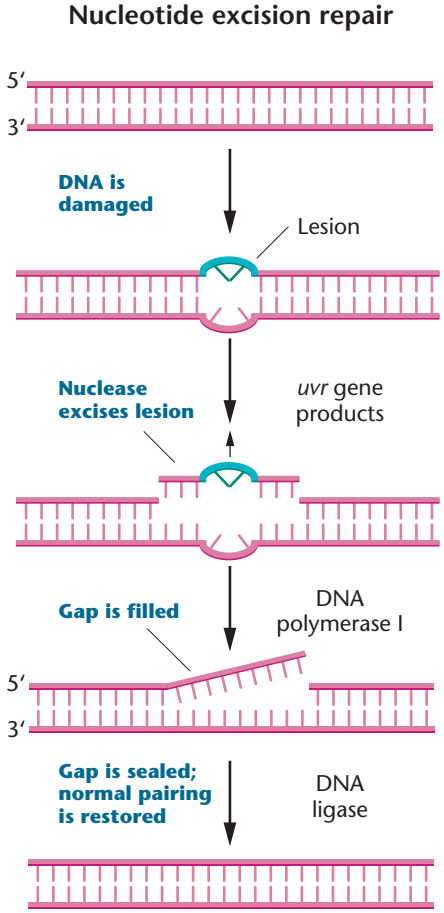


FIGURE 14–12 Nucleotide excision repair (NER) of a UV-induced thymine dimer. During repair, 13 nucleotides are excised in prokaryotes, and 28 nucleotides are excised in eukaryotes.

Nucleotide Excision Repair and Xeroderma Pigmentosum in Humans

The mechanism of NER in eukaryotes is much more complicated than that in prokaryotes and involves many more proteins, encoded by about 30 genes. Much of what is known about the system in humans has come from detailed studies of individuals with **xeroderma pigmentosum** (**XP**), a rare recessive genetic disorder that predisposes individuals to severe skin abnormalities, skin cancers, and a wide range of other symptoms including developmental and neurological defects. Patients with XP are extremely sensitive to UV radiation in sunlight. In addition, they have a 2000-fold higher rate of cancer, particularly skin cancer, than the general population. The condition is severe and may be lethal, although early detection and protection from sunlight can arrest it (**Figure 14–13**).

The repair of UV-induced lesions in XP has been investigated *in vitro,* using human fibroblast cell cultures derived from normal individuals and those with XP. (Fibroblasts are undifferentiated connective tissue cells.) The results of these studies suggest that the XP phenotype is caused by defects in NER pathways and by mutations in more than one gene.

In 1968, James Cleaver showed that cells from XP patients were deficient in DNA synthesis other than that occurring during chromosome replication—a phenomenon known as **unscheduled DNA synthesis.** Unscheduled DNA synthesis is elicited in normal cells by UV radiation. Because this type of synthesis is thought to represent the activity of DNA polymerization during NER, the lack of unscheduled DNA synthesis in XP patients suggested that XP may be a deficiency in NER.

FIGURE 14–13 Two individuals with xeroderma pigmentosum. These XP patients show characteristic XP skin lesions induced by sunlight, as well as mottled redness (erythema) and irregular pigment changes to the skin, in response to cellular injury.

The involvement of multiple genes in NER and XP was further investigated by studies using **somatic cell hybridization.** Fibroblast cells from any two unrelated XP patients, when grown together in tissue culture, can fuse together, forming heterokaryons. A **heterokaryon** is a single cell with two nuclei from different organisms but a common cytoplasm. NER in the heterokaryon can be measured by the level of unscheduled DNA synthesis. If the mutation in each of the two XP cells occurs in the same gene, the heterokaryon, like the cells that fused to form it, will still be unable to undergo NER. This is because there is no normal copy of the relevant gene present in the heterokaryon. However, if NER does occur in the heterokaryon, the mutations in the two XP cells must have been present in two different genes. Hence, the two mutants are said to demonstrate **complementation,** a concept also discussed earlier in the text (see Chapter 4). Complementation occurs because the heterokaryon has at least one normal copy of each gene in the fused cell. By fusing XP cells from a large number of XP patients, researchers were able to determine how many genes contribute to the XP phenotype.

Based on these and other studies, XP patients were divided into seven complementation groups, indicating that at least seven different genes are involved in nucleotide excision repair in humans. A gene representing each of these complementation groups, *XPA* to *XPG* (*X*eroderma *P*igmentosum gene *A* to *G*), has now been identified, and a homologous gene for each has been identified in yeast. Approximately 20 percent of XP patients do not fall into any of the seven complementation groups. Cells from most of these patients have mutations in the gene coding for DNA polymerase η and are defective in repair DNA synthesis.

As a result of the study of defective genes in XP, a great deal is now known about how NER counteracts DNA damage in normal cells. The first step in humans is recognition of the damaged DNA by proteins encoded by the *XPC, XPE,* and *XPA* genes. These proteins then recruit the remainder of the repair proteins to the site of DNA damage. The *XPB* and *XPD* genes encode helicases, and the *XPF* and *XPG* genes encode nucleases. The excision repair complex containing these and other factors is responsible for the excision of an approximately 28-nucleotide-long fragment from the DNA strand that contains the lesion.

Double-Strand Break Repair in Eukaryotes

Thus far, we have discussed repair pathways that deal with damage or errors within one strand of DNA. We conclude our discussion of DNA repair by considering what happens when both strands of the DNA helix are cleaved—as a result of exposure to ionizing radiation, for example. These

types of damage are extremely dangerous to cells, leading to chromosome rearrangements, cancer, or cell death. In this section, we will discuss double-strand breaks in eukaryotic cells.

Specialized forms of DNA repair, the DNA **double-strand break repair** (**DSB repair**) pathways, are activated and are responsible for reattaching two broken DNA strands. Recently, interest in DSB repair has grown because defects in these pathways are associated with X-ray hypersensitivity and immune deficiency. Such defects may also underlie familial disposition to breast and ovarian cancer. Several human disease syndromes, such as Fanconi anemia and ataxia telangiectasia, result from defects in DSB repair.

One pathway involved in double-strand break repair is **homologous recombination repair.** The first step in this process involves the activity of an enzyme that recognizes the double-strand break, and then digests back the 5′-ends of the broken DNA helix, leaving overhanging 3′-ends (**Figure 14–14**). One overhanging end searches for a region of sequence complementarity on the sister chromatid and then invades the homologous DNA duplex, aligning the complementary sequences. Once aligned, DNA synthesis proceeds from the 3′ overhanging ends, using the undamaged DNA strands as templates. The interaction of two sister chromatids is necessary because, when both strands of one helix are broken, there is no undamaged parental DNA strand available to use as a source of the complementary template DNA sequence during repair. After DNA repair synthesis, the resulting heteroduplex molecule is resolved and the two chromatids separate.

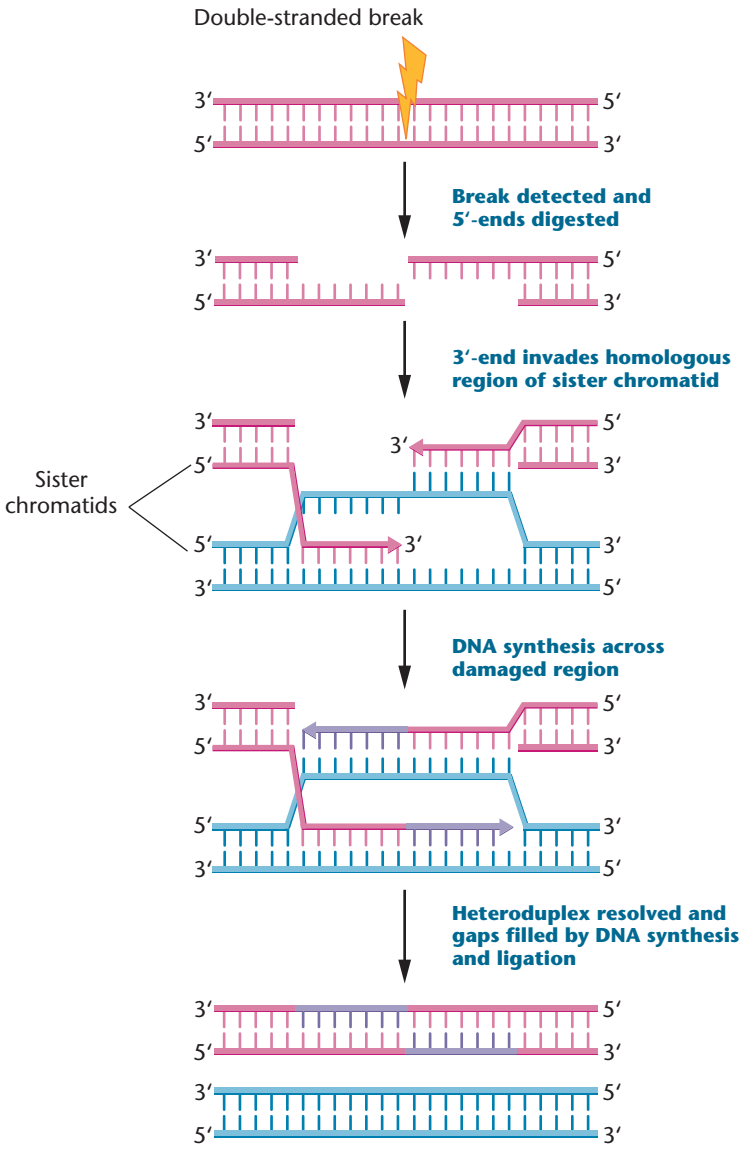


FIGURE 14–14 Steps in homologous recombination repair of double-stranded breaks.

DSB repair usually occurs during the late S or early G2 phase of the cell cycle, after DNA replication, a time when sister chromatids are available to be used as repair templates. Because an undamaged template is used during repair synthesis, homologous recombination repair is an accurate process.

A second pathway, called **nonhomologous end joining,** also repairs double-strand breaks. However, as the name implies, the mechanism does not recruit a homologous region of DNA during repair. This system is activated in G1, prior to DNA replication. End joining involves a complex of many proteins, and may include the DNA-dependent protein kinase and the breast cancer susceptibility gene product, BRCA1. These and other proteins bind to the free ends of the broken DNA, trim the ends, and ligate them back together. Because some nucleotide sequences are lost in the process of end joining, it is an error-prone repair system. In addition, if more than one chromosome suffers a double-strand break, the wrong ends could be joined together, leading to abnormal chromosome structures, such as those discussed earlier in the text (see Chapter 6).

NOW SOLVE THIS

14–4 Geneticists often use ethylmethane sulfonate (EMS) to induce mutations in *Drosophila*. Why is EMS a mutagen of choice for genetic research? What would be the effects of EMS in a strain of *Drosophila* lacking functional mismatch repair systems?

HINT: *This problem asks you to evaluate EMS as a useful mutagen and to determine its effects in the absence of DNA repair. The key to its solution is to consider the chemical effects of EMS on DNA. Also, consider the types of DNA repair that may operate on EMS-mutated DNA and the efficiency of these processes.*

ESSENTIAL POINT

Organisms counteract mutations by using a range of DNA repair systems. Errors in DNA synthesis can be repaired by proofreading, mismatch repair, and postreplication repair. DNA damage can be repaired by photoreactivation repair, SOS repair, base excision repair, nucleotide excision repair, and double-strand break repair. ■

14.6 The Ames Test Is Used to Assess the Mutagenicity of Compounds

There is great concern about the possible mutagenic properties of any chemical that enters the human body, whether through the skin, the digestive system, or the respiratory tract. Examples of synthetic chemicals that concern us are those found in air and water pollution, food preservatives, artificial sweeteners, herbicides, pesticides, and pharmaceutical products. Mutagenicity can be tested in various organisms, including fungi, plants, and cultured mammalian cells; however, one of the most common tests, which we describe here, uses bacteria.

The **Ames test** uses a number of different strains of the bacterium *Salmonella typhimurium* that have been selected for their ability to reveal the presence of specific types of mutations. For example, some strains are used to detect base-pair substitutions, and other strains detect various frameshift mutations. Each strain contains a mutation in one of the genes of the histidine operon. The mutant strains are unable to synthesize histidine (*his*$^-$ strains) and therefore require histidine for growth (**Figure 14–15**). The assay measures the frequency of reverse mutations that occur within the mutant gene, yielding wild-type bacteria (*his*$^+$ revertants). These *Salmonella* strains also have an increased sensitivity to mutagens due to the presence of mutations in genes involved in both DNA damage repair and the synthesis of the lipopolysaccharide barrier that coats bacteria and protects them from external substances.

Many substances entering the human body are relatively innocuous until activated metabolically, usually in the liver, to more chemically reactive products. Thus, the Ames test includes a step in which the test compound is incubated *in vitro* in the presence of a mammalian liver extract. Alternatively, test compounds may be injected into a mouse where they are modified by liver enzymes and then recovered for use in the Ames test.

In the initial use of Ames testing in the 1970s, a large number of known **carcinogens,** or cancer-causing agents, were examined, and more than 80 percent of these were shown to be strong mutagens. This is not surprising, as the transformation of cells to the malignant state occurs as a result of mutations. For example, more than 60 compounds found in cigarette smoke test positive in the Ames test and cause cancer in animal tests. Although a positive response in the Ames test does not prove that a compound is carcinogenic, the Ames test is useful as a preliminary screening device. The test is used extensively during the development of industrial and pharmaceutical chemical compounds.

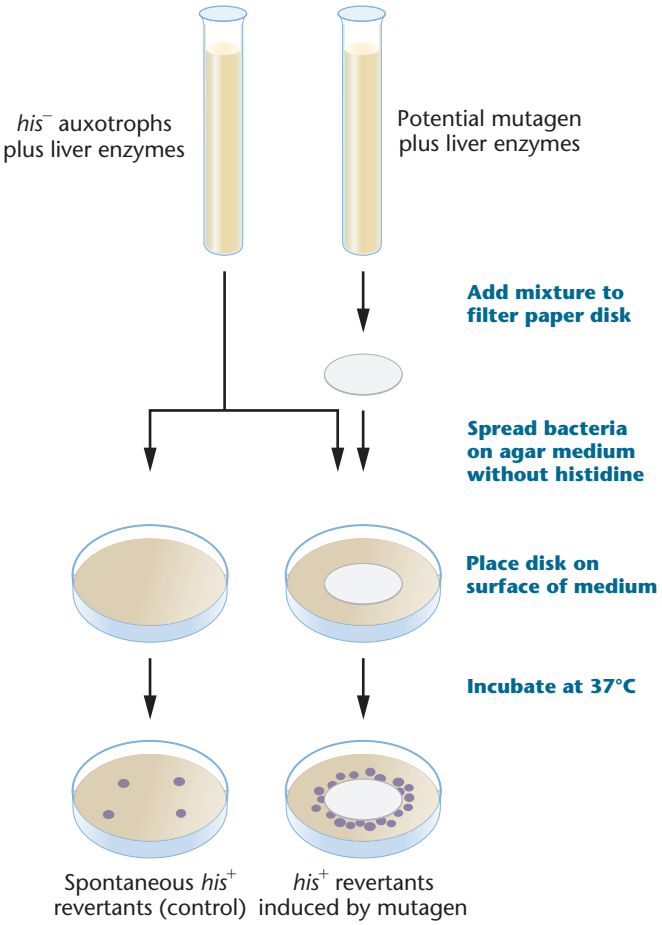


FIGURE 14–15 The Ames test, which screens compounds for potential mutagenicity.

14.7 Transposable Elements Move within the Genome and May Create Mutations

Transposable elements, also known as **transposons** or "jumping genes," can move or transpose within and between chromosomes, inserting themselves into various locations within the genome.

Transposable elements are present in the genomes of all organisms from bacteria to humans. Not only are they ubiquitous, but they also comprise large portions of some eukaryotic genomes. For example, almost 50 percent of the human genome is derived from transposable elements.

Some organisms with unusually large genomes, such as salamanders and barley, contain hundreds of thousands of copies of various types of transposable elements. Although the function of these elements is unknown, data from human genome sequencing suggest that some genes may have evolved from transposable elements and that the presence of these elements may help to modify and reshape the genome. Transposable elements are also valuable tools in genetic research. Geneticists harness transposons as mutagens, as cloning tags, and as vehicles for introducing foreign DNA into model organisms.

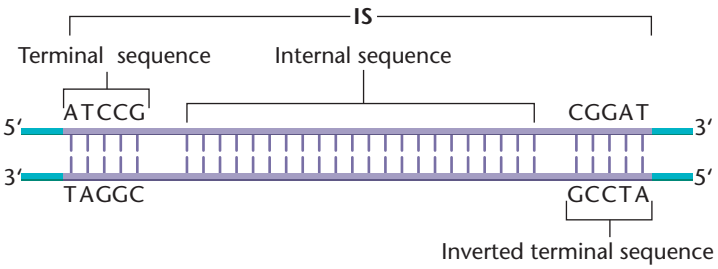


FIGURE 14–16 An insertion sequence (IS), shown in purple. The terminal sequences are perfect inverted repeats of one another.

In this section, we discuss transposable elements as naturally occurring mutagens. The movement of transposable elements from one place in the genome to another has the capacity to disrupt genes and cause mutations, as well as to create chromosomal damage such as double-strand breaks.

Insertion Sequences and Bacterial Transposons

There are two types of transposable elements in bacteria: insertion sequences and bacterial transposons. **Insertion sequences** (**IS elements**) can move from one location to another and, if they insert into a gene or gene-regulatory region, may cause mutations.

IS elements were first identified during analyses of mutations in the *gal* operon of *E. coli.* Researchers discovered that certain mutations in this operon were due to the presence of several hundred base pairs of extra DNA inserted into the beginning of the operon. Surprisingly, the segment of mutagenic DNA could spontaneously excise from this location, restoring wild-type function to the *gal* operon. Subsequent research revealed that several other DNA elements could behave in a similar fashion, inserting into bacterial chromosomes and affecting gene function.

IS elements are relatively short, not exceeding 2000 bp (2 kb). The first insertion sequence to be characterized in *E. coli,* IS1, is about 800 bp long. Other IS elements such as IS2, 3, 4, and 5 are about 1250 to 1400 bp in length. IS elements are present in multiple copies in bacterial genomes. For example, the *E. coli* chromosome contains five to eight copies of IS1, five copies each of IS2 and IS3, as well as copies of IS elements on plasmids such as F factors.

All IS elements contain two features that are essential for their movement. First, they contain a gene that encodes an enzyme called **transposase.** This enzyme is responsible for making staggered cuts in chromosomal DNA, into which the IS element can insert. Second, the ends of IS elements contain **inverted terminal repeats** (**ITRs**). ITRs are short segments of DNA that have the same nucleotide sequence as each other but are oriented in the opposite direction (**Figure 14–16**). Although Figure 14–16 shows the ITRs to consist of only a few nucleotides, IS ITRs usually contain about 20 to 40 nucleotide pairs. ITRs are essential for transposition and act as recognition sites for the binding of the transposase enzyme.

Bacterial transposons (**Tn elements**) are larger than IS elements and contain protein-coding genes that are unrelated to their transposition. Some Tn elements, such as Tn10, are composed of a drug-resistance gene flanked by two IS elements present in opposite orientations. The IS elements encode the transposase enzyme that is necessary for transposition of the Tn element. Other types of Tn elements, such as Tn3, have shorter inverted repeat sequences at their ends and encode their transposase enzyme from a transposase gene located in the middle of the Tn element. Like IS elements, Tn elements are mobile in both bacterial chromosomes and in plasmids and can cause mutations if they insert into genes or gene-regulatory regions.

Tn elements are currently of interest because they can introduce multiple drug resistance onto bacterial plasmids. These plasmids, called **R factors,** may contain many Tn elements conferring simultaneous resistance to heavy metals, antibiotics, and other drugs. These elements can move from plasmids onto bacterial chromosomes and can spread multiple drug resistance between different strains of bacteria.

The *Ac–Ds* System in Maize

About 20 years before the discovery of transposons in bacteria, Barbara McClintock discovered mobile genetic elements in corn plants (maize). She did this by analyzing the genetic behavior of two mutations, ***Dissociation*** (***Ds***) and ***Activator*** (***Ac***), expressed in either the endosperm or aleurone layers. She then correlated her genetic observations with cytological examinations of the maize

chromosomes. Initially, McClintock determined that *Ds* was located on chromosome 9. If *Ac* was also present in the genome, *Ds* induced breakage at a point on the chromosome adjacent to its own location. If chromosome breakage occurred in somatic cells during their development, progeny cells often lost part of the broken chromosome, causing a variety of phenotypic effects. The chapter-opening photo illustrates the types of phenotypic effects caused by *Ds* mutations in kernels of corn.

Subsequent analysis suggested to McClintock that both *Ds* and *Ac* elements sometimes moved to new chromosomal locations. While *Ds* moved only if *Ac* was also present, *Ac* was capable of autonomous movement. Where *Ds* came to reside determined its genetic effects—that is, it might cause chromosome breakage, or it might inhibit expression of a certain gene. In cells in which *Ds* caused a gene mutation, *Ds* might move again, restoring the gene mutation to wild type.

Figure 14–17 illustrates the types of movements and effects brought about by *Ds* and *Ac* elements. In McClintock's original observation, pigment synthesis was restored in cells in which the *Ds* element jumped out of chromosome 9. McClintock concluded that the *Ds* and *Ac* genes were **mobile controlling elements.** We now commonly refer to them as transposable elements, a term coined by another great maize geneticist, Alexander Brink.

Several *Ac* and *Ds* elements have now been analyzed, and the relationship between the two elements has been clarified. The first *Ds* element studied (*Ds*9) is nearly identical to *Ac* except for a 194-bp deletion within the transposase gene. The deletion of part of the transposase gene in the *Ds*9 element explains its dependence on the *Ac* element for transposition. Several other *Ds* elements have also been sequenced, and each contains an even larger deletion within the transposase gene. In each case, however, the ITRs are retained.

Although the significance of Barbara McClintock's mobile controlling elements was not fully appreciated following her initial observations, molecular analysis has since verified her conclusions. She was awarded the Nobel Prize in Physiology or Medicine in 1983.

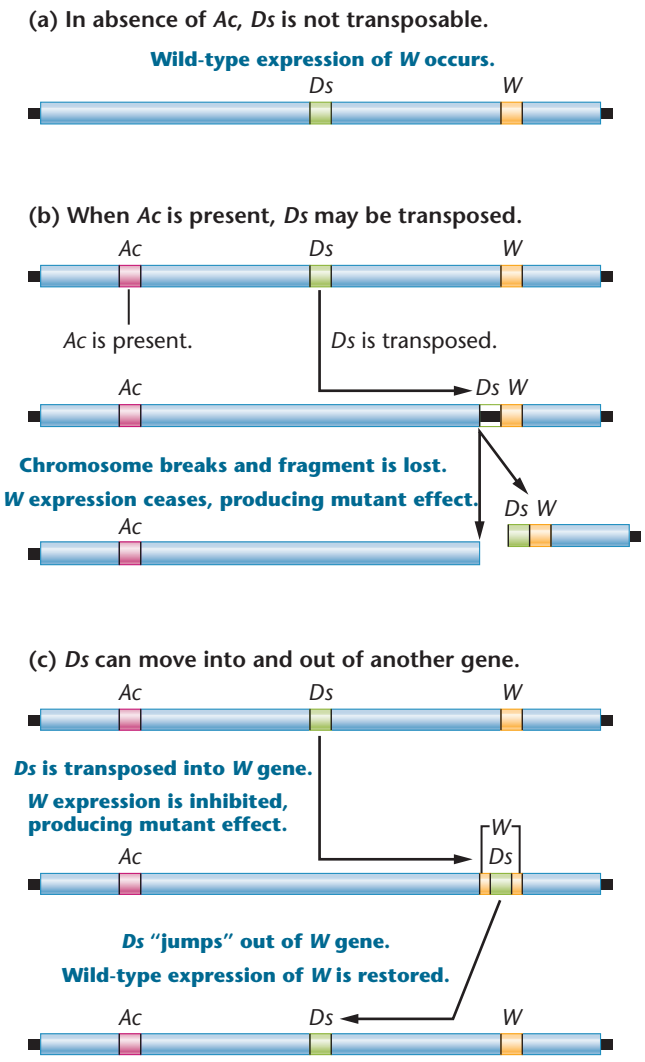


FIGURE 14–17 Effects of *Ac* and *Ds* elements on gene expression. (a) If *Ds* is present in the absence of *Ac*, there is normal expression of a distantly located hypothetical gene *W*. (b) In the presence of *Ac*, *Ds* may transpose to a region adjacent to *W*. *Ds* can induce chromosome breakage, which may lead to loss of a chromosome fragment bearing the *W* gene. (c) In the presence of *Ac*, *Ds* may transpose into the *W* gene, disrupting *W*-gene expression. If *Ds* subsequently transposes out of the *W* gene, *W*-gene expression may return to normal.

Copia and *P* Elements in *Drosophila*

There are more than 30 families of transposable elements in *Drosophila,* each of which is present in 20 to 50 copies in the genome. Together, these families constitute about 5 percent of the *Drosophila* genome and over half of the middle repetitive DNA of this organism. One study suggests that 50 percent of all visible mutations in *Drosophila* are the result of the insertion of transposons into otherwise wild-type genes.

In 1975, David Hogness and his colleagues David Finnegan, Gerald Rubin, and Michael Young identified a class of DNA elements in *Drosophila melanogaster* that they designated as ***copia.*** These elements are transcribed into "copious" amounts of RNA (hence their name). *Copia* elements are present in 10 to 100 copies in the genomes of *Drosophila* cells. Mapping studies show that they are transposable to different chromosomal locations and are dispersed throughout the genome.

Each *copia* element consists of approximately 5000 to 8000 bp of DNA, including a long **direct terminal repeat**

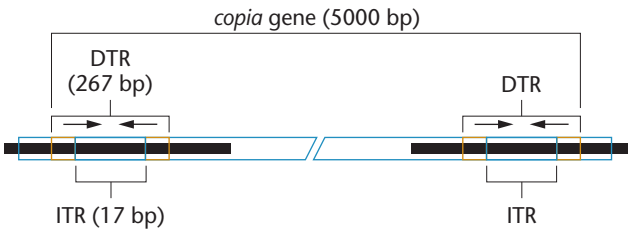


FIGURE 14–18 Structural organization of a *copia* transposable element in *Drosophila melanogaster,* showing the terminal repeats.

(**DTR**) sequence of 267 bp at each end. Within each DTR is an inverted terminal repeat (ITR) of 17 bp (**Figure 14–18**).

Insertion of *copia* is dependent on the presence of the ITR sequences and seems to occur preferentially at specific target sites in the genome. The *copia*-like elements demonstrate regulatory effects at the point of their insertion in the chromosome. Certain mutations, including those affecting eye color and segment formation, are due to *copia* insertions within genes. For example, the eye-color mutation *white-apricot* is caused by an allele of the *white* gene, which contains a *copia* element within the gene. Transposition of the *copia* element out of the *white-apricot* allele can restore the allele to wild type.

Perhaps the most significant *Drosophila* transposable elements are the ***P* elements.** These were discovered while studying the phenomenon of **hybrid dysgenesis,** a condition characterized by sterility, elevated mutation rates, and chromosome rearrangements in the offspring of crosses between certain strains of fruit flies. Hybrid dysgenesis is caused by high rates of *P* element transposition in the germ line, in which transposons insert themselves into or near genes, thereby causing mutations. *P* elements range from 0.5 to 2.9 kb long, with 31-bp ITRs. Full-length *P* elements encode at least two proteins, one of which is the transposase enzyme that is required for transposition, and another is a repressor protein that inhibits transposition. The transposase gene is expressed only in the germ line, accounting for the tissue specificity of *P* element transposition. Strains of flies that contain full-length *P* elements inserted into their genomes are resistant to further transpositions due to the presence of the repressor protein encoded by the *P* elements.

Mutations can arise from several kinds of insertional events. If a *P* element inserts into the coding region of a gene, it can terminate transcription of the gene and destroy normal gene expression. If it inserts into the promoter region of a gene, it can affect the level of expression of the gene. Insertions into introns can affect splicing or cause the premature termination of transcription.

Geneticists have harnessed *P* elements as tools for genetic analysis. One of the most useful applications of *P* elements is as vectors to introduce transgenes into *Drosophila*—a technique known as **germ-line transformation.** *P* elements are also used to generate mutations and to clone mutant genes. In addition, researchers are perfecting methods to target *P* element insertions to precise single-chromosomal sites, which should increase the precision of germ-line transformation in the analysis of gene activity.

Transposable Elements in Humans

The human genome, like that of other eukaryotes, is riddled with DNA derived from transposons. Recent genomic sequencing data reveal that approximately half of the human genome is composed of transposable element DNA. As discussed earlier in the text (see Chapter 11), the major families of human transposable elements are the long interspersed elements and short interspersed elements (**LINEs** and **SINEs,** respectively). Together, they comprise over 30 percent of the human genome.

Although most human transposable elements appear to be inactive, the potential mobility and mutagenic effects of these elements have far-reaching implications for human genetics, as can be seen in a recent example of a transposable element "caught in the act." The case involves a male child with hemophilia. One cause of hemophilia is a defect in blood-clotting factor VIII, the product of an X-linked gene. Haig Kazazian and his colleagues found LINEs inserted at two points within the gene. Researchers were interested in determining if one of the mother's X chromosomes also contained this specific LINE. If so, the unaffected mother would be heterozygous and pass the LINE-containing chromosome to her son. The surprising finding was that the LINE sequence was *not* present on either of her X chromosomes but *was* detected on chromosome 22 of both parents. This suggests that this mobile element may have transposed from one chromosome to another in the gamete-forming cells of the mother, prior to being transmitted to the son.

LINE insertions into the human *dystrophin* gene have resulted in at least two separate cases of Duchenne muscular dystrophy. In one case, a LINE inserted into exon 48, and in another case, it inserted into exon 44, both leading to frameshift mutations and premature termination of translation of the dystrophin protein. There are also reports that LINEs have inserted into the *APC* and *c-myc* genes, leading to mutations that may have contributed to the development of some colon and breast cancers. In the latter cases, the transposition had occurred within one or a few somatic cells. As of 2012, researchers have determined that at least 25 LINE element insertions have resulted in single-gene human diseases.

SINE insertions are also responsible for more than 30 cases of human disease. In one case, an ***Alu* element**

integrated into the *BRCA2* gene, inactivating this tumor suppressor gene and leading to a familial case of breast cancer. Other genes that have been mutated by *Alu* integrations are the *factor IX* gene (leading to hemophilia B), the *ChE* gene (leading to acholinesterasemia), and the *NF1* gene (leading to neurofibromatosis).

Transposons, Mutations, and Evolution

Transposons can have a wide range of effects on genes. The insertion of a transposon into the coding region of a gene may disrupt the gene's normal translation reading frame or may induce premature termination of translation of the mRNA transcribed from the gene. Many transposable elements contain their own promoters and enhancers, as well as splice sites and polyadenylation signals. The presence of these regulatory sequences can have effects on nearby genes. The insertion of a transposable element containing polyadenylation or transcription termination signals into a gene's intron may bring about termination of the gene's transcription within the element. In addition, it can cause aberrant splicing of an RNA transcribed from the gene. Insertions of a transposon into a gene's transcription regulatory region may disrupt the gene's normal regulation or may cause the gene to be expressed differently as a result of the presence of the transposon's own promoter or enhancer sequences. The presence of two or more identical transposons in a genome creates the potential for recombination between the transposons, leading to duplications, deletions, inversions, or chromosome translocations. Any of these rearrangements may bring about phenotypic changes or disease.

It is thought that about 0.2 percent of detectable human mutations may be due to transposable element insertions. Other organisms appear to suffer more damage due to transposition. About 10 percent of new mouse mutations and 50 percent of *Drosophila* mutations are caused by insertions of transposable elements in or near genes.

Because of their ability to alter genes and chromosomes, transposons may contribute to the variability that underlies evolution. For example, the Tn elements of bacteria carry antibiotic resistance genes between organisms, conferring a survival advantage to the bacteria under certain conditions. Another example of a transposon's contribution to evolution is provided by *Drosophila* telomeres. LINE-like elements are present at the ends of *Drosophila* chromosomes, and these elements act as telomeres, maintaining the length of *Drosophila* chromosomes over successive cell divisions. Other examples of evolved transposons are the *RAG1* and *RAG2* genes in humans. These genes encode **recombinase** enzymes that are essential to the development of the immune system. These two genes appear to have evolved from transposons.

Transposons may also affect the evolution of genomes by altering gene-expression patterns in ways that are subsequently retained by the host. For example, the human *amylase* gene contains an enhancer that causes the gene to be expressed in the parotid gland. This enhancer evolved from transposon sequences that were inserted into the generegulatory region early in primate evolution. Other examples of gene-expression patterns that were affected by the presence of transposon sequences are T-cell-specific expression of the *CD8* gene and placenta-specific expression of the *leptin* and *CYP19* genes.

ESSENTIAL POINT

Transposable elements can move within a genome, creating mutations and altering gene expression. Besides creating mutations, transposons may contribute to evolution. Geneticists use transposons as research tools to create mutations, clone genes, and introduce genes into model organisms. ■

CASE STUDY | Genetic dwarfism

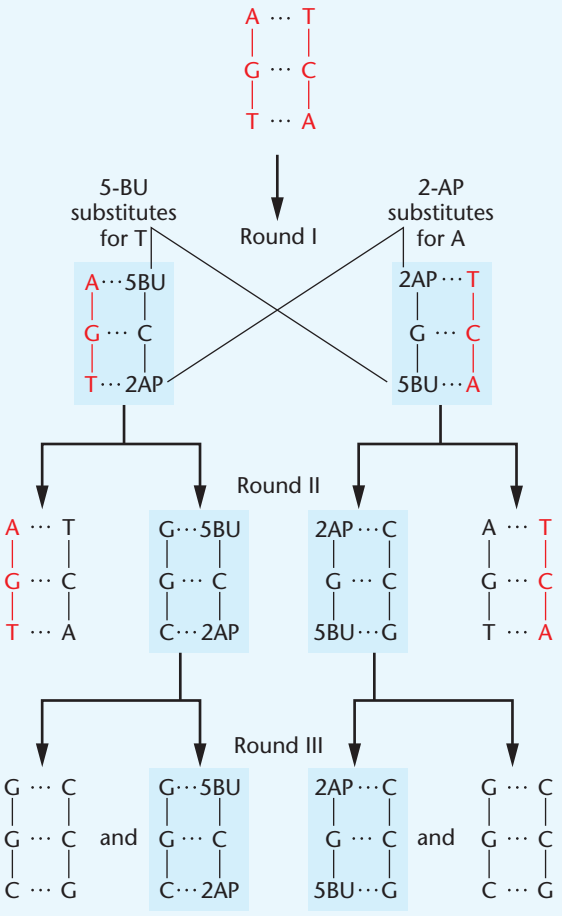
Seven months pregnant, an expectant mother was undergoing a routine ultrasound. While prior tests had been normal, this one showed that the limbs of the fetus were unusually short. The doctor suspected that the baby might have a genetic form of dwarfism called achondroplasia. He told her that the disorder was due to an autosomal dominant mutation and occurred with a frequency of about 1 in 25,000 births. The expectant mother had studied genetics in college and immediately raised several questions. How would you answer them?

1. How could her baby have a dominantly inherited disorder if there was no history of this condition on either side of the family?
2. Is the mutation more likely to have come from the mother or the father?
3. If this child has achondroplasia, is there an increased chance that their next child would also have this disorder?
4. Could this disorder have been caused by X rays or ultrasounds she had earlier in pregnancy?

INSIGHTS AND SOLUTIONS

1. The base analog 2-amino purine (2-AP) substitutes for adenine during DNA replication, but it may base-pair with cytosine. The base analog 5-bromouracil (5-BU) substitutes for thymidine, but it may base-pair with guanine. Follow the double-stranded trinucleotide sequence shown here through three rounds of replication, assuming that, in the first round, both analogs are present and become incorporated wherever possible. Before the second and third round of replication, any unincorporated base analogs are removed. What final sequences occur?

Solution:



2. A rare dominant mutation expressed at birth was studied in humans. Records showed that six cases were discovered in 40,000 live births. Family histories revealed that in two cases, the mutation was already present in one of the parents. Calculate the spontaneous mutation rate for this mutation. What are some underlying assumptions that may affect our conclusions?

Solution: Only four cases represent a new mutation. Because each live birth represents two gametes, the sample size is from 80,000 meiotic events. The rate is equal to

$$4/80{,}000 = 1/20{,}000 = 5 \times 10^{-5}$$

We have assumed that the mutant gene is fully penetrant and is expressed in each individual bearing it. If it is not fully penetrant, our calculation may be an underestimate because one or more mutations may have gone undetected. We have also assumed that the screening was 100 percent accurate. One or more mutant individuals may have been "missed," again leading to an underestimate. Finally, we assumed that the viability of the mutant and nonmutant individuals is equivalent and that they survive equally *in utero*. Therefore, our assumption is that the number of mutant individuals at birth is equal to the number at conception. If this were not true, our calculation would again be an underestimate.

3. Consider the following estimates:

   a. There are 7×10^9 humans living on this planet.

   b. Each individual has about 20,000 (0.2×10^5) genes.

   c. The average mutation rate at each locus is 10^{-5}.

   How many spontaneous mutations are currently present in the human population? Assuming that these mutations are equally distributed among all genes, how many new mutations have arisen in each gene in the human population?

Solution: First, since each individual is diploid, there are two copies of each gene per person, each arising from a separate gamete. Therefore, the total number of spontaneous mutations is

$$(2 \times 0.2 \times 10^5 \text{ genes}) \times (7 \times 10^9 \text{ humans}) \times (10^{-5} \text{ mutations})$$
$$= (0.4 \times 10^5) \times (7 \times 10^9) \times (10^{-5}) \text{ mutations}$$
$$= 2.8 \times 10^9 \text{ mutations in the population}$$
$$2.8 \times 10^9 \text{ mutations}/0.2 \times 10^5 \text{ genes}$$
$$= 14 \times 10^4 \text{ mutations per gene in the population}$$

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we focused on how gene mutations arise and how cells repair DNA damage. In particular, we discussed spontaneous and induced mutations, DNA repair methods, and transposable elements. Based on your knowledge of these topics, answer several fundamental questions:
   (a) How do we know that mutations occur spontaneously?
   (b) How do we know that certain chemicals and wavelengths of radiation induce mutations in DNA?
   (c) How do we know that DNA repair mechanisms detect and correct the majority of spontaneous and induced mutations?

CONCEPT QUESTION

2. Review the Chapter Concepts list on page 273. These concepts relate to how gene mutations occur, their phenotypic effects, and how mutations can be repaired. The first four concepts focus on the effects of gene mutations in diploid organisms. Write a short essay describing how these concepts would apply, or not apply, to a haploid organism such as *E. coli*. ■

3. Distinguish between spontaneous and induced mutations. Give some examples of mutagens that cause induced mutations.
4. Why would a mutation in a somatic cell of a multicellular organism escape detection?
5. Why is a random mutation more likely to be deleterious than beneficial?
6. Why are organisms that have a haploid life cycle valuable tools for mutagenesis studies?
7. What is meant by a conditional mutation?
8. Describe a tautomeric shift and how it may lead to a mutation.
9. Contrast and compare the mutagenic effects of deaminating agents, alkylating agents, and base analogs.
10. Why are frameshift mutations likely to be more detrimental than point mutations, in which a single pyrimidine or purine has been substituted?
11. In which phases of the cell cycle would you expect double-strand break repair and nonhomologous end joining to occur and why?
12. DNA damage brought on by a variety of natural and artificial agents elicits a wide variety of cellular responses. In addition to the activation of DNA repair mechanisms, there can be activation of pathways leading to apoptosis (programmed cell death) and cell-cycle arrest. Why would apoptosis and cell-cycle arrest often be part of a cellular response to DNA damage?
13. Distinguish between proofreading and mismatch repair.
14. How would you expect the misincorporation of bases by a DNA polymerase to change if the relative ratios of the dNTPs were A = T = G but a five-fold excess of C?
15. A chemist has synthesized a novel chemical, which he suspects to be a potential mutagen. Name and explain a popular test that can be used to test the mutagenicity of this product in bacteria.
16. What genetic defects result in the disorder xeroderma pigmentosum (XP) in humans? How do these defects create the phenotypes associated with the disorder?
17. In a bacterial culture in which all cells are unable to synthesize leucine (*leu*$^-$), a potent mutagen is added, and the cells are allowed to undergo one round of replication. At that point, samples are taken, a series of dilutions is made, and the cells are plated on either minimal medium or minimal medium containing leucine. The first culture condition (minimal medium) allows the growth of only *leu*$^+$ cells, while the second culture condition (minimum medium with leucine added) allows the growth of all cells. The results of the experiment are as follows:

Culture Condition	Dilution	Colonies
Minimal medium	10^{-1}	18
Minimal + leucine	10^{-7}	6

What is the rate of mutation at the locus associated with leucine biosynthesis?

18. DNA mismatch repair is a mechanism of DNA repair that has been observed in *E. coli*. Give a list of genes in *E. coli*, mutations in which can adversely affect DNA mismatch repair. Give a list of equivalent genes in humans.
19. A number of different types of mutations in the *HBB* gene can cause human β-thalassemia, a disease characterized by various levels of anemia. Many of these mutations occur within introns or in upstream noncoding sequences. Explain why mutations in these regions often lead to severe disease, although they may not directly alter the coding regions of the gene.
20. Some mutations that lead to diseases such as Huntington disease are caused by the insertion of trinucleotide repeats. Describe how the process of DNA replication could lead to expansions of trinucleotide repeat regions.
21. In maize, a *Ds* or *Ac* transposon can cause mutations in genes at or near the site of transposon insertion. It is possible for these elements to transpose away from their original site, causing a reversion of the mutant phenotype. In some cases, however, even more severe phenotypes appear, due to events at or near the mutant allele. What might be happening to the transposon or the nearby gene to create more severe mutations?
22. Presented here are hypothetical findings from studies of heterokaryons formed from seven human xeroderma pigmentosum cell strains:

	XP1	*XP2*	*XP3*	*XP4*	*XP5*	*XP6*	*XP7*
XP1	−						
XP2	−	−					
XP3	−	−	−				
XP4	+	+	+	−			
XP5	+	+	+	+	−		
XP6	+	+	+	+	−	−	
XP7	+	+	+	+	−	−	−

Note: "+" = complementation; "−" = no complementation

These data are measurements of the occurrence or nonoccurrence of unscheduled DNA synthesis in the fused heterokaryon. None of the strains alone shows any unscheduled DNA synthesis. What does unscheduled DNA synthesis represent? Which strains fall into the same complementation groups? How many different groups are revealed based on these data? What can we conclude about the genetic basis of XP from these data?

23. Cystic fibrosis (CF) is a severe autosomal recessive disorder in humans that results from a chloride ion channel defect in epithelial cells. More than 500 mutations have been identified in the 24 exons of the responsible gene (*CFTR,* or cystic fibrosis transmembrane regulator), including dozens of different missense mutations, frameshift mutations, and splice-site defects. Although all affected CF individuals demonstrate chronic obstructive lung disease, there is variation in whether or not they exhibit pancreatic enzyme insufficiency (PI). Speculate as to which types of mutations are likely to give rise to less severe symptoms of CF, including only minor PI. Some of the 300 sequence alterations that have been detected within the exon regions of the *CFTR* gene do not give rise to cystic fibrosis. Taking into account your knowledge of the genetic code, gene expression, protein function, and mutation, describe why this might be so.
24. Electrophilic oxidants are known to create the modified base named 7,8-dihydro-8-oxoguanine (oxoG) in DNA. Whereas guanine base-pairs with cytosine, oxoG base-pairs with either cytosine or adenine.
    (a) What are the sources of reactive oxidants within cells that cause this type of base alteration?
    (b) Drawing on your knowledge of nucleotide chemistry, draw the structure of oxoG, and, below it, draw guanine. Opposite guanine, draw cytosine, including the hydrogen bonds that allow these two molecules to base-pair. Does the structure of oxoG, in contrast to guanine, provide any hint as to why it base-pairs with adenine?

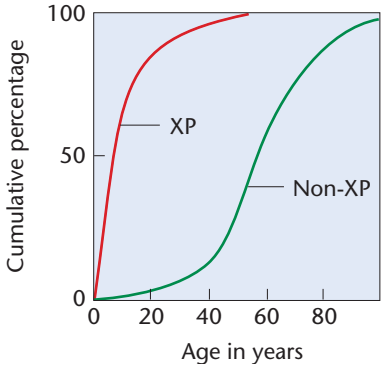
(c) Assume that an unrepaired oxoG lesion is present in the helix of DNA opposite cytosine. Predict the type of mutation that will occur following several rounds of replication.

(d) Which DNA repair mechanisms might work to counteract an oxoG lesion? Which of these is likely to be most effective?

25. Skin cancer carries a lifetime risk nearly equal to that of all other cancers combined. Following is a graph (modified from Kraemer, 1997. *Proc. Natl. Acad. Sci. (USA)* 94: 11–14) depicting the age of onset of skin cancers in patients with or without XP, where cumulative percentage of skin cancer is plotted against age. The non-XP curve is based on 29,757 cancers surveyed by the National Cancer Institute, and the curve representing those with XP is based on 63 skin cancers from the Xeroderma Pigmentosum Registry.

(a) Provide an overview of the information contained in the graph.

(b) Explain why individuals with XP show such an early age of onset.

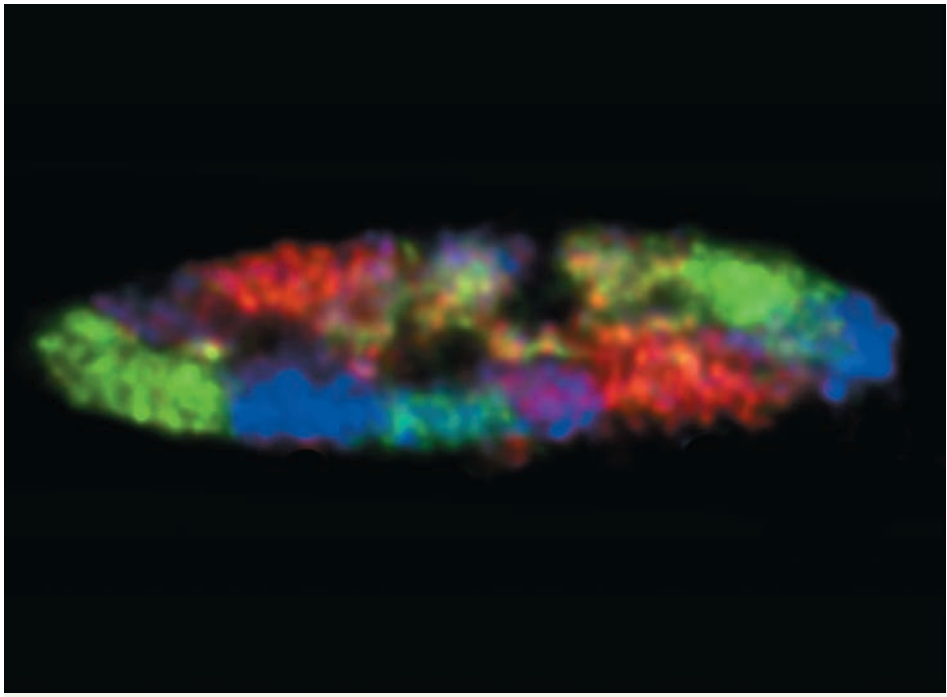


26. The initial discovery of IS elements in bacteria revealed the presence of an element upstream (5′) of three genes controlling galactose metabolism. All three genes were affected simultaneously, although there was only one IS insertion. Offer an explanation as to why this might occur.
27. Suppose you are studying a DNA repair system, such as the nucleotide excision repair *in vitro*. By mistake, you add DNA ligase from a tube that has already expired. What would be the result?
28. It has been noted that most transposons in humans and other organisms are located in noncoding regions of the genome—regions such as introns, pseudogenes, and stretches of particular types of repetitive DNA. There are several ways to interpret this observation. Describe two possible interpretations. Which interpretation do you favor? Why?
29. Two related forms of muscular dystrophy—Duchenne muscular dystrophy (DMD) and Becker muscular dystrophy (BMD)—are both recessive, X-linked, single-gene conditions caused by point mutations, deletions, and insertion in the *dystrophin* gene. Each mutated form of *dystrophin* is one allele. Of the two diseases, DMD is much more severe. Given your knowledge of mutations, the genetic code, and translation, propose an explanation for why the two disorders differ greatly in severity.

15 Regulation of Gene Expression

CHAPTER CONCEPTS

- Expression of genetic information is regulated by intricate regulatory mechanisms that control transcription, mRNA stability, translation, and posttranslational modifications.
- In prokaryotes, genes that encode proteins with related functions tend to be organized in clusters and are often under coordinated control. Such clusters, including their associated regulatory sequences, are called operons.
- Transcription within operons is either inducible or repressible and is often regulated by the metabolic substrate or end product of the pathway.
- Eukaryotic gene regulation is more complex than prokaryotic gene regulation.
- The organization of eukaryotic chromatin in the nucleus plays a role in regulating gene expression. Chromatin must be remodeled to provide access to regulatory DNA sequences within it.
- Eukaryotic transcription initiation requires the presence of transcription regulators at enhancer sites and general transcription complexes at promoter sites.
- Eukaryotic gene expression is also regulated at posttranscriptional steps, including splicing of pre-mRNA, mRNA stability, translation, and posttranslational processing.



Chromosome territories in an interphase chicken cell nucleus. Each chromosome is stained with a different-colored probe.

In previous chapters, we described how DNA is organized into genes, how genes store genetic information, and how this information is expressed through the processes of transcription and translation. We now consider one of the most fundamental questions in molecular genetics: *How is gene expression regulated?*

It is clear that not all genes are expressed at all times in all situations. For example, some proteins in the bacterium *E. coli* are present in as few as 5 to 10 molecules per cell, whereas others, such as ribosomal proteins and the many proteins involved in the glycolytic pathway, are present in as many as 100,000 copies per cell. Although many prokaryotic gene products are present continuously at low levels, the level of these products can increase dramatically when required. In multicellular eukaryotes, differential gene expression is also essential, not only to allow appropriate and rapid responses to their environments, but also as the basis for embryonic development and adult organ function.

The activation and repression of gene expression are part of a delicate balancing act for both prokaryotic and eukaryotic organisms. Expression of a gene at the wrong time, in the wrong cell type, or in abnormal amounts can lead to a deleterious phenotype, cancer, or cell death—even when the gene itself is normal.

In this chapter, we will explore the ways in which prokaryotic and eukaryotic organisms regulate gene expression. We will describe some of the fundamental components of gene regulation, including the *cis*-acting DNA elements and *trans*-acting factors that regulate transcription

initiation. We will then explain how these components interact with each other and with other factors such as activators, repressors, and chromatin proteins. We will also consider the roles that posttranscriptional mechanisms play in the regulation of eukaryotic gene expression. Please note that some of the topics discussed in this chapter are explored in greater depth later in the text (see Special Topic Chapter 1—Epigenetics and Special Topic Chapter 2—Emerging Roles of RNA.)

15.1 Prokaryotes Regulate Gene Expression in Response to Both External and Internal Conditions

Not only do bacteria respond metabolically to changes in their environment, but they also regulate gene expression in order to synthesize products required for a variety of normal cellular activities, including DNA replication, recombination, repair, and cell division. In the following sections, we will focus on prokaryotic gene regulation at the level of transcription, which is the predominant level of regulation in prokaryotes. Keep in mind, however, that posttranscriptional regulation also occurs in bacteria. We will defer discussion of posttranscriptional gene-regulatory mechanisms to subsequent sections dealing with eukaryotic gene expression.

The idea that microorganisms regulate the synthesis of gene products is not a new one. As early as 1900, it was shown that when lactose (a galactose and glucose-containing disaccharide) is present in the growth medium of yeast, the organisms synthesize enzymes required for lactose metabolism. When lactose is absent, the enzymes are not manufactured. Soon thereafter, investigators were able to generalize that bacteria also adapt to their environment, producing certain enzymes only when specific chemical substrates are present. These enzymes are referred to as **inducible enzymes,** reflecting the role of the substrate, which serves as the **inducer** in enzyme production. In contrast, those enzymes that are produced continuously, regardless of the chemical makeup of the environment, are called **constitutive enzymes.**

More recent investigation has revealed a contrasting system whereby the presence of a specific molecule inhibits gene expression. This is usually true for molecules that are end products of anabolic biosynthetic pathways. For example, the amino acid tryptophan can be synthesized by bacterial cells. If a sufficient supply of tryptophan is present in the environment or culture medium, it is energetically inefficient for the organism to synthesize the enzymes necessary for tryptophan production. A mechanism has evolved whereby tryptophan plays a role in repressing transcription of genes that encode the appropriate biosynthetic enzymes. In contrast to the inducible system controlling lactose metabolism, the system governing tryptophan expression is said to be **repressible.**

Regulation, whether it is inducible or repressible, may be under either **negative** or **positive control.** Under negative control, gene expression occurs *unless it is shut off by some form of a regulator molecule.* In contrast, under positive control, transcription occurs *only if a regulator molecule directly stimulates RNA production.* In theory, either type of control or a combination of the two can govern inducible or repressible systems.

15.2 Lactose Metabolism in *E. coli* Is Regulated by an Inducible System

Beginning in 1946, the studies of Jacques Monod (with later contributions by Joshua Lederberg, François Jacob, and André Lwoff) revealed genetic and biochemical insights into the mechanisms of lactose metabolism in bacteria. These studies explained how gene expression is repressed when lactose is absent, but induced when it is available. In the presence of lactose, concentrations of the enzymes responsible for lactose metabolism increase rapidly from a few molecules to thousands per cell. The enzymes responsible for lactose metabolism are thus *inducible,* and lactose serves as the *inducer.*

In prokaryotes, genes that code for enzymes with related functions (in this case, the genes involved with lactose metabolism) tend to be organized in clusters on the bacterial chromosome. In addition, transcription of these genes is often under the coordinated control of a single transcription regulatory region. The location of this regulatory region is almost always on the same DNA molecule and upstream of the gene cluster it controls. We refer to this type of regulatory region as a ***cis*-acting site.** *Cis*-acting regulatory regions bind molecules that control transcription of the gene cluster. Such molecules are called ***trans*-acting molecules.** Actions at the *cis*-acting regulatory site determine whether the genes are transcribed into RNA and thus whether the corresponding enzymes or other protein products are synthesized from the mRNA. Binding of a *trans*-acting molecule at a *cis*-acting site can regulate the gene cluster either negatively (by turning off transcription) or positively (by turning on transcription of genes in the cluster). In this section, we discuss how transcription of such bacterial gene clusters is coordinately regulated.

ESSENTIAL POINT

Research on the *lac* operon in *E. coli* pioneered our understanding of gene regulation in bacteria. ■

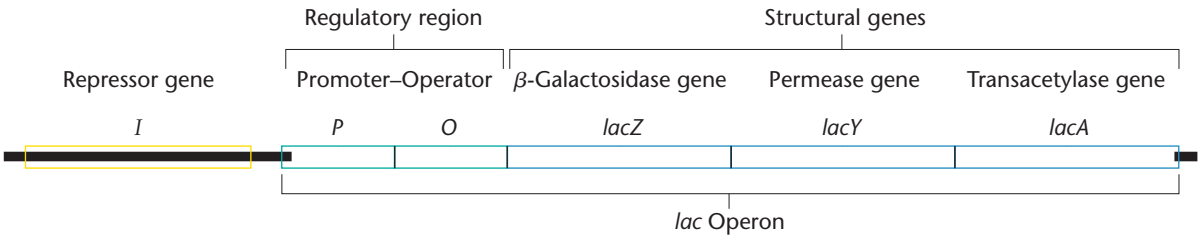


FIGURE 15–1 A simplified overview of the genes and regulatory units involved in the control of lactose metabolism. (This region of DNA is not drawn to scale.) A more detailed model is described later in this chapter.

Structural Genes

As illustrated in **Figure 15–1**, three genes and an adjacent regulatory region constitute the **lactose,** or ***lac,* operon.** Together, the entire gene cluster functions in an integrated fashion to provide a rapid response to the presence or absence of lactose.

Genes coding for the primary structure of the enzymes are called **structural genes.** There are three structural genes in the *lac* operon. The *lacZ* gene encodes ***β*-galactosidase,** an enzyme whose role is to convert the disaccharide lactose to the monosaccharides glucose and galactose (**Figure 15–2**). This conversion is essential if lactose is to serve as an energy source in glycolysis. The second gene, *lacY,* encodes the amino acid sequence of **permease,** an enzyme that facilitates the entry of lactose into the bacterial cell. The third gene, *lacA,* codes for the enzyme **transacetylase.** Although its physiological role is still not completely clear, it may be involved in the removal of toxic by-products of lactose digestion from the cell.

To study the genes encoding these three enzymes, researchers isolated numerous mutants, each of which eliminated the function of one of the enzymes. These mutants were first isolated and studied by Joshua Lederberg. Mutant cells that fail to produce active β-galactosidase (*lacZ*$^-$) or permease (*lacY*$^-$) are unable to use lactose as an energy source and are collectively known as *lac*$^-$ mutants. Mapping studies by Lederberg established that all three genes are closely linked or contiguous to one another on the bacterial chromosome, in the order *Z–Y–A*. (See Figure 15–1.) All three genes are transcribed as a single unit, resulting in a polycistronic mRNA **(Figure 15–3)**. This results in the coordinated regulation of all three genes, since a single messenger RNA is simultaneously translated into all three gene products.

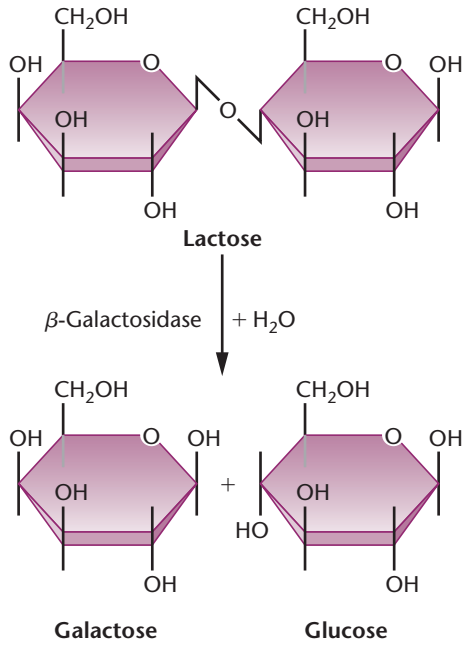


FIGURE 15–2 The catabolic conversion of the disaccharide lactose into its monosaccharide units, galactose and glucose.

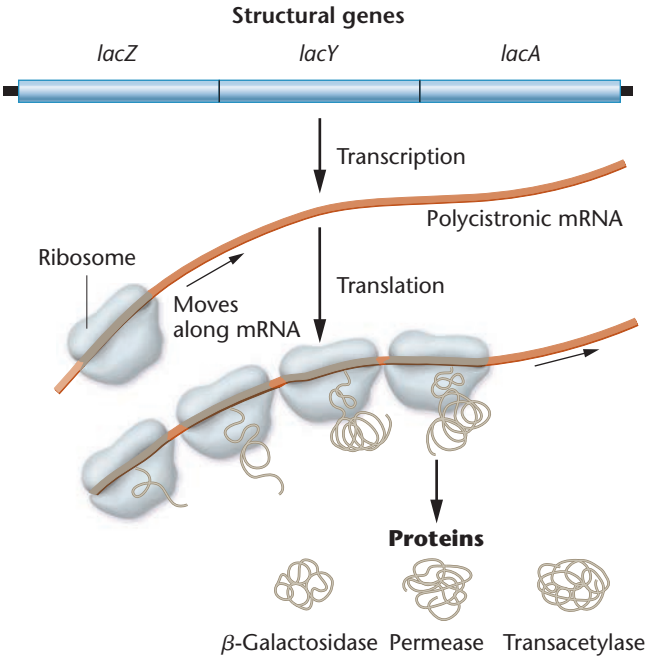


FIGURE 15–3 The structural genes of the *lac* operon are transcribed into a single polycistronic mRNA, which is translated simultaneously by several ribosomes into the three enzymes encoded by the operon.

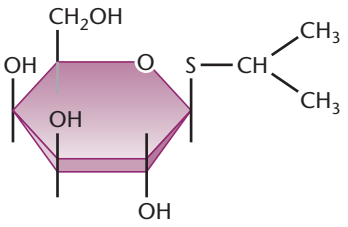


FIGURE 15–4 The gratuitous inducer isopropylthiogalactoside (IPTG).

The Discovery of Regulatory Mutations

How does lactose stimulate transcription of the *lac* operon and induce the synthesis of the related enzymes? A partial answer comes from studies using **gratuitous inducers,** chemical analogs of lactose such as the sulfur analog **isopropylthiogalactoside** (**IPTG**), shown in **Figure 15–4**. Gratuitous inducers behave like natural inducers, but they do not serve as substrates for the enzymes that are subsequently synthesized.

What, then, is the role of lactose in gene regulation? The answer to this question required the study of a class of mutants called **constitutive mutants.** In cells bearing constitutive mutations, enzymes are produced regardless of the presence or absence of lactose. Studies of the constitutive mutation *lacI*$^-$ mapped the mutation to a site on the bacterial chromosome close to, but distinct from, the *lacZ, lacY,* and *lacA* genes. This mutation defined the *lacI* gene, which is appropriately called a **repressor gene.** Another set of constitutive mutations that produce identical effects to those of *lacI*$^-$ occurs in a region immediately adjacent to the structural genes. This class of mutations, designated *lacO*^{C}, occurs in the **operator region** of the operon. In both types of constitutive mutants, the enzymes are produced continuously, inducibility is eliminated, and gene regulation is lost.

The Operon Model: Negative Control

Around 1960, Jacob and Monod proposed a scheme involving negative control called the **operon model,** whereby a group of genes is regulated and expressed together as a unit. As we saw in Figure 15–1, the *lac* operon consists of the *Z, Y,* and *A* structural genes, as well as the adjacent sequences of DNA referred to as the *operator region.* They argued that the *lacI* gene regulates the transcription of the structural genes by producing a **repressor molecule,** and that the repressor is **allosteric,** meaning that it reversibly interacts with another molecule, causing both a conformational change in the repressor's three-dimensional shape and a change in its chemical activity. **Figure 15–5** illustrates the components of the *lac* operon as well as the action of the *lac* repressor in the presence and absence of lactose.

Jacob and Monod suggested that the repressor normally binds to the DNA sequence of the operator region. When it does so, it inhibits the action of RNA polymerase, effectively repressing the transcription of the structural genes [**Figure 15–5(b)**]. However, when lactose is present, the sugar binds to the repressor molecule and causes an allosteric conformational change. This change renders the repressor incapable of interacting with operator DNA [**Figure 15–5(c)**]. In the absence of the repressor–operator interaction, RNA polymerase transcribes the structural genes, and the enzymes necessary for lactose metabolism are produced. Because transcription occurs only when the repressor *fails* to bind to the operator region, regulation is said to be under *negative control.*

To summarize, the operon model invokes a series of molecular interactions between proteins, inducers, and DNA to explain the efficient regulation of structural gene expression. In the absence of lactose, the enzymes encoded by the genes are not needed, and expression of genes encoding these enzymes is repressed. When lactose is present, it indirectly induces the transcription of the structural genes by interacting with the repressor.* If all lactose is metabolized, none is available to bind to the repressor, which is again free to bind to operator DNA and repress transcription.

Both the *I*$^-$ and *O*^{C} constitutive mutations interfere with these molecular interactions, allowing continuous transcription of the structural genes. In the case of the *I*$^-$ mutant, seen in **Figure 15–6(a)**, the repressor protein is altered or absent and cannot bind to the operator region, so the structural genes are always transcribed. In the case of the *O*^{C} mutant [**Figure 15–6(b)**], the nucleotide sequence of the operator DNA is altered and will not bind with a normal repressor molecule. The result is the same: the structural genes are always transcribed.

Genetic Proof of the Operon Model

The operon model leads to three major predictions that can be tested to determine its validity. The major predictions to be tested are that (1) the *I* gene produces a diffusible product; (2) the *O* region is involved in regulation but does not produce a product; and (3) the *O* region must be adjacent to the structural genes in order to regulate transcription.

The construction of partially diploid bacteria allows us to assess these assumptions, particularly those that predict *trans*-acting regulatory molecules. For example, as introduced in previously (see Chapter 8), the F plasmid may contain chromosomal genes, in which case it is designated F′. When an F$^-$ cell acquires such a plasmid, it contains its

* Technically, allolactose, an isomer of lactose, is the inducer. When lactose enters the bacterial cell, some of it is converted to allolactose by the β-galactosidase enzyme.

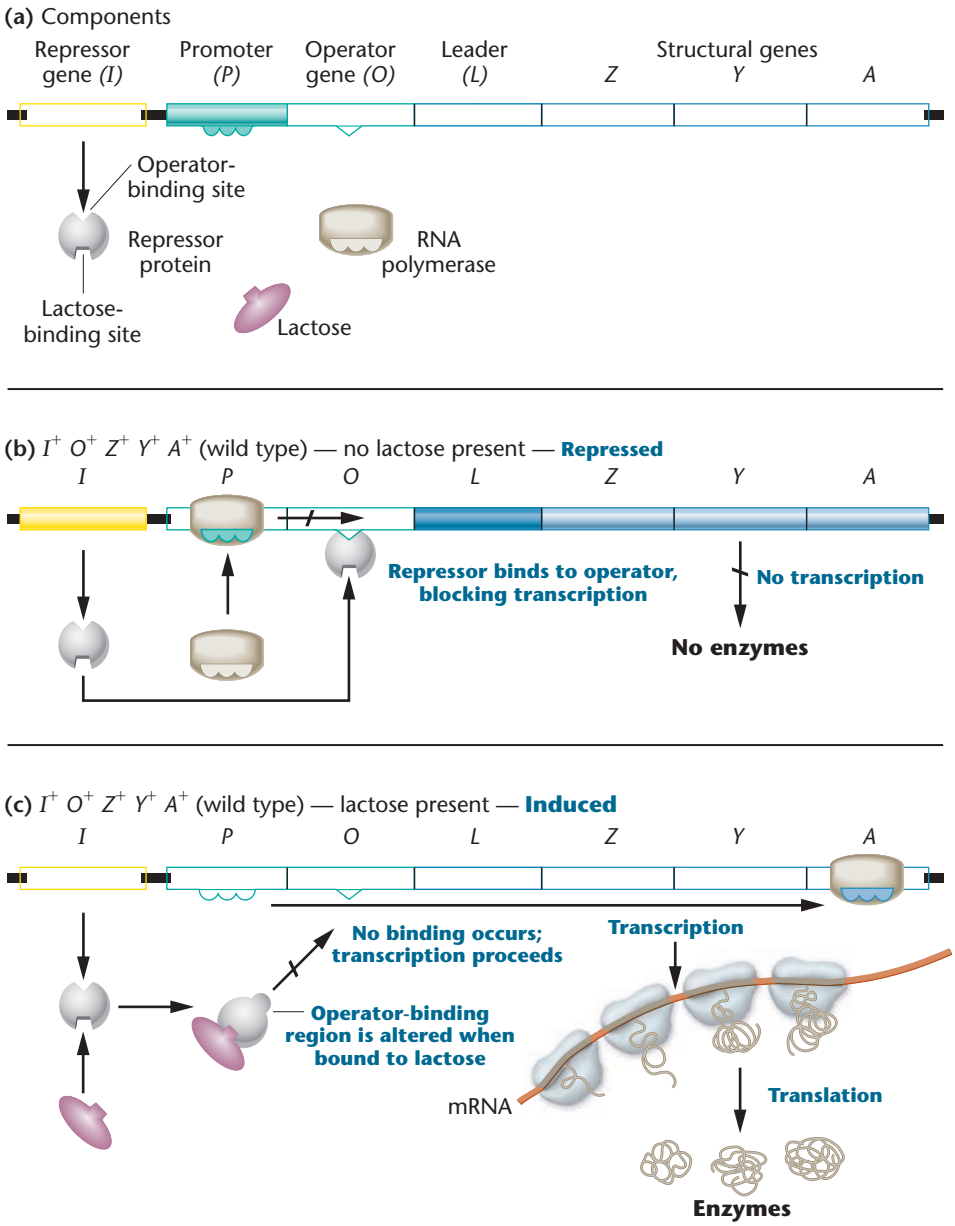


FIGURE 15–5 The components of the wild-type *lac* operon (a) and the response of the *lac* operon to the absence (b) and presence (c) of lactose. The Leader (L) sequence encodes a short region of mRNA that is 5′ of the AUG translation start codon and is not translated.

own chromosome plus one or more additional genes present in the plasmid. This creates a host cell, called a **merozygote,** that is diploid for those genes. The use of such a plasmid makes it possible, for example, to introduce an I^+ gene into a host cell whose genotype is I^- or to introduce an O^+ region into a host cell of genotype O^C . The Jacob–Monod operon model predicts how regulation should be affected in such cells. Adding an I^+ gene to an I^- cell should restore inducibility because the normal wild-type repressor, which is a *trans*-acting factor, would be produced by the inserted I^+ gene. Adding an O^+ region to an O^C cell should have no effect on constitutive enzyme production, since regulation depends on the presence of an O^+ region immediately adjacent to the structural genes—that is, O^+ is a *cis*-acting regulator.

The results of these experiments are shown in Table 15.1, where Z represents the structural genes. The inserted genes are listed after the designation F′. In both cases described here, the Jacob–Monod model is upheld (part B of Table 15.1). Part C shows the reverse experiments, where either an I^- gene or an O^C region is added to cells of normal inducible genotypes. As the model predicts, inducibility is maintained in these partial diploids.

Another prediction of the operon model is that certain mutations in the I gene should have the opposite effect of I^-. That is, instead of being constitutive because the repressor

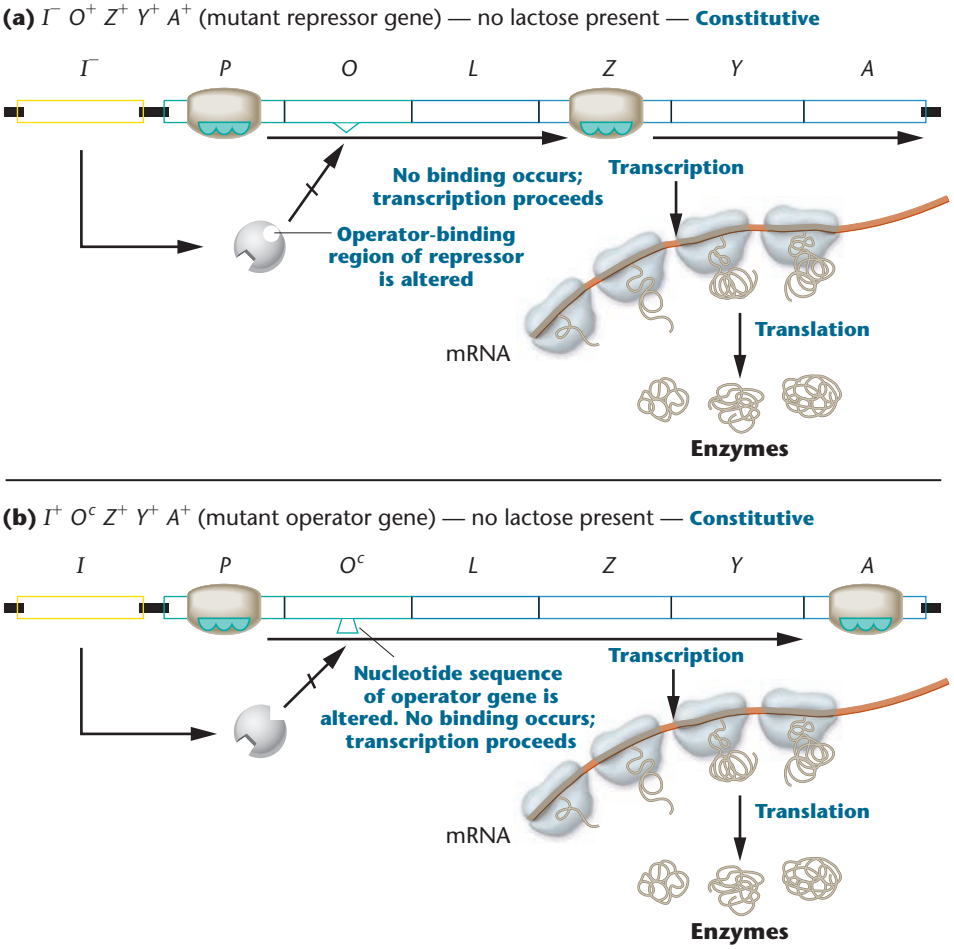


FIGURE 15–6 The response of the *lac* operon in the absence of lactose when a cell bears either the I^- (a) or the O^C (b) mutation.

TABLE 15.1 A Comparison of Gene Activity (+ or −) in the Presence or Absence of Lactose for Various *E. coli* Genotypes

		Presence of β-Galactosidase Activity	
	Genotype	Lactose Present	Lactose Absent
	$I^+O^+Z^+$	+	−
A.	$I^+O^+Z^-$	−	−
	$I^-O^+Z^+$	+	+
	$I^+O^cZ^+$	+	+
B.	$I^-O^+Z^+/F'I^+$	+	−
	$I^+O^cZ^+/F'O^+$	+	+
C.	$I^+O^+Z^+/F'I^-$	+	−
	$I^+O^+Z^+/F'O^c$	+	−
D.	$I^SO^+Z^+$	−	−
	$I^SO^+Z^+/F'I^+$	−	−

Note: In parts B to D, most genotypes are partially diploid, containing an F factor plus attached genes (F′).

can't bind the operator, mutant repressor molecules should be produced that cannot interact with the inducer, lactose. As a result, the repressor would always bind to the operator sequence, and the structural genes would be permanently repressed (Figure 15–7). If this were the case, the presence of an additional I^+ gene would have little or no effect on repression.

In fact, such a mutation, I^S, was discovered wherein the operon is "super-repressed," as shown in part D of Table 15.1. An additional I^+ gene does not effectively relieve repression of gene activity. These observations are consistent with the idea that the repressor contains separate DNA-binding domains and inducer-binding domains. The binding of lactose to the inducer-binding domain causes an allosteric change in the DNA-binding domain.

ESSENTIAL POINT

Genes involved in the metabolism of lactose are coordinately regulated by a negative control system that responds to the presence or absence of lactose. ■

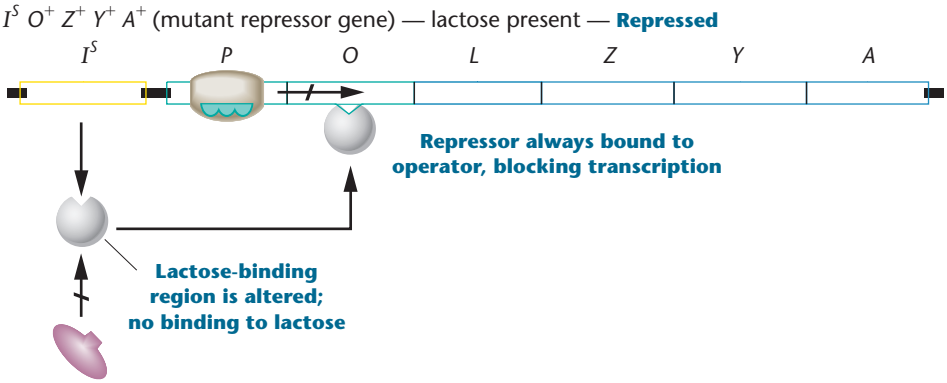


FIGURE 15–7 The response of the *lac* operon in the presence of lactose in a cell bearing the I^S mutation.

Isolation of the Repressor

Although Jacob and Monod's operon theory succeeded in explaining many aspects of genetic regulation in prokaryotes, the nature of the repressor molecule was not known when their landmark paper was published in 1961. While they had assumed that the allosteric repressor was a protein, RNA was also a candidate because activity of the molecule required the ability to bind to DNA. A single *E. coli* cell contains no more than ten or so molecules of the *lac* repressor; therefore, direct chemical identification of ten molecules in a population of millions of proteins and RNAs in a single cell presented a tremendous challenge. Nevertheless, in 1966, Walter Gilbert and Benno Müller-Hill reported the isolation of the *lac* repressor. Once the repressor was purified, it was shown to have various characteristics of a protein. The isolation of the repressor thus confirmed the operon model, which had been put forward strictly on genetic grounds.

NOW SOLVE THIS

15–1 The *lac Z*, *Y*, and *A* structural genes are transcribed as a single polycistronic mRNA; however, each structural gene contains its own initiation and termination signals essential for translation. Predict what will happen when cells growing in the presence of lactose contain a deletion of one nucleotide (a) early in the *Z* gene and (b) early in the *A* gene.

■ **HINT:** *This problem requires you to combine your understanding of the genetic expression of the lac operon, the genetic code, frameshift mutations, and termination of transcription. The key to its solution is to consider the effect of the loss of one nucleotide within a polycistronic mRNA.*

15.3 The Catabolite-Activating Protein (CAP) Exerts Positive Control over the *lac* Operon

As we discussed previously, the role of β-galactosidase is to cleave lactose into its components, glucose and galactose. However, for galactose to be used by the cell, it also must be converted to glucose. What if the cell found itself in an environment that contained an ample amount of lactose *and* glucose? Given that glucose is the preferred carbon source for *E. coli,* it would not be energetically efficient for a cell to induce transcription of the *lac* operon, make β-galactosidase, and metabolize lactose, since what it really needs—glucose—is already present. As we shall see next, a molecule called the **catabolite-activating protein** (**CAP**) helps activate expression of the *lac* operon but is able to inhibit expression when glucose is present. This inhibition is called **catabolite repression.**

To understand CAP and its role in regulation, let's backtrack for a moment. When the *lac* repressor is bound to the inducer, RNA polymerase transcribes the *lac* operon structural genes. As stated earlier in the text (see Chapter 12), transcription is initiated as a result of the binding that occurs between RNA polymerase and the nucleotide sequence of the promoter region, found upstream (5′) from the initial coding sequences. Within the *lac* operon, the promoter is located upstream of the *lac* operator region (*O*). (See Figure 15–1.) Careful examination has revealed that RNA polymerase binding is never very efficient unless CAP is also present to facilitate the process.

The mechanism is summarized in **Figure 15–8**. In the absence of glucose and under inducible conditions, CAP exerts positive control by binding to the CAP site, facilitating RNA polymerase binding at the *lac* operon promoter and thus transcription. Therefore, for maximal transcription of the structural genes, the repressor must be bound by lactose (so as not to repress *lac* operon transcription), *and* CAP must be bound to the CAP-binding site.

What role does glucose play in inhibiting CAP binding? The answer involves still another molecule, **cyclic adenosine monophosphate** (**cAMP**), upon which CAP binding is dependent. In order to bind to the *lac* operon promoter, CAP must be bound to cAMP. The level of cAMP is itself dependent on an enzyme, **adenyl cyclase,** which catalyzes the conversion of ATP to cAMP.

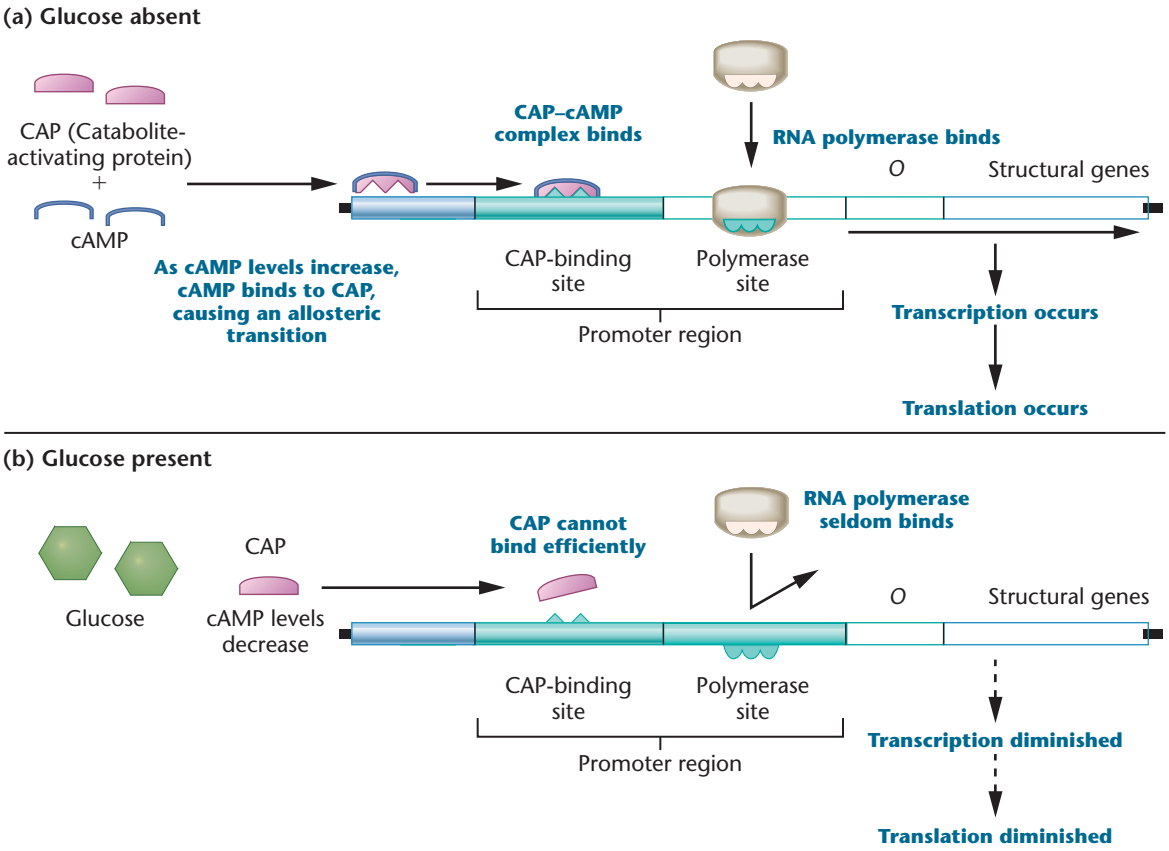


FIGURE 15–8 Catabolite repression. (a) In the absence of glucose, cAMP levels increase, resulting in the formation of a cAMP–CAP complex, which binds to the CAP site of the promoter, stimulating transcription. (b) In the presence of glucose, cAMP levels decrease, cAMP–CAP complexes are not formed, and transcription is not stimulated.

The role of glucose in catabolite repression is to inhibit the activity of adenyl cyclase, causing a decline in the level of cAMP in the cell. Under this condition, CAP cannot form the cAMP–CAP complex that is essential to the positive control of transcription of the *lac* operon.

The structures of CAP and cAMP–CAP have been examined by using X-ray crystallography. CAP is a dimer that binds adjacent regions of a specific nucleotide sequence of the DNA making up the *lac* promoter. The cAMP–CAP complex, when bound to DNA, bends the DNA, causing it to assume a new conformation.

Binding studies in solution further clarify the mechanism of gene activation. Alone, neither cAMP–CAP nor RNA polymerase has a strong affinity to bind to *lac* promoter DNA, nor does either molecule have a strong affinity to bind to the other. However, when both are together in the presence of the *lac* promoter DNA, a tightly bound complex is formed, an example of what is called **cooperative binding.** The control conferred by the cAMP–CAP provides another illustration of how the regulation of one small group of genes can be fine-tuned by several simultaneous influences.

In contrast to the negative regulation conferred by the *lac* repressor, the action of cAMP–CAP constitutes positive regulation. Thus, a combination of positive and negative regulatory mechanisms determines transcription levels of the *lac* operon. Catabolite repression involving CAP has also been observed for other inducible operons, including those controlling the metabolism of galactose and arabinose.

NOW SOLVE THIS

15–2 Predict the level of gene expression of the *lac* operon, as well as the status of the *lac* repressor and the CAP protein, when bacterial growth media contain the following sugars: (a) no lactose or glucose, (b) lactose but no glucose, (c) glucose but no lactose, (d) both lactose and glucose.

■ **HINT:** *This problem asks you to combine your knowledge of* lac *operon regulation with your understanding of how catabolite repression affects this regulation. The key to its solution is to keep in mind that regulation involving lactose is a negative control system, while regulation involving glucose and catabolite repression is a positive control system.*

ESSENTIAL POINT

The catabolite-activating protein (CAP) exerts positive control over *lac* gene expression by interacting with RNA polymerase at the *lac* promoter and by responding to the levels of cyclic AMP in the bacterial cell. ■

15.4 The Tryptophan (*trp*) Operon in *E. coli* Is a Repressible Gene System

Although inducible gene regulation had been known for some time, it was not until 1953 that Monod and colleagues discovered a repressible system. Studies on the biosynthesis of the essential amino acid tryptophan revealed that, if tryptophan is present in sufficient quantity in the growth medium, the enzymes necessary for its synthesis (such as **tryptophan synthase**) are not produced. It is energetically advantageous for bacteria to repress expression of genes involved in tryptophan synthesis when ample tryptophan is present in the growth medium.

Further investigation showed that enzymes encoded by five contiguous genes on the *E. coli* chromosome are involved in tryptophan synthesis. These genes are part of an operon and, in the presence of tryptophan, all are coordinately repressed, and none of the enzymes is produced. Because of the great similarity between this repression and the induction of enzymes for lactose metabolism, Jacob and Monod proposed a model of gene regulation resembling that of the *lac* system (**Figure 15–9**).

The model suggests the presence of a *normally inactive repressor* that alone cannot interact with the operator region of the operon. However, the repressor is an allosteric molecule that can bind to tryptophan. When tryptophan is present, the resultant complex of repressor and tryptophan attains a new conformation that binds to the operator, repressing transcription. Thus, when tryptophan, the end product of this anabolic pathway, is present, the operon is repressed and enzymes are not made. Since the regulatory complex inhibits transcription of the operon, this repressible system is under negative control. And as tryptophan participates in repression, it is referred to as a **corepressor** in this regulatory scheme.

Evidence for the *trp* Operon

Support for the concept of a repressible operon is based primarily on the isolation of two distinct categories of constitutive mutations. The first class, *trpR*$^-$, maps at a considerable distance from the structural genes. This locus represents the gene coding for the repressor. Presumably, the mutation either inhibits the interaction of the repressor with tryptophan or inhibits repressor formation entirely. Whichever the case, no repression is present in cells with the *trpR*$^-$ mutation. As expected, if the *trpR*$^+$ gene encodes a functional repressor molecule, the presence of a copy of this gene will restore repressibility.

The second constitutive mutant is analogous to the O^C mutant of the lactose operon because it maps immediately adjacent to the structural genes. Furthermore, the addition of a wild-type operator gene into mutant cells (as an external element) does not restore repression. This is predictable if the mutant operator, which must be present in *cis*, no longer interacts with the repressor–tryptophan complex.

The entire *trp* operon has now been well defined, as shown in Figure 15–9. Five contiguous structural genes (*trpE, D, C, B,* and *A*) are transcribed as a polycistronic mRNA directing translation of the enzymes that catalyze the biosynthesis of tryptophan. As in the *lac* operon, a promoter region (*trpP*) represents the binding site for RNA polymerase, and an operator region (*trpO*) is the binding site for the repressor. In the absence of repressor binding, transcription initiates within the overlapping *trpP–trpO* region and proceeds along a **leader sequence** 162 nucleotides prior to the first structural gene (*trpE*). Within that leader sequence, still another regulatory site exists, called an *attenuator,* which we describe in the next section of this chapter. As we shall see, the attenuator is also an integral part of the control mechanism of the operon.

ESSENTIAL POINT

Unlike the inducible *lac* operon, the *trp* operon is repressible. In the presence of tryptophan, the repressor binds to the regulatory region of the *trp* operon and represses transcription initiation. ■

EVOLVING CONCEPT OF THE GENE

The groundbreaking work of Jacob, Monod, and Lwoff in the early 1960s, which established the operon model for the regulation of gene expression in bacteria, expanded the concept of the gene to include noncoding regulatory sequences that are present upstream (5′) from the coding region. In bacterial operons, the transcription of several contiguous structural genes whose products are involved in the same biochemical pathway is regulated by a single set of regulatory sequences. ■

15.5 Alterations to RNA Secondary Structure Also Contribute to Prokaryotic Gene Regulation

In the preceding sections of this chapter, we focused on gene regulation brought about by DNA-binding regulatory proteins that interact with promoter and operator

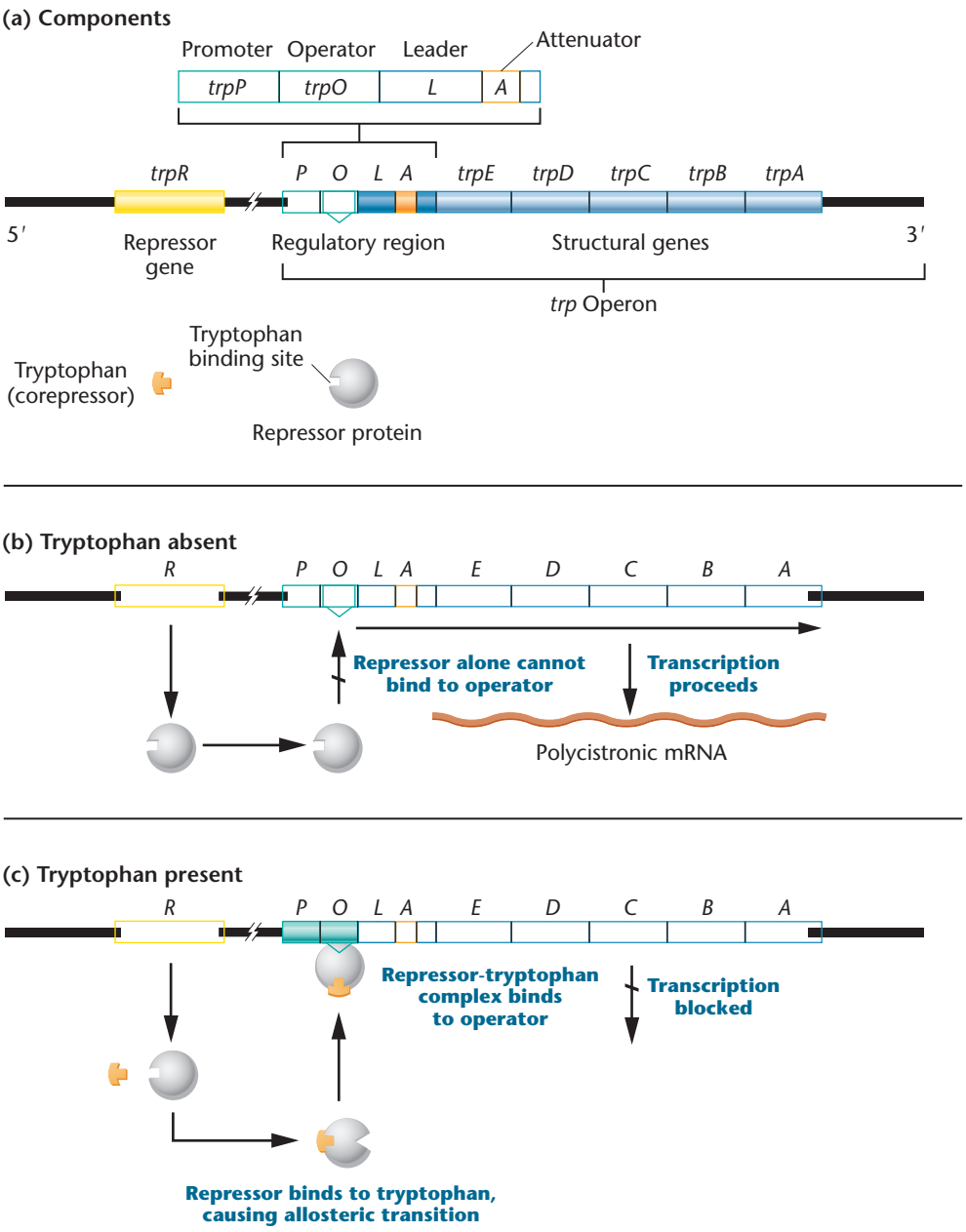


FIGURE 15–9 (a) The components involved in the regulation of the tryptophan operon. (b) Regulatory conditions are depicted that involve either activation or (c) repression of the structural genes. In the absence of tryptophan, an inactive repressor is made that cannot bind to the operator (*O*), thus allowing transcription to proceed. In the presence of tryptophan, it binds to the repressor, causing an allosteric transition to occur. This complex binds to the operator region, leading to repression of the operon.

regions of the genes to be regulated. These regulatory proteins, such as the *lac* repressor and the CAP protein, act to decrease or increase transcription initiation from their target promoters by affecting the binding of RNA polymerase to the promoter.

Gene regulation in prokaryotes can also occur through the interactions of regulatory molecules with specific regions of a nascent mRNA, after transcription has been initiated. The binding of these regulatory molecules alters the secondary structure of the mRNA, leading to premature transcription termination or repression of translation. We will discuss two types of regulation by RNA secondary structure—attenuation and riboswitches. Both types help to fine-tune prokaryotic gene regulation and are used in addition to regulation of transcription initiation.

Transcription Attenuation

Charles Yanofsky, Kevin Bertrand, and their colleagues defined the mechanisms of bacterial attenuation. They observed that, when tryptophan is present and the *trp* operon is repressed, initiation of transcription still occurs at a low level but is subsequently terminated at a point about 140 nucleotides along the transcript. They called this process **attenuation,** as it further diminishes expression of the operon. In contrast, when tryptophan is absent or present in very low concentrations, transcription is initiated but is *not* subsequently terminated, instead continuing beyond the leader sequence into the structural genes.

The site involved in attenuation is located 115 to 140 nucleotides into the leader sequence and is referred to as the **attenuator.** (See Figure 15–9.)

Yanofsky and colleagues presented a model to explain how attenuation occurs and is regulated. The initial DNA sequence that is transcribed gives rise to an mRNA sequence that has the potential to fold into two mutually exclusive stem-loop structures referred to as "hairpins." In the presence of excess tryptophan, the mRNA hairpin that is formed behaves as a **terminator** structure, and transcription is almost always terminated prematurely, just beyond the attenuator. On the other hand, if tryptophan is scarce, an alternative mRNA hairpin referred to as the **antiterminator hairpin** is formed. Transcription proceeds past the antiterminator hairpin region, and the entire mRNA is subsequently produced.

A key point in Yanofsky's model is that the leader transcript must be translated in order for the antiterminator hairpin to form. The leader transcript contains two triplets (UGG) that encode tryptophan, and these are present just downstream of the initial AUG sequence that signals the initiation of translation by ribosomes. When adequate tryptophan is present, charged $tRNA^{Trp}$ is present in the cell. As a result, ribosomes translate these UGG triplets, proceed through the attenuator, and allow the *terminator hairpin* to form. The terminator hairpin signals RNA polymerase to prematurely terminate transcription, and the operon is not transcribed. If cells are starved of tryptophan, charged $tRNA^{Trp}$ is unavailable. As a result, ribosomes "stall" during translation of the UGG triplets. The presence of ribosomes in this region of the mRNA interferes with the formation of the terminator hairpin, but allows the formation of the antiterminator hairpin within the leader transcript. As a result, transcription proceeds, leading to expression of the entire set of structural genes.

Many other bacterial operons use attenuation to control gene expression. These include operons that encode enzymes involved in the biosynthesis of amino acids such as threonine, histidine, leucine, and phenylalanine. As with the *trp* operon, attenuation occurs in a leader sequence that contains an attenuator region.

Riboswitches

Since the elucidation of attenuation in the *trp* operon, numerous cases of gene regulation that also depend on alternative forms of mRNA secondary structure have been documented. These involve what are called **riboswitches,** which are mRNA sequences (or elements) present in the 5′-untranslated region (5′-UTR) upstream from the coding sequences. These elements are capable of binding with small molecule ligands, such as metabolites, whose synthesis or activity is controlled by the genes encoded by the mRNA. Such binding causes a conformational change in one domain of the riboswitch element, which induces another change at a second RNA domain, most often creating a transcription *terminator structure.* This terminator structure interfaces directly with the transcriptional machinery and shuts it down.

Riboswitches can recognize a broad range of ligands, including amino acids, purines, vitamin cofactors, amino sugars, and metal ions, among others. They are widespread in bacteria. In *Bacillus subtilis,* for example, approximately 5 percent of this bacterium's genes are regulated by riboswitches. They are also found in archaea, fungi, and plants, and may be present in animals as well.

The two important domains within a riboswitch are the ligand-binding site, called the **aptamer,** and the **expression platform,** which is capable of forming the terminator structure. Figure 15–10 illustrates the principles involved in

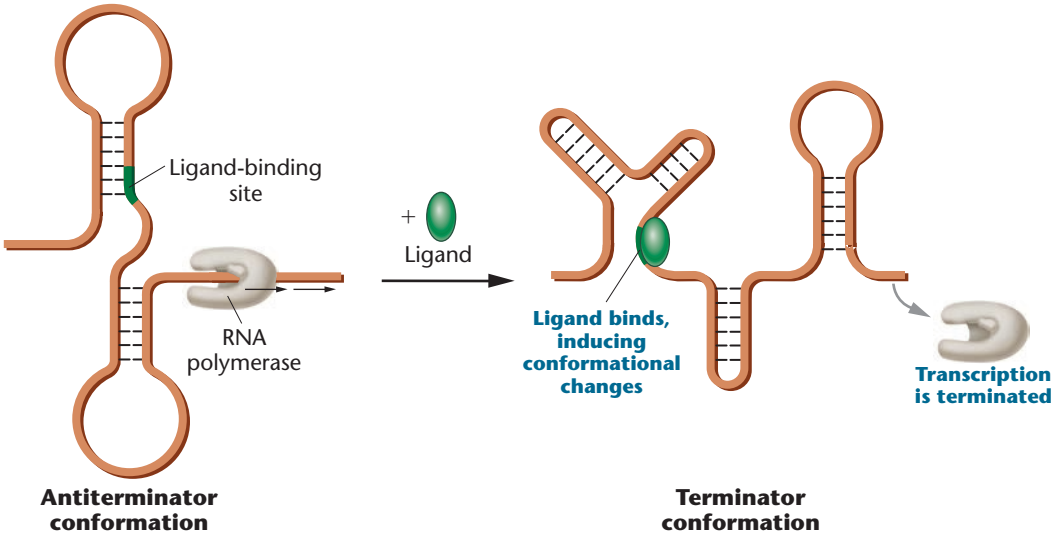


FIGURE 15–10 Illustration of the mechanism of riboswitch regulation of gene expression, where the default position (left) is in the antiterminator conformation. Upon binding by the ligand, the mRNA adopts the terminator conformation (right).

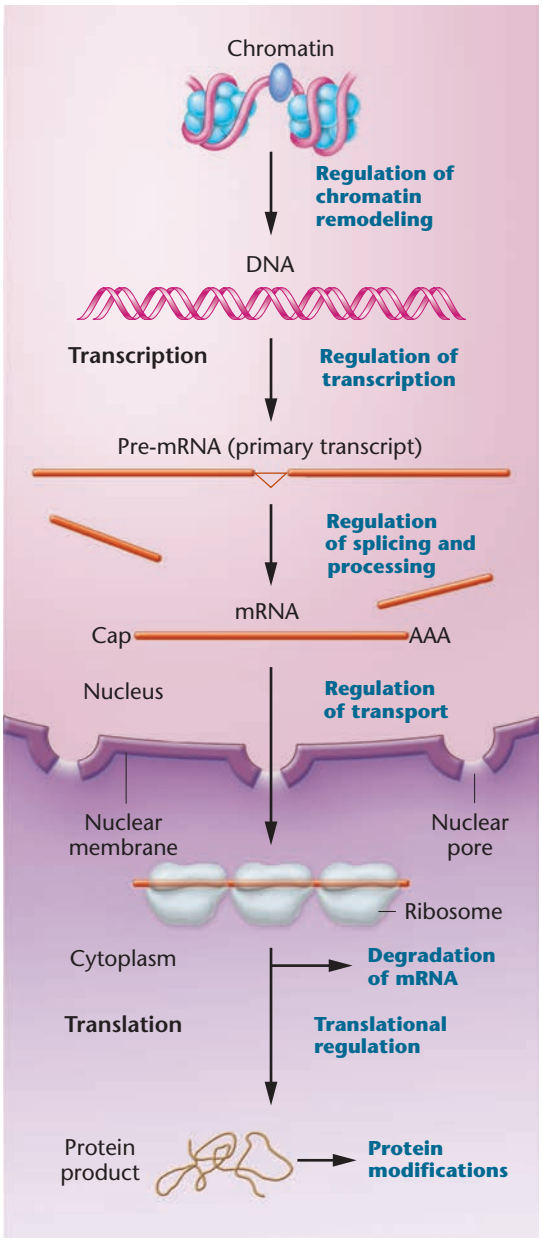


FIGURE 15–11 Regulation can occur at any stage in the expression of genetic material in eukaryotes. All these forms of regulation affect the degree to which a gene is expressed.

riboswitch control. The 5′-UTR of an mRNA is shown on the left side of the figure in the absence of the ligand (metabolite). RNA polymerase has transcribed the unbound ligand-binding site, and in the *default conformation,* the expression domain adopts an *antiterminator conformation.* Thus, transcription continues through the expression platform and into the coding region. On the right side of the figure, the presence of the ligand on the ligand-binding site induces an alternative conformation in the expression platform, creating the *terminator conformation.* RNA polymerase is effectively blocked and transcription ceases.

We will discuss other recent findings involving RNA-mediated gene expression later in the text (see Section 15.12 and Special Topic Chapter 2—Emerging Roles of RNA).

ESSENTIAL POINT

Attenuation and riboswitches regulate gene expression by inducing alterations to mRNA secondary structure, leading to premature termination of transcription. ■

15.6 Eukaryotic Gene Regulation Differs from That in Prokaryotes

Virtually all cells in a multicellular eukaryotic organism contain a complete genome; however, only a subset of genes is expressed in any particular cell type. For example, some white blood cells express genes encoding certain immunoglobulins, allowing these cells to synthesize antibodies that defend the organism from infection and foreign agents. However, skin, kidney, and liver cells do not express immunoglobulin genes. Pancreatic islet cells synthesize and secrete insulin in response to the presence of blood sugars; however, they do not manufacture immunoglobulins. In addition, they do not synthesize insulin when it is not required. Eukaryotic cells, as part of multicellular organisms, do not grow solely in response to the availability of nutrients. Instead, they regulate their growth and division to occur at appropriate places in the body and at appropriate times during development. The loss of gene regulation that controls normal cell growth and division may lead to developmental defects or cancer.

Eukaryotes employ a wide range of mechanisms for altering the expression of genes. In contrast to prokaryotic gene regulation, which occurs primarily at the level of transcription initiation, regulation of gene expression in eukaryotes can occur at many different levels. These include the initiation of transcription, mRNA modifications and stability, and the synthesis, modification, and stability of the protein product (**Figure 15–11**).

Several features of eukaryotic cells make it possible for them to use more types of gene regulation than are possible in prokaryotic cells:

- Eukaryotic cells contain a much greater amount of DNA than do prokaryotic cells. This DNA is associated with histones and other proteins to form highly compact chromatin structures within an enclosed nucleus. Eukaryotic cells modify this structural organization in order to influence gene expression.
- The mRNAs of most eukaryotic genes must be spliced, capped, and polyadenylated prior to transport from the nucleus. Each of these processes can be regulated

in order to influence the numbers and types of mRNAs available for translation.

- Genetic information in eukaryotes is carried on many chromosomes (rather than just one), and these chromosomes are enclosed within a double-membrane-bound nucleus. After transcription, transport of RNAs into the cytoplasm can be regulated in order to modulate the availability of mRNAs for translation.
- Eukaryotic mRNAs can have a wide range of half-lives ($t_{1/2}$). In contrast, the majority of prokaryotic mRNAs decay very rapidly. Rapid turnover of mRNAs allows prokaryotic cells to rapidly respond to environmental changes. In eukaryotes, the complement of mRNAs in each cell type can be more subtly manipulated by altering mRNA decay rates over a larger range.
- In eukaryotes, translation rates can be modulated, as well as the way proteins are processed, modified, and degraded.

In the following sections, we examine some of the major ways in which eukaryotic gene expression is regulated. As most eukaryotic genes are regulated, at least in part, at the transcriptional level, we will emphasize transcriptional control. In addition, we will limit our discussion to regulation of genes transcribed by RNA polymerase II. As we previously described in Chapter 12, three different RNA polymerases transcribe eukaryotic genes. RNA polymerase II transcribes all mRNAs and some small nuclear RNAs, whereas RNA polymerases I and III transcribe ribosomal RNAs, some small nuclear RNAs, and transfer RNAs. Transcription by each of these RNA polymerases is regulated differently, with RNA polymerase II having the most diverse and complex mechanisms.

15.7 Eukaryotic Gene Expression Is Influenced by Chromatin Modifications

Two structural features of eukaryotic genes distinguish them from the genes of prokaryotes. First, eukaryotic genes are situated on chromosomes that occupy a distinct location within the cell—the nucleus. This sequestering of genetic information in a discrete compartment allows the proteins that directly regulate transcription to be kept apart from those involved with translation and other aspects of cellular metabolism. Second, as described earlier in the text (see Chapter 11), eukaryotic DNA is combined with histones and nonhistone proteins to form chromatin. Chromatin's basic structure is characterized by repeating units called nucleosomes that are wound into 30-nm fibers, which in turn form other, even more compact structures. The presence of these compact chromatin structures is inhibitory to many processes, including transcription, replication, and DNA repair. In this section, we outline some of the ways in which eukaryotic cells modify chromatin in order to regulate gene expression.

Chromosome Territories and Transcription Factories

Recent research has revealed that the interphase nucleus is not a bag of tangled chromosome arms, but has a highly organized structure. In the interphase nucleus, each chromosome occupies a discrete domain called a **chromosome territory** and stays separate from other chromosomes. Channels between chromosomes contain little or no DNA and are called **interchromosomal domains.**

Transcriptionally active genes appear to be located at the edges of chromosome territories next to interchromosomal domain channels. Scientists hypothesize that this organization may bring actively expressed genes into closer association with transcription factors, or with other actively expressed genes, thereby facilitating their coordinated expression.

Another feature within the nucleus—the **transcription factory**—may also contribute to regulating gene expression. Transcription factories are nuclear sites at which most RNA polymerase II transcription occurs. These sites also contain the majority of active RNA polymerase and other transcription factors. It is not yet clear whether the formation of transcription factories is a prerequisite or a consequence of transcription initiation; however, by concentrating transcription proteins and actively transcribed genes in specific locations in the nucleus, the cell may enhance the expression of these genes.

Histone Modifications and Nucleosomal Chromatin Remodeling

Chromatin modification is an important step in gene regulation. Chromatin modification appears to be a prerequisite for transcription of some eukaryotic genes, although it can occur simultaneously with transcription of other genes.

Chromatin can be modified in two general ways. The first involves changes to nucleosomes, and the second involves modifications to DNA. In this subsection, we will discuss changes to the nucleosomal component of chromatin. In the next subsection, we present DNA modifications, specifically DNA methylation.

The tight association of DNA with nucleosomes and other chromatin-binding proteins inhibits access of the DNA to the proteins involved in many functions, including transcription. This inhibitory structure is often referred to as "closed" chromatin. Before transcription can be initiated

within nucleosomal chromatin, the structure of chromatin must become "open" to transcription regulatory factors and enzymes such as RNA polymerases.

Nucleosomal chromatin can be modified in three ways. The first involves changes in nucleosome composition that can affect gene transcription. For example, most nucleosomes contain the normal histone H2A. Some gene promoter regions may be flanked by nucleosomes containing variant histones, such as H2A.Z. The presence of the H2A.Z variant within these nucleosomes affects nucleosome mobility and positioning on DNA. As a result, a gene promoter associated with these variant nucleosomes may be either transcriptionally activated or repressed, depending on the nucleosome position.

A second mechanism of chromatin alteration involves histone modification. Histone modification involves the covalent bonding of functional groups onto the N-terminal tails of histone proteins. The most common histone modifications are the addition of acetyl, methyl, or phosphate groups onto the basic amino acids of histone tails. Acetylation decreases the positive charge on histones, resulting in a reduced affinity of the histone for DNA. In turn, this may assist the formation of open chromatin conformations, which would allow the binding of transcription regulatory proteins to DNA.

Histone acetylation is catalyzed by **histone acetyltransferase enzymes** (**HATs**). In some cases, HATs are recruited to genes by the presence of certain transcription activator proteins that bind to transcription regulatory regions. In other cases, transcription activator proteins themselves have HAT activity. Of course, what can be opened can also be closed. In that case, histone deacetylases (HDACs) remove acetyl groups from histone tails. HDACs can be recruited to genes by the presence of certain repressor proteins on regulatory regions.

The third mechanism of chromatin alteration is chromatin remodeling, which involves the repositioning or removal of nucleosomes on DNA, brought about by chromatin remodeling complexes. Chromatin remodeling complexes are large multi-subunit complexes that use the energy of ATP hydrolysis to move and rearrange nucleosomes along the DNA. Repositioned nucleosomes make regions of the chromosome accessible to transcription regulatory proteins, such as transcription activators and RNA polymerase II. One of the best-studied remodeling complexes is the SWI/SNF complex. Remodelers such as SWI/SNF can act in several different ways (**Figure 15–12**). They may loosen the attachment between histones and DNA, resulting in the nucleosome sliding along the DNA and exposing regulatory regions. Alternatively, they may loosen the DNA strand from the nucleosome core, or they may cause reorganization of the internal nucleosome components. In all cases, the DNA is left transiently exposed to association with transcription factors and RNA polymerase. Like HATs, chromatin remodeling

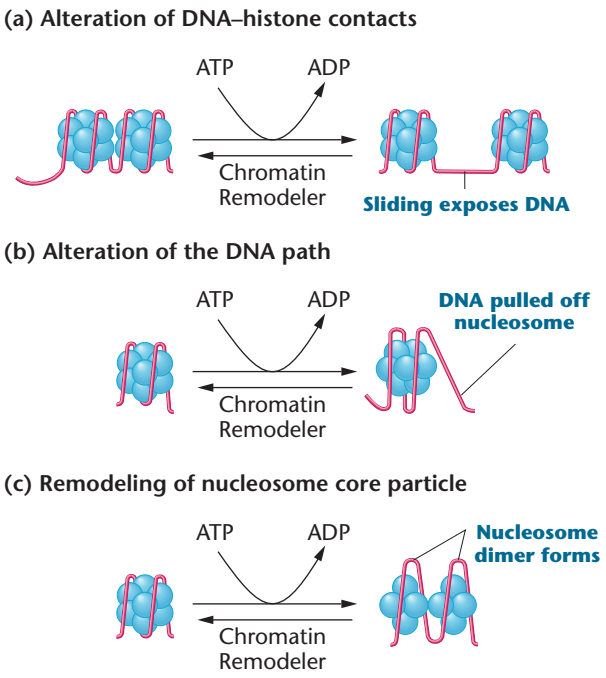


FIGURE 15–12 Three ways by which chromatin remodelers, such as the SWI/SNF complex, alter the association of nucleosomes with DNA. (a) The DNA–histone contacts may be loosened, allowing the nucleosomes to slide along the DNA, exposing DNA regulatory regions. (b) The path of the DNA around a nucleosome core particle may be altered. (c) Components of the core nucleosome particle may be rearranged, resulting in a modified nucleosome structure.

complexes can be recruited to DNA by transcription activator proteins that are bound to specific regions of DNA. The interactions of remodeling complexes are also affected by the presence or absence of histone modifications.

DNA Methylation

Another type of change in chromatin that plays a role in gene regulation is the addition or removal of methyl groups to or from bases in DNA. **DNA methylation** most often involves cytosine. In the genome of any given eukaryotic species, approximately 5 percent of the cytosine residues are methylated. However, the extent of methylation can be tissue specific and can vary from less than 2 percent to more than 7 percent of cytosine residues.

Evidence of a role for methylation in eukaryotic gene expression is based on a number of observations. First, an inverse relationship exists between the degree of methylation and the degree of expression. Large transcriptionally inert regions of the genome, such as the inactivated X chromosome in mammalian female cells, are often heavily methylated. Second, methylation patterns are tissue specific and, once established, are heritable for all cells of that tissue. It appears that proper patterns of DNA methylation are essential for normal mammalian development. Undifferentiated

embryonic cells that are not able to methylate DNA die when they are required to differentiate into specialized cell types.

Perhaps the most direct evidence for the role of methylation in gene expression comes from studies using base analogs. The nucleotide **5-azacytidine** can be incorporated into DNA in place of cytidine during DNA replication. This analog cannot be methylated, causing the undermethylation of the sites where it is incorporated. The incorporation of 5-azacytidine into DNA changes the pattern of gene expression and stimulates expression of alleles on inactivated X chromosomes. In addition, the presence of 5-azacytidine in DNA can induce the expression of genes that would normally be silent in certain differentiated cells.

How might methylation affect gene regulation? Data from *in vitro* studies suggest that methylation can repress transcription by inhibiting the binding of transcription factors to DNA. Methylated DNA may also recruit repressive chromatin remodeling complexes to gene-regulatory regions.

ESSENTIAL POINT

Eukaryotic gene regulation at the level of chromatin may involve gene-specific chromatin remodeling, histone modifications, or DNA modifications. These modifications may either allow or inhibit access of promoters and enhancers to transcription factors, resulting in increased or decreased levels of transcription initiation. ■

NOW SOLVE THIS

15–3 Cancer cells often have abnormal patterns of chromatin modifications. In some cancers, the DNA repair genes *MLH1* and *BRCA1* are hypermethylated on their promoter regions. Explain how this abnormal methylation pattern could contribute to cancer.

■ **HINT:** *This problem involves an understanding of the types of genes that are mutated in cancer cells. The key to its solution is to consider how methylation affects gene expression of cancer-related genes.*

15.8 Eukaryotic Transcription Regulation Requires Specific *Cis*-Acting Sites

As in prokaryotes, eukaryotic transcription regulation is controlled by *trans*-acting regulatory proteins that bind to specific *cis*-acting DNA sequences located in and around eukaryotic genes. Although these *cis*-acting sequences do not, by themselves, regulate gene transcription, they are essential because they position regulatory proteins in regions where those proteins can act to stimulate or repress transcription of the associated gene. In this section, we will discuss some of these *cis*-acting DNA sequences including promoters, enhancers, and silencers.

Promoters

A **promoter** is a region of DNA that binds one or more proteins that regulate transcription initiation. Promoters are located immediately adjacent to the genes they regulate. They may be up to several hundred nucleotides in length and specify where transcription begins and the direction of transcription along the DNA. Within promoters are a number of **promoter elements**—short nucleotide sequences that bind specific regulatory factors.

There are two subcategories within eukaryotic promoters. First, the **core promoter** determines the accurate initiation of transcription by RNA polymerase II. Second, **proximal promoter elements** are those that modulate the efficiency of basal levels of transcription.

Recent bioinformatic research reveals that there is a great deal of diversity in eukaryotic core promoters in terms of both their structures and functions. Core promoters are now thought to be either *focused* or *dispersed*. **Focused promoters** specify transcription initiation at a single specific nucleotide (the transcription start site). In contrast, **dispersed promoters** direct initiation from a number of weak transcription start sites located over a 50- to 100-nucleotide region (**Figure 15–13**). Focused transcription initiation is the major type of initiation for most genes of lower eukaryotes, but for only about 30 percent of vertebrate genes. Focused promoters are usually associated with genes whose transcription levels are highly regulated, whereas dispersed promoters are associated with genes that are transcribed constitutively.

Little is known about the DNA elements that make up dispersed promoters. These promoters are usually found

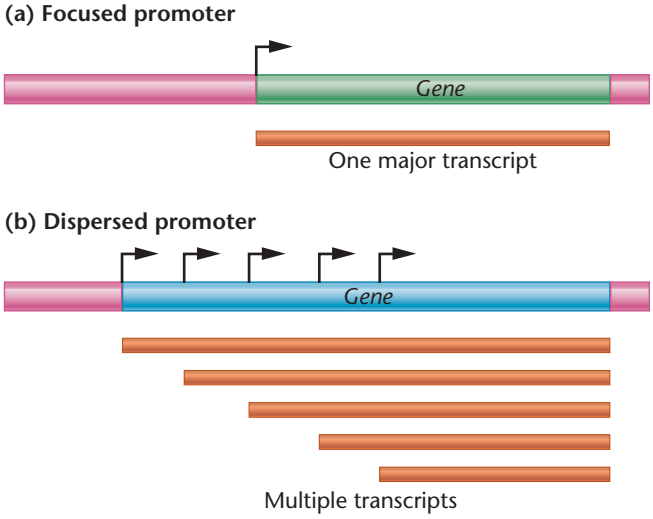


FIGURE 15–13 Focused and dispersed promoters. Focused promoters (a) specify one specific transcription initiation site. Dispersed promoters (b) specify weak transcription initiation at multiple start site positions over an approximately 100-bp region. Transcription start sites and the directions of transcription are indicated with arrows.

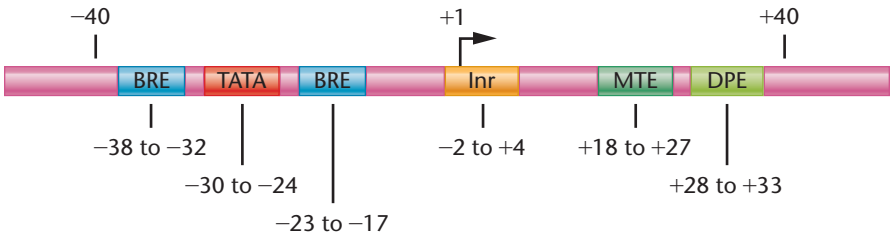


FIGURE 15–14 Core-promoter elements found in focused promoters. Core-promoter elements are usually located between −40 and +40 nucleotides, relative to the transcription start site, indicated as +1. None of these elements is universal, and a core promoter may contain only one, or several, of these elements. BRE is the TFIIB recognition element, TATA is the TATA box, Inr is the initiator element, MTE is the motif ten element, and DPE is the downstream promoter element.

within CG-rich regions, suggesting that chromatin modifications may influence initiation from these promoters. Some data suggest that dispersed promoters contain the same types of elements as focused promoters; however, these elements contain multiple mismatches to the element consensus sequences, perhaps accounting for the low levels of transcription initiation from these types of promoters.

Much more is known about the structure of focused promoters. These promoters are made up of one or more DNA sequence elements, as summarized in **Figure 15–14**. Each of these elements is found in only some core promoters, with no element being a universal component of all focused promoters.

The **Inr element** encompasses the transcription start site, from approximately nucleotides −2 to +4, relative to the start site. In humans, the Inr consensus sequence is YYAN$^{A}/_{T}$YY (where Y indicates any pyrimidine nucleotide and N indicates any nucleotide). The transcription start site is the first A residue at +1. The **TATA box** element is located at approximately −30 relative to the transcription start site and has the consensus sequence TATA$^{A}/_{T}$AAR (where R indicates any purine nucleotide). The **BRE** is found in some core promoters at positions either immediately upstream or downstream from the TATA box. The **MTE** and **DPE** sequence motifs are located downstream of the transcription start site, at approximately +18 to +27 and +28 to +32 respectively.

In addition to core-promoter elements, many promoters also contain **proximal-promoter elements** located upstream of the TATA box and BRE. Proximal-promoter elements act along with the core-promoter elements to increase the levels of basal transcription. For example, the **CAAT box** is a common proximal-promoter element. It has the consensus sequence CAAT or CCAAT and is usually located about 70 to 80 base pairs upstream from the start site. Mutational analysis suggests that CAAT boxes (when present) are critical to the promoter's ability to initiate transcription. Mutations on either side of this element have no effect on transcription, whereas mutations within the CAAT sequence dramatically lower the rate of transcription. **Figure 15–15** summarizes the transcriptional effects of mutations in the CAAT box and other promoter elements. The **GC box** is another element often found in proximal promoter regions and has the consensus sequence GGGCGG. It is located, in one or more copies, at about position −110.

Enhancers and Silencers

Although eukaryotic promoter elements are essential for basal or low levels of transcription initiation, more dramatic changes in transcription initiation require the presence of other sequence elements known as enhancers and silencers.

Like promoters, **enhancers** are *cis* regulators because they function when adjacent to the structural genes they regulate. However, unlike promoters, enhancers can be located on either side of a gene, at some distance from the gene, or even within the gene. Enhancers are necessary for achieving the maximum level of transcription. In addition, enhancers are responsible for time- and tissue-specific gene expression. Thus, there is some degree of analogy between enhancers and operator regions in prokaryotes. However, enhancers are more complex in both structure and function.

Several features distinguish promoters from enhancers:

1. The position of an enhancer need not be limited in position; it will function whether it is upstream, downstream, or within the gene it regulates.
2. The orientation of an enhancer can be inverted without significant effect on its action.
3. If an enhancer is experimentally moved adjacent to a gene elsewhere in the genome, or if an unrelated gene is placed near an enhancer, the transcription of the newly adjacent gene is enhanced.

Another type of *cis*-acting transcription regulatory element, the **silencer,** acts upon eukaryotic genes to repress the level of transcription initiation. Silencers, like enhancers, are short DNA sequence elements that affect the rate of transcription initiated from an associated promoter. They often act in tissue- or temporal-specific ways to control gene expression.

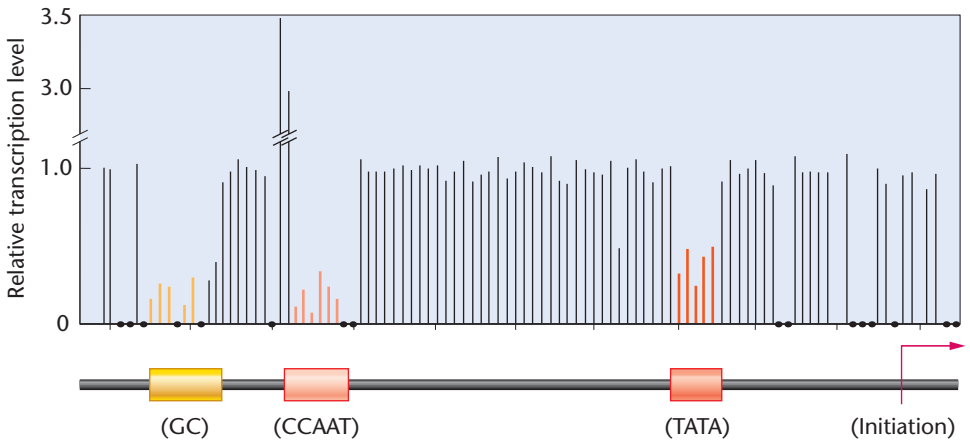


FIGURE 15–15 Summary of the effects on transcription levels of different point mutations in the promoter region of the β-globin gene. Each line represents the level of transcription produced in a separate experiment by a single-nucleotide mutation (relative to wild-type) at a particular location. Dots represent nucleotides for which no mutation was obtained. Note that mutations within specific elements of the promoter have the greatest effects on the level of transcription.

ESSENTIAL POINT

Eukaryotic transcription regulation requires gene-specific promoter, enhancer, and silencer elements. The presence of these *cis*-acting regulatory sites can affect transcription in tissue- and temporal-specific ways. ■

15.9 Eukaryotic Transcription Initiation is Regulated by Transcription Factors That Bind to *Cis*-Acting Sites

Eukaryotic promoters, enhancers, and silencers influence transcription initiation by acting as binding sites for transcription regulatory proteins. These transcription regulatory proteins, known as **transcription factors,** can have diverse and complicated effects on transcription. Some transcription factors increase the levels of transcription initiation and are known as **activators,** whereas others reduce transcription levels and are known as **repressors.**

Some transcription factors are expressed in tissue-specific ways, regulating their target genes for tissue-specific levels of expression. In addition, some transcription factors are expressed in cells only at certain times during development or in response to external physiological signals. In some cases, a transcription factor that binds to a *cis*-acting site and regulates a certain gene may be present in a cell and may even bind to its appropriate *cis*-acting site but will only become active when modified structurally (for example, by phosphorylation or by binding to a coactivator such as a hormone). These modifications to transcription factors can also be regulated in tissue- or temporal-specific ways. In addition, different transcription factors may compete for binding to a given DNA sequence or to one of two overlapping sequences. In these cases, transcription factor concentrations and the strength with which each factor binds to DNA will dictate which factor binds. The same site may also bind different factors in different tissues. Finally, multiple transcription factors that bind to several different enhancers and promoter elements within a gene-regulatory region can interact with each other to fine-tune the levels and timing of transcription initiation.

The Human Metallothionein IIA Gene: Multiple *Cis*-Acting Elements and Transcription Factors

The **human metallothionein IIA gene** (*hMTIIA*) provides an example of how one gene can be transcriptionally regulated through the interplay of multiple promoter and enhancer elements and the transcription factors that bind to them. The product of the *hMTIIA* gene is a protein that binds to heavy metals such as zinc and cadmium, thereby protecting cells from the toxic effects of high levels of these metals. The gene is expressed at low levels in all cells but is transcriptionally induced to express at high levels when cells are exposed to heavy metals and steroid hormones such as glucocorticoids.

The *cis*-acting regulatory elements controlling transcription of the *hMTIIA* gene include promoter, enhancer, and silencer elements (**Figure 15–16**). Each *cis*-acting element is a short DNA sequence that has specificity for binding to one or more transcription factors.

The *hMTIIA* gene contains the promoter elements TATA box and start site, which specify the start of transcription. The proximal promoter element, GC, binds the SP1 factor, which is present in most eukaryotic cells and stimulates transcription at low levels in most cells. Basal levels of expression are also regulated by the BLE (basal element) and ARE (AP factor response element) regions. These *cis*-elements bind the activator proteins 1, 2, and 4 (AP1, AP2, and AP4), which are present in various levels in different cell types and can be activated in response to extracellular growth signals. The BLE contains overlapping binding sites for the AP1 and AP4 factors, providing some degree of selectivity in how these factors stimulate transcription of *hMTIIA*

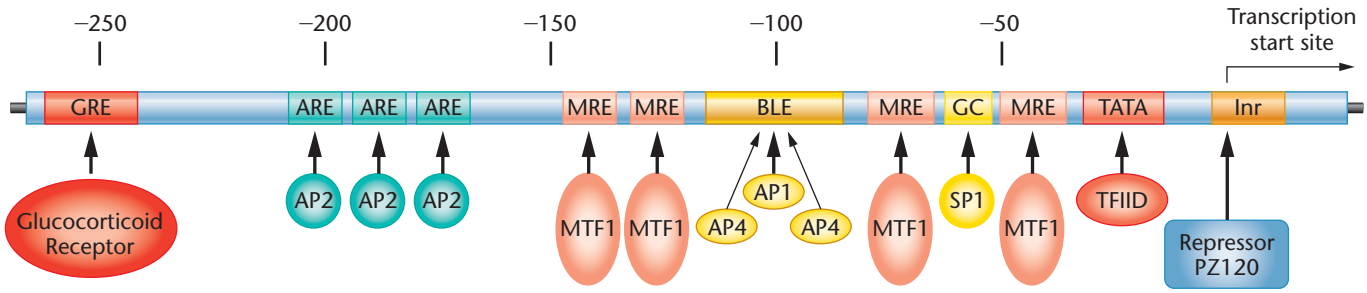


FIGURE 15–16 The human metallothionein IIA gene promoter and enhancer regions, containing multiple *cis*-acting regulatory sites. The transcription factors controlling both basal and induced levels of MTIIA transcription, and their binding sites, are indicated below the gene and are described in the text.

when bound to the BLE in different cell types. High levels of transcription are conferred by the presence of the enhancers MRE (metal response element) and GRE (glucocorticoid response element). The metal-inducible transcription factor (MTF1) binds to the MRE in response to the presence of heavy metals. The glucocorticoid receptor protein binds to the GRE, but only when the receptor protein is also bound to the glucocorticoid steroid hormone. The glucocorticoid receptor is normally located in the cytoplasm of the cell. However, when glucocorticoid hormone enters the cytoplasm, it binds to the receptor and causes a conformational change that allows the receptor to enter the nucleus, bind to the GRE, and stimulate *hMTIIA* gene transcription. In addition to induction, transcription of the *hMTIIA* gene can be repressed by the actions of the repressor protein PZ120, which binds over the transcription start region.

The presence of multiple regulatory elements and transcription factors that bind to them allows the *hMTIIA* gene to be transcriptionally induced or repressed in response to subtle changes in both extracellular and intracellular conditions.

ESSENTIAL POINT

Transcription factors influence transcription rates by binding to *cis*-acting regulatory sites within or adjacent to a gene promoter. ■

15.10 Activators and Repressors Interact with General Transcription Factors and Affect Chromatin Structure

We have now discussed the first steps in eukaryotic transcription regulation: first, chromatin must be remodeled and modified to allow transcription proteins to bind to their specific *cis*-acting sites; second, transcription factors bind to *cis*-acting sites and bring about positive and negative effects on the transcription initiation rate—often in response to extracellular signals or in tissue- or time-specific ways. The next question is, how do these *cis*-acting regulatory elements and their DNA-binding factors act to influence transcription initiation? To answer this question, we must first discuss how eukaryotic RNA polymerase II and its basal transcription factors assemble at promoters.

Formation of the Transcription Pre-Initiation Complex

A number of proteins called **general transcription factors** are needed to initiate both basal-level and enhanced levels of transcription. These proteins assemble at the promoter in a specific order, forming a transcriptional **pre-initiation complex** (**PIC**) that in turn provides a platform for RNA polymerase to recognize and bind to the promoter. We will restrict our discussion of PIC formation to focused promoters with TATA boxes—the type of promoter for which the most information is available.

The general transcription factors and their interactions with the core promoter and RNA polymerase II are outlined in **Figure 15–17**. The first step in the formation of a PIC is the binding of TFIID to the TATA box of the core promoter. TFIID is a multi-subunit complex that contains **TBP** (*T*ATA *B*inding *P*rotein) and approximately 13 proteins called **TAFs** (*T*BP *A*ssociated *F*actors). As its name implies, TBP binds to the TATA box. In addition, a subset of TAFs binds to Inr elements, as well as DPEs and MTEs. TFIIA interacts with TFIID and assists the binding of TFIID to the core promoter. Once TFIID has made contact with the core promoter, TFIIB binds to BREs on one or both sides of the TATA box. Once TFIID and TFIIB have bound the core promoter, the other general transcription factors interact with RNA polymerase II and help recruit it to the promoter. The fully formed PIC mediates the unwinding of promoter DNA at the start site and the transition of RNA polymerase II from transcription initiation to elongation. On many promoters of higher eukaryotes, RNA polymerase II remains paused at about 50 bp downstream of the transcription start site, awaiting

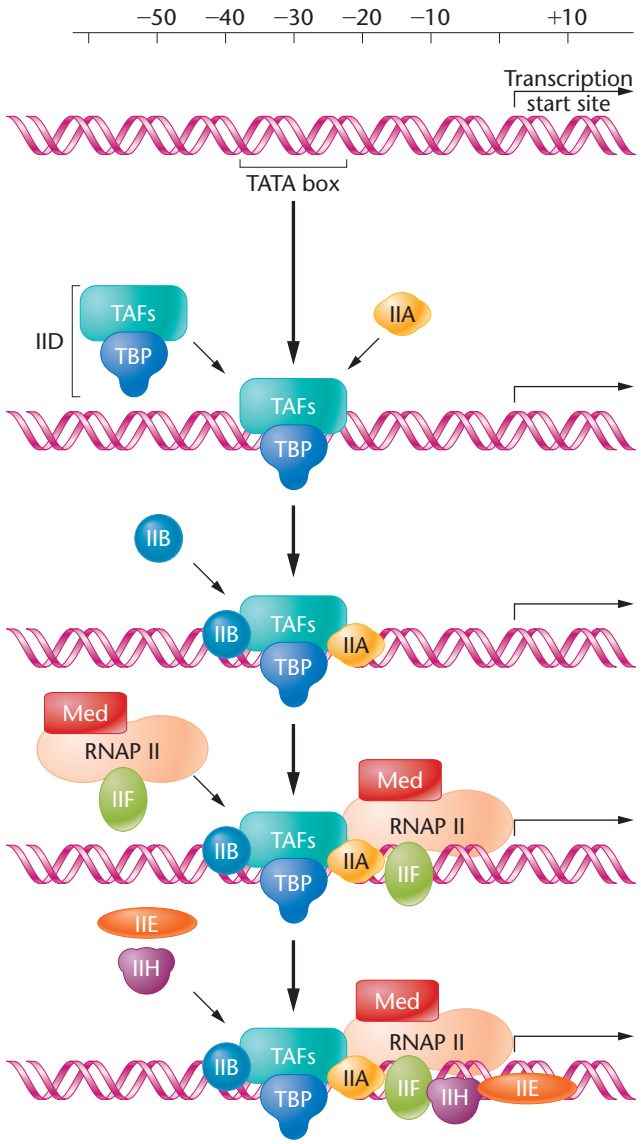


FIGURE 15–17 The assembly of general transcription factors required for the initiation of transcription by RNA polymerase II (RNAP II).

signals that release it into transcription elongation. In other gene promoters of higher eukaryotes and in all promoters in yeast, RNA polymerase II immediately leaves the promoter region and proceeds down the DNA template in an **elongation complex.** Several of the general transcription factors, specifically TFIID, TFIIE, TFIIH, and Mediator, remain on the core promoter to help set up the next PIC.

Mechanisms of Transcription Activation and Repression

Researchers have proposed several models to explain how transcription activators and repressors bring about changes to RNA polymerase II transcription. In most cases, these models involve the formation of DNA loops that bring distant enhancer or silencer elements into close physical contact with the promoter regions of genes that they regulate.

In one model of transcription activation and repression, DNA looping may deliver activators, repressors, and general transcription factors to the vicinity of promoters that must be activated or repressed. In this *recruitment model*, enhancer and silencer elements act as donors that increase the concentrations of important regulatory proteins at gene promoters. By enhancing the rate of PIC assembly or stability, or by accelerating the release of RNA polymerase II from a promoter, transcription activators bound at enhancers may stimulate the rate of transcription initiation. In order to make contact with promoter-bound factors, activators are thought to interact with other proteins called **coactivators** that form a complex known as an **enhanceosome** (Figure 15–18). Enhanceosomes may directly contact the PIC through subunits of the mediator and TFIID. In a similar way, repressors bound at silencer elements may decrease the rate of PIC assembly and the release of RNA polymerase II.

In a second model, DNA looping may result in *chromatin alterations* that either stimulate or repress transcription of target genes. Chromatin remodeling complexes or chromatin modifiers, once delivered to the vicinity of a promoter, may open or close the promoter to interactions with general transcription factors and RNA polymerase II, or may inhibit the release of paused RNA polymerase II from pre-initiation complexes.

A third model of transcription activation and repression states that enhancer or repressor looping may relocate a target gene to a nuclear region that is favorable or inhibitory to transcription. This *nuclear relocation* model would be consistent with the presence of transcription factories—regions of the nucleus that contain concentrations of RNA polymerase II and transcription regulatory factors.

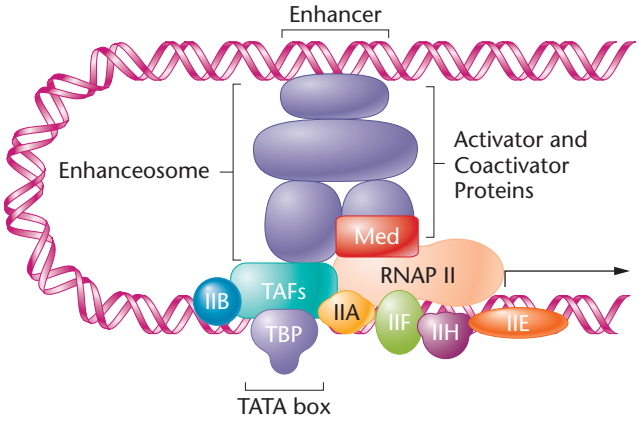


FIGURE 15–18 Formation of DNA loops allows factors that bind to an enhancer (or silencer) at a distance from a promoter to interact with general transcription factors and RNA polymerase II (RNAP II) in the pre-initiation complex and to regulate the level of transcription.

ESSENTIAL POINT

Transcription factors act by enhancing or repressing the association of general transcription factors at the promoter. They may also assist in chromatin remodeling and the relocation of a target gene to specific nuclear sites. ■

NOW SOLVE THIS

15–4 The hormone estrogen converts the estrogen receptor (ER) protein from an inactive molecule to an active transcription factor. The ER binds to *cis*-acting sites that act as enhancers, located near the promoters of a number of genes. In some tissues, the presence of estrogen appears to activate transcription of ER-target genes, whereas in other tissues, it appears to repress transcription of those same genes. Offer an explanation as to how this may occur.

■ **HINT:** *This problem involves an understanding of how transcription enhancers and repressors work. The key to its solution is to consider the many ways that* trans*-acting factors can interact at enhancers to bring about changes in transcription initiation.*

15.11 Posttranscriptional Gene Regulation Occurs at Many Steps from RNA Processing to Protein Modification

Although transcriptional control is a major type of gene regulation in eukaryotes, **posttranscriptional regulation** plays an equal, and in some cases more significant, role. Modification of eukaryotic nuclear RNA transcripts prior to translation includes the removal of noncoding introns, the precise splicing together of the remaining exons, and the addition of a cap at the mRNA's 5′ end and a poly-A tail at its 3′-end. The messenger RNA is then exported to the cytoplasm, where it is translated and degraded. Each of the mRNA processing steps can be regulated to control the quantity of functional mRNA available for synthesis of a protein product. In addition, the rate of translation, as well as the stability and activity of protein products, can be regulated. We will examine several mechanisms of posttranscriptional gene regulation that are especially important in eukaryotes—alternative splicing, mRNA stability, translation, and protein stability.

Alternative Splicing of mRNA

Alternative splicing can generate different forms of mRNA from identical pre-mRNA molecules, so that expression of one gene can give rise to a number of proteins with similar or different functions. Changes in splicing patterns can have many different effects on the translated protein. For example, they can alter the protein's enzymatic activity, receptor-binding capacity, or protein localization in the cell.

Figure 15–19 presents an example of alternative splicing of the pre-mRNA transcribed from the **calcitonin/calcitonin gene-related peptide gene** (***CT/CGRP* gene**). In thyroid cells, the *CT/CGRP* primary transcript is spliced in such a way that the mature mRNA contains the first four exons only. In these cells, the exon 4 polyadenylation signal is used to process the mRNA and add the poly-A tail. This mRNA is translated into the calcitonin peptide, a 32-amino acid peptide hormone that is involved in regulating calcium. In the brain and peripheral nervous system, the *CT/CGRP* primary transcript is spliced to include exons 5 and 6, but not exon 4. In these cells, the exon 6 polyadenylation site is recognized. The *CGRP* mRNA encodes a 37-amino acid peptide with hormonal activities in a wide range of tissues. Through alternative splicing, two peptide hormones with different structures, locations, and functions are synthesized from the same gene. Even more complex alternative splicing patterns occur in some genes, such as the example in **Figure 15–20**.

Alternative splicing increases the number of proteins that can be made from each gene. As a result, the number of

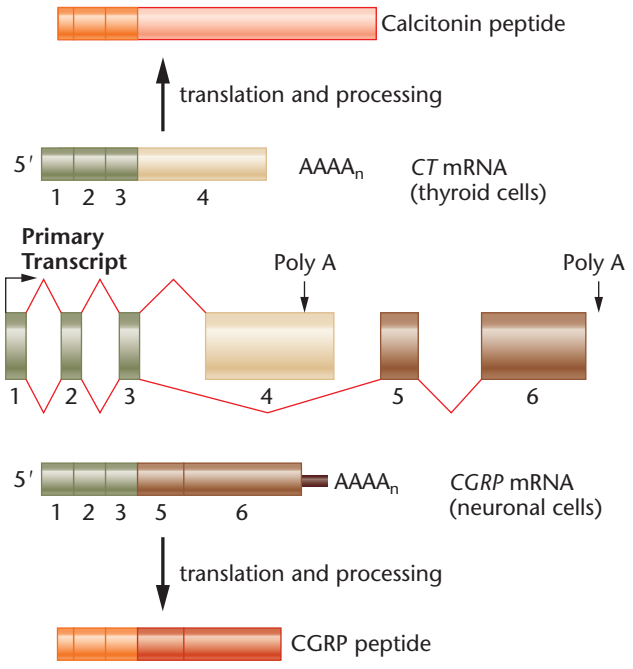
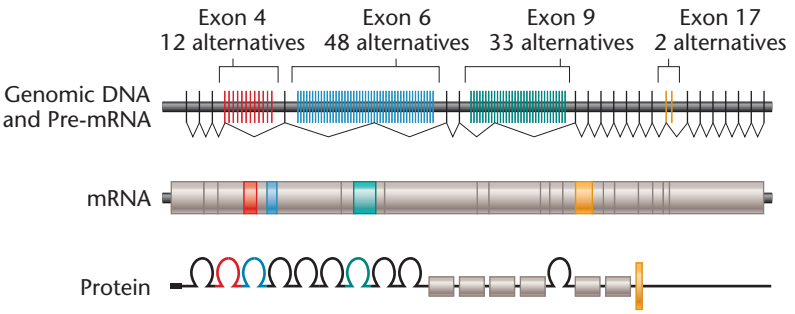


FIGURE 15–19 Alternative splicing of the *CT/CGRP* gene transcript. The primary transcript, which is shown in the middle of the diagram, contains six exons. The primary transcript can be spliced into two different mRNAs, both containing the first three exons but differing in their final exons. The *CT* mRNA contains exon 4, with polyadenylation occurring at the end of the fourth exon. The *CGRP* mRNA contains exons 5 and 6, and polyadenylation occurs at the end of exon 6. The *CT* mRNA is produced in thyroid cells. After translation, the resulting protein is processed into the calcitonin peptide. In contrast, the *CGRP* mRNA is produced in neuronal cells, and after translation, its protein product is processed into the CGRP peptide.



© 2000 Elsevier

FIGURE 15–20 Alternative splicing of the *Dscam* gene mRNA. The *Dscam* gene encodes a protein that guides axon growth during development. Each mRNA will contain one of the 12 possible exons for exon 4 (red), one of the 48 possible exons for exon 6 (blue), one of the 33 possible exons for exon 9 (green), and one of the 2 possible exons for exon 17 (yellow). Counting all possible combinations of these exons, the *Dscam* gene could encode 38,016 different versions of the DSCAM protein.

proteins that an organism can make—its **proteome**—is not the same as the number of genes in the genome, and protein diversity can exceed gene number by an order of magnitude. Alternative splicing is found in all metazoans but is especially common in vertebrates, including humans. It has been estimated that at least two-thirds of protein-coding genes in the human genome can undergo alternative splicing. Thus, humans can produce several hundred thousand different proteins (or perhaps more) from the approximately 20,000 genes in the haploid genome.

Mutations that affect regulation of splicing contribute to several genetic disorders. One of these disorders, **myotonic dystrophy** (**DM**), provides an example of how defects in alternative RNA splicing can lead to a wide range of symptoms. Myotonic dystrophy is the most common form of adult muscular dystrophy, affecting 1 in 8000 individuals. It is an autosomal dominant disorder that occurs in two forms—DM1 and DM2. Both of these diseases show a wide range of symptoms, including muscle wasting, **myotonia** (difficulty relaxing muscles), insulin resistance, behavior and cognitive defects, and cardiac muscle problems.

DM1 is caused by the expansion of the trinucleotide repeat CTG in the 3′-untranslated region of the ***DMPK* gene.** In unaffected individuals, the *DMPK* gene contains between 5 and 35 copies of the CTG repeat sequence, whereas in DM1 patients, the gene contains between 150 and 2000 copies. The severity of the symptoms is directly related to the number of copies of the repeat sequence. DM2 is caused by an expansion of the repeat sequence CCTG within the first intron of the ***ZNF9* gene.** Affected individuals may have up to 11,000 copies of the repeat sequence in the *ZNF9* intron. In DM2, the severity of symptoms is not related to the number of repeats.

Recently, scientists have discovered that DM1 and DM2 are caused not by changes in the protein products of the *DMPK* or *ZNF9* genes, but by the toxic effects of their repeat-containing RNAs. These RNAs accumulate and form inclusions within the nucleus. In the case of *ZNF9*, only the CCUG sequence repeat itself accumulates in the nucleus, as the remainder of the intron is degraded after splicing of the mRNA. It appears that the accumulated RNA fragments bind to, and sequester, proteins that would normally be involved in regulating the alternative splicing patterns of a large number of other RNAs. These RNAs include those whose products are required for the proper functioning of muscle and neural tissue. So far, scientists have discovered over 20 genes that are inappropriately spliced in the muscle, heart, and brain of DM1 patients. Often, the fetal splicing patterns occur in DM1 and DM2 patients, and the normal transitions to adult splicing patterns are lacking. Such defects in the regulation of RNA splicing are known as **spliceopathies.**

Control of mRNA Stability

The **steady-state level** of an mRNA is its amount in the cell as determined by a combination of the rate at which the gene is transcribed and the rate at which the mRNA is degraded. The steady-state level determines the amount of mRNA that is available for translation. All mRNA molecules are degraded at some point after their synthesis, but the lifetime of an mRNA, defined in terms of its **half-life,** or $t_{1/2}$, can vary widely between different mRNAs and can be regulated in response to the needs of the cell. Some mRNAs are degraded within minutes after their synthesis, whereas others can remain stable for hours, months, or even years (in the case of mRNAs stored in oocytes).

Regulation of mRNA stability is often linked with the process of translation. Several observations demonstrate this link between translation and mRNA stability. First, most mRNA molecules become stable in cells that are treated with translation inhibitors. Second, the presence of premature stop codons in the body of an mRNA, as well as premature translation termination, causes rapid degradation of mRNAs. Third, many of the ribonucleases and mRNA-binding proteins that affect mRNA stability associate with ribosomes.

Another way that an mRNA's half-life can be altered is through specific RNA sequence elements that recruit degrading or stabilizing complexes. One well-studied mRNA stability element is the adenosine-uracil rich element (ARE)—a stretch of ribonucleotides that consist of A and U ribonucleotides. These AU-rich elements are usually located in the 3′-untranslated regions of mRNAs that have short, regulated half-lives. These ARE-containing mRNAs encode proteins that are involved in cell growth or transcription control and need to be rapidly modulated in abundance. In cells that are not growing or require low levels of gene expression, specific complexes bind to the ARE elements of these mRNA molecules, bringing about shortening of the poly-A tail and rapid mRNA degradation. It is estimated that approximately 10 percent of mammalian mRNAs contain these instability elements.

Translational and Posttranslational Controls

In some cases, the translation of an mRNA can be regulated by the extent of the cell's requirement for the gene product. In other cases, the stability of a protein can be modulated or the protein can be modified after translation to change its structure and affect its activity.

An example of regulated stability and modification is that of the **p53 protein.** The p53 protein is essential to protect normal cells from the effects of DNA damage and other stresses. It is a transcription factor that increases the transcription of a number of genes whose products are involved in cell-cycle arrest, DNA repair, and programmed cell death. Under normal conditions, the levels of p53 protein are extremely low in cells, and the p53 that is present is inactive. When cells suffer DNA damage or metabolic stress, the amount of p53 protein increases dramatically, and p53 becomes an active transcription factor.

The changes in the abundance and activity of p53 are due to a combination of increased protein stability and modifications to the protein. In unstressed cells, p53 is bound by another protein called **Mdm2.** The Mdm2 protein binds to the p53 protein, blocking its ability to induce transcription. In addition, Mdm2 adds ubiquitin residues onto the p53 protein. **Ubiquitin** is a small protein that tags other proteins for degradation by proteolytic enzymes. The presence of ubiquitin on p53 results in p53 degradation. When cells are stressed, Mdm2 and p53 become modified by phosphorylation and acetylation, resulting in the release of Mdm2 from p53. As a consequence, p53 proteins become stable, the levels of p53 increase, and the protein is able to act as a transcription factor. An added level of control is that p53 is a transcription factor that induces the transcription of the *Mdm2* gene. Hence, the presence of active p53 triggers a negative feedback loop that creates more Mdm2 protein, which rapidly returns p53 to its rare and inactive state.

ESSENTIAL POINT

Posttranscriptional gene regulation can involve alternative splicing of nascent RNA, changes in mRNA stability, translational control, and posttranslational modifications. These mechanisms may alter the type, quantity, or activity of a gene's protein product. ■

15.12 RNA-Induced Gene Silencing Controls Gene Expression in Several Ways

In the last decade, the discovery that **small noncoding RNA** (**sncRNA**) molecules control gene expression has given rise to a new field of research. First discovered in plants, sncRNAs are now known to regulate gene expression in the cytoplasm of plants, animals, and fungi by repressing translation and triggering the degradation of mRNAs. This form of sequence-specific posttranscriptional regulation is known as **RNA interference** (**RNAi**). More recently, sncRNAs have been shown to act in the nucleus to alter chromatin structure and bring about repression of transcription. Together, these phenomena are known as **RNA-induced gene silencing.** Later in the text (see Special Topic Chapter 2—Emerging Roles of RNA), we present a comprehensive description of gene regulation by various types of RNA molecules, including sncRNAs.

Recent studies are demonstrating that RNA-induced gene-silencing mechanisms operate during normal development and control the expression of batteries of genes involved in tissue-specific cellular differentiation. In addition, scientists have discovered that abnormal activities of sncRNAs contribute to the occurrence of cancers, diabetic complications, and heart disease.

Geneticists are applying RNAi as a powerful research tool. RNAi technology allows investigators to create specific single-gene defects without having to induce inherited gene mutations. RNAi-mediated gene silencing is relatively specific and inexpensive, and it allows scientists to rapidly analyze gene function.

In addition to its use in laboratory research, RNAi is being developed as a potential pharmaceutical agent. In theory, any disease caused by overexpression of a specific gene, or even normal expression of an abnormal gene product, could be attacked by therapeutic RNAi. Viral infections are obvious targets, and scientists have had promising results using RNAi in tissue cultures to reduce the severity of infection by several types of viruses such as HIV, influenza, and polio. In animal models, siRNA molecules have successfully treated virus infections, eye diseases, cancers, and inflammatory bowel disease.

New as it is, the science of RNAi holds powerful promise for molecular medicine. The uses of RNAi in therapeutics are also discussed in the Genetics, Technology, and Society feature in Chapter 12 on p. 250.

GENETICS, TECHNOLOGY, AND SOCIETY

Quorum Sensing: Social Networking in the Bacterial World

For decades, scientists regarded bacteria as independent single-celled organisms, incapable of cell-to-cell communication. However, recent research is revealing that many bacteria can regulate gene expression and coordinate group behavior through a form of communication termed *quorum sensing.* Through this process, bacteria send and receive chemical signals called autoinducers that relay information about population size. When the population size reaches a "quorum," defined in the business world as the minimum number of members of an organization that must be present to conduct business, the autoinducers regulate gene expression in a way that benefits the group as a whole. Quorum sensing has been described in more than 70 species of bacteria, and its uses range from controlling bioluminescence in marine bacteria to regulating the expression of virulence factors in pathogenic bacteria. Our understanding of quorum sensing has altered our perceptions of prokaryotic gene regulation and is leading to the development of practical applications, including new antibiotic drugs.

Quorum sensing was discovered in the 1960s, during research on the bioluminescent bacterium *Vibrio fischeri,* which lives in a symbiotic relationship with the Hawaiian bobtail squid, *Euprymna scolopes.* While hunting for food at night, the squid uses light emitted by the *V. fischeri* present in its light organ to illuminate the ocean floor and to counter the shadows created by moonlight that normally act as a beacon for the squid's predators. In return, the bacteria gain a protected, nutrient-rich environment in the squid's light organ. During the day, the bacteria do not glow as a result of the squid's ability to reduce the concentration of bacteria in its light organ, which in turn prevents expression of the bacterial luciferase (*lux*) operon.

What turns the bacteria's *lux* genes on in response to high cell density and off in response to low cell density? In *V. fischeri*, the responsible autoinducer is a secreted homoserine lactone molecule. At a critical population size, these molecules accumulate, are taken up by bacteria within the population, and regulate the *lux* operon by binding directly to transcription factors that stimulate *lux* gene expression.

Since the discovery of quorum sensing in *Vibrio fischeri*, scientists have identified similar microbial communication systems in other bacteria, including significant human pathogens such as pseudomonal, staphylococcal, and streptococcal species. The expression of as many as 15 percent of bacterial genes may be regulated by quorum sensing.

Quorum sensing molecules may also mediate communication among members of different species. In 1994, Bonnie Bassler and her colleagues at Princeton University discovered an autoinducer molecule in the marine bacterium *Vibrio harveyi* that was also present in many diverse types of bacteria. This molecule, autoinducer-2 (AI-2), has the potential to mediate "quorum-sensing cross talk" between species and thus serve as a universal language for bacterial communication. Because the accumulation of AI-2 is proportional to cell number, and because the structure of AI-2 may vary slightly between different species, the current hypothesis is that AI-2 can transmit information about both the cell density and species composition of a bacterial community.

Pathogenic bacteria use quorum sensing to regulate the expression of genes whose products help these bacteria invade a host and avoid immune system detection. For example, *Vibrio cholerae*, the causative agent of cholera, uses AI-2 and an additional species-specific autoinducer to activate the genes controlling the production of cholera toxin. *Pseudomonas aeruginosa*, the Gram-negative bacterium that often affects cystic fibrosis patients, uses quorum sensing to regulate the production of elastase, a protease that disrupts the respiratory epithelium and interferes with ciliary function. *P. aeruginosa* also uses autoinducers to control the production of biofilms, tough protective shells that resist host defenses and make treatment with antibiotics nearly impossible. Other bacteria determine cell density through quorum sensing to delay the production of toxic substances until the colony is large enough to overpower the host's immune system and establish an infection. Because many bacteria rely on quorum sensing to regulate disease-causing genes, therapeutics that block quorum sensing may help combat infections. Research into these potential therapies is now in progress, and several are now approaching the clinical trial phase. Thus, what began as a fascinating observation in the glowing squid has launched an exciting era of research in bacterial genetics that may one day prove of great clinical significance.

Your Turn

Take time, individually or in groups, to answer the following questions. Investigate the references and links to help you understand the mechanisms and potential uses of quorum sensing in bacteria.

1. Inhibitors of quorum sensing molecules have potential as antibacterial agents. What are some ways in which quorum sensing inhibitors could work to combat bacterial infections? Have any of these therapeutics reached clinical trials?

A recent review of quorum sensing therapeutics can be found in Njoroge, J. and Sperandio, V. 2009. Jamming bacterial communication: New approaches for the treatment of infectious diseases. *EMBO Mol. Med.* 1(4): 201–210.

2. Regulation of bacterial gene expression by autoinducer molecules involves several different mechanisms. Describe these mechanisms and how each could be used as a target for the control of bacterial infections.

The mechanisms by which autoinducers regulate gene expression are summarized in Asad, S. and Opal, S.M. 2008. Bench-to-bedside review: Quorum sensing and the role of cell-to-cell communication during invasive bacterial infection. *Critical Care* 12: 236–247.

3. Quorum sensing systems are also capable of detecting and responding

to chemical signals given off by host cells. Explain how this works and how this might benefit pathogenic bacteria.

A review article dealing with interkingdom communication and quorum sensing can be found at Wagner, V.E. et al. 2006. Quorum sensing: dynamic response of *Pseudomonas aeruginosa* to external signals. *Trends Microbiol.* 14(2): 55–58.

CASE STUDY | A mysterious muscular dystrophy

A man in his early 30s suddenly developed weakness in his hands and neck, followed a few weeks later by burning muscle pain—all symptoms of late-onset muscular dystrophy. His internist ordered genetic tests to determine whether he had one of the inherited muscular dystrophies, focusing on Becker muscular dystrophy, myotonic dystrophy Type I, and myotonic dystrophy Type II. These tests were designed to detect mutations in the related *dystrophin, DMPK,* and *ZNF9* genes. The testing ruled out Becker muscular dystrophy. While awaiting the results of the *DMPK* and *ZNF9* gene tests, the internist explained that the possible mutations were due to expanded tri- and tetranucleotide repeats, but not in the protein-coding portion of the genes. She went on to say that the resulting disorders were due not to changes in the encoded proteins, which appear to be normal, but instead to altered RNA splicing patterns, whereby the RNA splicing remnants containing the nucleotide repeats disrupt normal splicing of the transcripts of other genes. This discussion raises several interesting questions about the diagnosis and genetic basis of the disorders.

1. What is alternative splicing, where does it occur, and how could disrupting it affect the expression of the affected gene(s)?
2. What role might the expanded tri- and tetranucleotide repeats play in the altered splicing?
3. How does this contrast with other types of muscular dystrophy, such as Becker muscular dystrophy and Duchenne muscular dystrophy?

INSIGHTS AND SOLUTIONS

1. A theoretical operon (*theo*) in *E. coli* contains several structural genes encoding enzymes that are involved sequentially in the biosynthesis of an amino acid. Unlike the *lac* operon, in which the repressor gene is separate from the operon, the gene encoding the regulator molecule is contained within the *theo* operon. When the end product (the amino acid) is present, it combines with the regulator molecule, and this complex binds to the operator, repressing the operon. In the absence of the amino acid, the regulatory molecule fails to bind to the operator, and transcription proceeds.

   Characterize this operon, then consider the following mutations, as well as the situation in which the wild-type gene is present along with the mutant gene in partially diploid cells (F′):

   (a) Mutation in the operator region.

   (b) Mutation in the promoter region.

   (c) Mutation in the regulator gene.

   In each case, will the operon be active or inactive in transcription, assuming that the mutation affects the regulation of the *theo* operon? Compare each response with the equivalent situation of the *lac* operon.

   Solution: The *theo* operon is repressible and under negative control. When there is no amino acid present in the medium (or the environment), the product of the regulatory gene cannot bind to the operator region, and transcription proceeds under the direction of RNA polymerase. The enzymes necessary for synthesis of the amino acid are produced, as is the regulator molecule. If the amino acid *is* present, or is present after sufficient synthesis occurs, the amino acid binds to the regulator, forming a complex that interacts with the operator region, causing repression of transcription of the genes within the operon.

   The *theo* operon is similar to the tryptophan system, except that the regulator gene is within the operon rather than separate from it. Therefore, in the *theo* operon, the regulator gene is itself regulated by the presence or absence of the amino acid.

   (a) As in the *lac* operon, a mutation in the *theo* operator region inhibits binding with the repressor complex, and transcription occurs constitutively. The presence of an F′ plasmid bearing the wild-type allele would have no effect, since it is not adjacent to the structural genes.

   (b) A mutation in the *theo* promoter region would no doubt inhibit binding to RNA polymerase and therefore inhibit transcription. This would also happen in the *lac* operon. A wild-type allele present in an F′ plasmid would have no effect.

   (c) A mutation in the *theo* regulator gene, as in the *lac* system, may inhibit either its binding to the repressor or its binding to the operator gene. In both cases, transcription will be constitutive because the *theo* system is repressible. Both cases result in the failure of the regulator to bind to the operator, allowing transcription to proceed. In the *lac* system, failure to bind the corepressor lactose would permanently repress the system. The addition of a wild-type allele would restore repressibility, provided that this gene was transcribed constitutively.

2. Regulatory sites for eukaryotic genes are usually located within a few hundred nucleotides of the transcription start site, but they can be located up to several kilobases away. DNA sequence-specific binding assays have been used to detect and isolate protein factors present at low concentrations in nuclear extracts. In these experiments, short DNA molecules containing

(continued)

Insights and Solutions—continued

DNA-binding sequences are attached to material that is packed into a glass column, and nuclear extracts are passed over the column. The idea is that if proteins that specifically bind to the DNA sequence are present in the nuclear extract, they will bind to the DNA, and they can be recovered from the column after all other nonbinding material has been washed away. Once a DNA-binding protein has been isolated and identified, the problem is to devise a general method for screening cloned libraries for the genes encoding the DNA-binding factors. Determining the amino acid sequence of the protein and constructing synthetic oligonucleotide probes are time consuming and useful for only one factor at a time. Knowing the strong affinity for binding between the protein and its DNA-recognition sequence, how would you screen for genes encoding binding factors?

Solution: Several general strategies have been developed, and one of the most promising was devised by Steve McKnight's laboratory at the Fred Hutchinson Cancer Center. The cDNA isolated from cells expressing the binding factor is cloned into the lambda vector, gt11. Plaques of this library, containing proteins derived from expression of cDNA inserts, are adsorbed onto nitrocellulose filters and probed with double-stranded radioactive DNA corresponding to the binding site. If a fusion protein corresponding to the binding factor is present, it will bind to the DNA probe. After the unbound probe is washed off, the filter is subjected to autoradiography and the plaques corresponding to the DNA-binding proteins can be identified. An added advantage of this strategy is filter recycling by washing the bound DNA from the filters prior to their reuse. Such an ingenious procedure is similar to the colony-hybridization and plaque-hybridization procedures described for library screening in Chapter 17, and it provides a general method for isolating genes encoding DNA-binding factors.

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we have focused on how prokaryotic and eukaryotic organisms regulate the expression of genetic information. In particular, we discussed both transcriptional and posttranscriptional gene regulation. Based on your knowledge of these topics, answer several fundamental questions:
   (a) How do we know that bacteria regulate the expression of certain genes in response to the environment?
   (b) How do we know that bacterial gene clusters are often coordinately regulated by a regulatory region that must be located adjacent to the cluster?
   (c) What led researchers to conclude that a *trans*-acting repressor molecule regulates the *lac* operon?
   (d) How do we know that promoters and enhancers regulate transcription of eukaryotic genes?
   (e) How do we know that DNA methylation plays a role in the regulation of eukaryotic gene expression?

CONCEPT QUESTION

2. Review the Chapter Concepts list on p. 296. These all relate to the regulation of gene expression in prokaryotes and eukaryotes. Write a brief essay that discusses why you think gene-regulatory systems evolved in bacteria, and why genes related to common functions are found together in operons. ■

3. Describe which enzymes are required for lactose and tryptophan metabolism in bacteria when lactose and tryptophan, respectively, are (a) present and (b) absent.
4. Contrast positive versus negative regulation of gene expression. Describe the role of the repressor in an inducible system and in a repressible system.
5. Both attenuation and riboswitches rely on changes in the secondary structure of the leader regions of mRNA to regulate gene expression. Compare and contrast the specific mechanisms in these two types of regulation.
6. For the *lac* genotypes shown in the accompanying table, predict whether the structural gene (*Z*) is constitutive, permanently repressed, or inducible in the presence of lactose.

Genotype	Constitutive	Repressed	Inducible
$I^+O^+Z^+$			X
$I^-O^+Z^+$			
$I^+O^CZ^+$			
$I^-O^+Z^+/F'I^+$			
$I^+O^CZ^+/F'O^C$			
$I^SO^+Z^+$			
$I^SO^+Z^+/F'I^+$			

7. For the genotypes and conditions (lactose present or absent) shown in the accompanying table, predict whether functional enzymes, nonfunctional enzymes, or no enzymes are made.

Genotype	Condition	Functional Enzyme Made	Nonfunctional Enzyme Made	No Enzyme Made
$I^+O^+Z^+$	No lactose			X
$I^+O^CZ^+$	Lactose			
$I^-O^+Z^-$	No lactose			
$I^-O^+Z^-$	Lactose			
$I^-O^+Z^+/F'I^+$	No lactose			
$I^+O^CZ^+/F'O^+$	Lactose			
$I^+O^+Z^-/F'I^+O^+Z^+$	Lactose			
$I^-O^+Z^-/F'I^+O^+Z^+$	No lactose			
$I^SO^+Z^+/F'O^+$	No lactose			
$I^+O^CZ^+/F'O^+Z^+$	Lactose			

8. The locations of numerous *lacI*$^-$ and *lacI*^{S} mutations have been determined within the DNA sequence of the *lacI* gene. Among these, *lacI*$^-$ mutations were found to occur in the 5′-upstream region of the gene, while *lacI*^{S} mutations were found to occur farther downstream in the gene. Are the locations of the two types of mutations within the gene consistent with what is known about the function of the repressor that is the product of the *lacI* gene?

9. Explain why catabolite repression is used in regulating the *lac* operon and describe how it fine-tunes β-galactosidase synthesis.
10. Describe experiments that would confirm whether or not two transcription regulatory molecules act through the mechanism of cooperative binding.
11. Predict the level of genetic activity of the *lac* operon as well as the status of the *lac* repressor and the CAP protein under the cellular conditions listed in the accompanying table.

	Lactose	Glucose
(a)	−	−
(b)	+	−
(c)	−	+
(d)	+	+

12. Predict the effect on the inducibility of the *lac* operon of a mutation that disrupts the function of (a) the *crp* gene, which encodes the CAP protein, and (b) the CAP-binding site within the promoter.
13. Describe the role of attenuation in the regulation of tryptophan biosynthesis.
14. Imagine that a new operon is discovered in a certain microorganism. The promoter sequence, the operator sequence, and the structural gene are represented by *P*, *O*, and *S*, respectively. Products of two other genes, *X* and *Y*, help in the regulation of this operon. It is induced by a molecule *A*. Its function is very similar to that of the *lac* operon, that is, the *S* gene product (enzyme) helps in metabolizing a molecule *X*. From the data provided in the accompanying table, explain how the *X* and *Y* gene products help in the regulation of this operon.

Genotype	In the presence of *A*	In the absence of *A*
$P^+O^+S^+X^+Y^+$	enzyme is produced	no enzyme is produced
$P^+O^+S^+X^-Y^+$	enzyme is produced	enzyme is produced
$P^+O^+S^+X^+Y^-$	no enzyme is produced	no enzyme is produced
$P^+O^+S^+X^-Y^-$	enzyme is produced	enzyme is produced
$P^+O^-S^+X^+Y^+$	enzyme is produced	enzyme is produced
$P^+O^+S^+X^-Y^-/F'X^+$	no enzyme is produced	no enzyme is produced
$P^+O^+S^+X^-Y^-/F'Y^+$	enzyme is produced	enzyme is produced
$P^+O^+S^+X^-Y^-/F'X^+Y^+$	enzyme is produced	no enzyme is produced

15. A bacterial operon is responsible for production of the biosynthetic enzymes needed to make the theoretical amino acid tisophane (tis). The operon is regulated by a separate gene, *R,* deletion of which causes the loss of enzyme synthesis. In the wild-type condition, when tis is present, no enzymes are made; in the absence of tis, the enzymes are made. Mutations in the operator gene (O^-) result in repression regardless of the presence of tis.

    Is the operon under positive or negative control? Propose a model for (a) repression of the genes in the presence of tis in wild-type cells and (b) the mutations.
16. A marine bacterium is isolated and is shown to contain an inducible operon whose genetic products metabolize oil when it is encountered in the environment. Investigation demonstrates that the operon is under positive control and that there is a *reg* gene whose product interacts with an operator region (*o*) to regulate the structural genes designated *sg*.

    In an attempt to understand how the operon functions, a constitutive mutant strain and several partial diploid strains were isolated and tested with the results shown here:

Host Chromosome	F′ Factor	Phenotype
wild type	none	inducible
wild type	*reg* gene from mutant strain	inducible
wild type	operon from mutant strain	constitutive
mutant strain	*reg* gene from wild type	constitutive

    Draw all possible conclusions about the mutation as well as the nature of regulation of the operon. Is the constitutive mutation in the *trans*-acting *reg* element or in the *cis*-acting *o* operator element?

17. What is the mechanism by which the chemical 5-azacytidine enhances gene expression?
18. List and define the levels of eukaryotic gene regulation discussed in this chapter.
19. What are the subcategories within eukaryotic promoters? How do enhancers and silencers differ from promoters?
20. What are transcription factors? What *cis*-acting elements do they bind?
21. Compare the control of gene regulation in eukaryotes and prokaryotes at the level of initiation of transcription. How do the regulatory mechanisms work? What are the similarities and differences in these two types of organisms in terms of the specific components of the regulatory mechanisms? Address how the differences or similarities relate to the biological context of the control of gene expression.
22. Many eukaryotic promoter regions contain CAAT boxes with consensus sequences CAAT or CCAAT approximately 70 to 80 bases upstream from the transcription start site. How might one determine the influence of CAAT boxes on the transcription rate of a given gene?
23. What is RNA-induced gene silencing in eukaryotes? How do sncRNAs affect gene regulation and how are they currently used in research and medicine?
24. Although it is customary to consider transcriptional regulation in eukaryotes as resulting from the positive or negative influence of different factors binding to DNA, a more complex picture is emerging. For instance, researchers have described the action of a transcriptional repressor (Net) that is regulated by nuclear export (Ducret et al., 1999. *Mol. and Cell. Biol.* 19: 7076–7087). Under neutral conditions, Net inhibits transcription of target genes; however, when phosphorylated, Net stimulates transcription of target genes. When stress conditions exist in a cell (for example, from ultraviolet light or heat shock), Net is excluded from the nucleus, and target genes are transcribed. Devise a model (using diagrams) that provides a consistent explanation of these three conditions.
25. DNA methylation is commonly associated with a reduction of transcription. The following data come from a study of the impact of the location and extent of DNA methylation on gene activity in human cells. A bacterial gene, luciferase, was cloned next to eukaryotic promoter fragments that were methylated to various degrees, *in vitro.* The chimeric plasmids were then introduced into tissue culture cells, and the luciferase activity was assayed. These data compare the degree of expression of luciferase with differences in the location of DNA methylation

(Irvine et al., 2002. *Mol. and Cell. Biol.* 22: 6689–6696). What general conclusions can be drawn from these data?

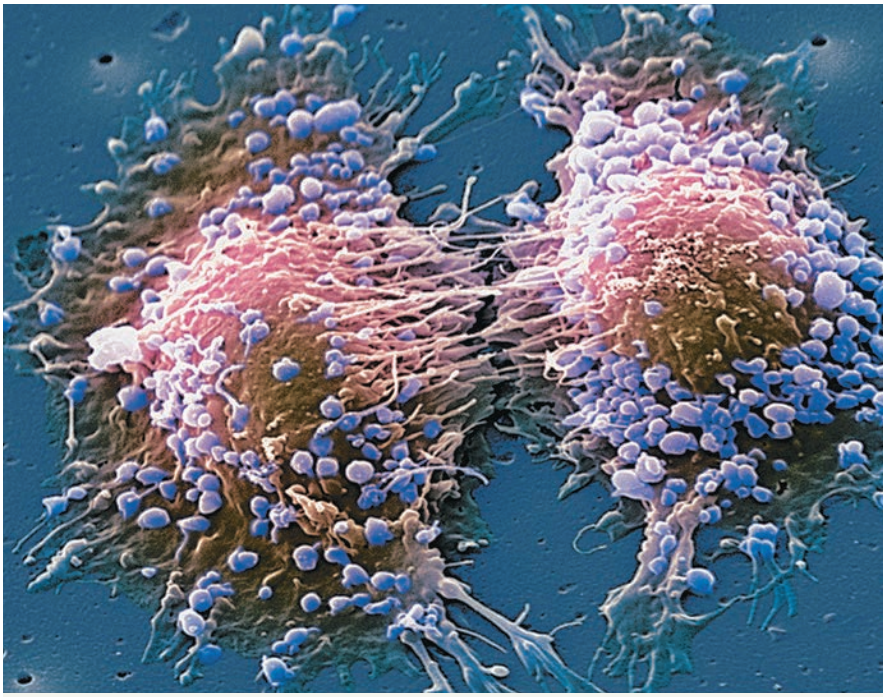
DNA Segment	Patch Size of Methylation (kb)	Number of Methylated CpGs	Relative Luciferase Expression
Outside transcription unit (0–7.6 kb away)	0.0	0	490X
	2.0	100	290X
	3.1	102	250X
	12.1	593	2X
Inside transcription unit	0.0	0	490X
	1.9	108	80X
	2.4	134	5X
	12.1	593	2X

26. The interphase nucleus appears to be a highly structured organelle with chromosome territories, interchromosomal compartments, and transcription factories. In cultured human cells, researchers have identified approximately 8000 transcription factories per cell, each containing an average of eight tightly associated RNA polymerase II molecules actively transcribing RNA. If each RNA polymerase II molecule is transcribing a different gene, how might such a transcription factory appear? Provide a simple diagram that shows eight different genes being transcribed in a transcription factory and include the promoters, structural genes, and nascent transcripts in your presentation.
27. It has been estimated that at least two-thirds of human genes produce alternatively spliced mRNA isoforms. In some cases, incorrectly spliced RNAs lead to human pathologies. Scientists have examined human cancer cells for splice-specific changes and found that many of the changes disrupt tumor-suppressor gene function (Xu and Lee, 2003. *Nucl. Acids Res.* 31: 5635–5643). In general, what would be the effects of splicing changes on these RNAs and the function of tumor-suppressor gene function? How might loss of splicing specificity be associated with cancer?

16 The Genetics of Cancer

CHAPTER CONCEPTS

- Cancer is characterized by genetic defects in fundamental aspects of cellular function, including DNA repair, chromatin modification, cell-cycle regulation, apoptosis, and signal transduction.
- Most cancer-causing mutations occur in somatic cells; only about 5 percent of cancers have a hereditary component.
- Mutations in cancer-related genes lead to abnormal proliferation and loss of control over how cells spread and invade surrounding tissues.
- The development of cancer is a multistep process requiring mutations in genes controlling many aspects of cell proliferation and metastasis.
- Cancer cells show high levels of genomic instability, leading to the accumulation of multiple mutations, some in cancer-related genes.
- DNA methylation and histone modifications play significant roles in the development of cancers.
- Mutations in proto-oncogenes and tumor-suppressor genes contribute to the development of cancers.
- Cancer-causing viruses and environmental agents contribute to the development of human cancers.



Colored scanning electron micrograph of two prostate cancer cells in the final stages of cell division (cytokinesis). The cells are still joined by strands of cytoplasm.

Cancer is the leading cause of death in Western countries. It strikes people of all ages, and one out of three people will experience a cancer diagnosis sometime in his or her lifetime. Each year, more than 1 million cases of cancer are diagnosed in the United States, and more than 500,000 people die from the disease.

Over the last 30 years, scientists have discovered that cancer is a genetic disease at the somatic cell level, characterized by the presence of gene products derived from mutated or abnormally expressed genes. The combined effects of numerous abnormal gene products lead to the uncontrolled growth and spread of cancer cells. Although some mutated cancer genes may be inherited, most are created within somatic cells that then divide and form tumors. Completion of the Human Genome Project and numerous large-scale rapid DNA sequencing studies have opened the door to a wealth of new information about the mutations that trigger a cell to become cancerous. This new understanding of cancer genetics is also leading to new gene-specific treatments, some of which are now entering clinical trials. Some scientists predict that gene-targeted therapies will replace chemotherapies within the next 25 years.

The goal of this chapter is to highlight our current understanding of the nature and causes of cancer. As we will see, cancer is a genetic disease that arises from the accumulation of mutations in genes controlling many basic aspects of cellular function. Please note that some of the topics discussed

in this chapter are explored in greater depth later in the text (see Special Topic Chapter 1—Epigenetics and Special Topic Chapter 4—Genomics and Personalized Medicine).

16.1 Cancer Is a Genetic Disease at the Level of Somatic Cells

Perhaps the most significant development in understanding the causes of cancer is the realization that cancer is a genetic disease. Genomic alterations that are associated with cancer range from single-nucleotide substitutions to large-scale chromosome rearrangements, amplifications, and deletions (**Figure 16–1**). However, unlike other genetic diseases, cancer is caused by mutations that arise predominantly in somatic cells. Only about 5 percent of cancers are associated with germ-line mutations that increase a person's susceptibility to certain types of cancer. Another important difference between cancers and other genetic diseases is that cancers rarely arise from a single mutation in a single gene, but from the accumulation of many mutations. The mutations that lead to cancer affect multiple cellular functions, including repair of DNA damage, cell division, apoptosis, cellular differentiation, migratory behavior, and cell–cell contact.

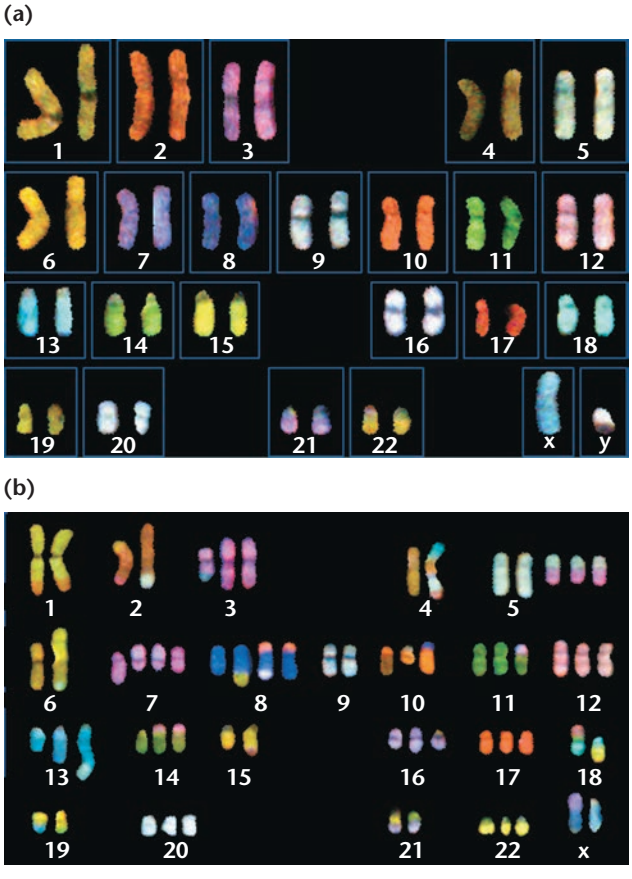


FIGURE 16–1 (a) Spectral karyotype of a normal cell. (b) Karyotype of a cancer cell showing translocations, deletions, and aneuploidy—characteristic features of cancer cells.

What Is Cancer?

Clinically, cancer is defined as a large number of complex diseases, up to a hundred, that behave differently depending on the cell types from which they originate and the types of genetic alterations that occur within each cancer type. Cancers vary in their ages of onset, growth rates, invasiveness, prognoses, and responsiveness to treatments. However, at the molecular level, all cancers exhibit common characteristics that unite them as a family.

All cancer cells share two fundamental properties: (1) abnormal cell growth and division (**proliferation**), and (2) defects in the normal restraints that keep cells from spreading and colonizing other parts of the body (**metastasis**). In normal cells, these functions are tightly controlled by genes that are expressed appropriately in time and place. In cancer cells, these genes are either mutated or are expressed inappropriately.

It is this combination of uncontrolled cell proliferation and metastatic spread that makes cancer cells dangerous. When a cell simply loses genetic control over cell growth, it may grow into a multicellular mass, a **benign tumor.** Such a tumor can often be removed by surgery and may cause no serious harm. However, if cells in the tumor also have the ability to break loose, enter the bloodstream, invade other tissues, and form secondary tumors (**metastases**), they become malignant. **Malignant tumors** are often difficult to treat and may become life threatening. As we will see later in the chapter, there are multiple steps and genetic mutations that convert a benign tumor into a dangerous malignant tumor.

ESSENTIAL POINT

Cancer cells show two fundamental properties: abnormal cell proliferation and a propensity to spread and invade other parts of the body (metastasis). ■

The Clonal Origin of Cancer Cells

Although malignant tumors may contain billions of cells, and may invade and grow in numerous parts of the body, all cancer cells in the primary and secondary tumors are clonal, meaning that they originated from a common ancestral cell that accumulated specific cancer-causing mutations. This is an important concept in understanding the molecular causes of cancer and has implications for its diagnosis.

Numerous data support the concept of cancer clonality. For example, reciprocal chromosomal translocations are characteristic of many cancers, including leukemias

and lymphomas (two cancers involving white blood cells). Cancer cells from patients with **Burkitt lymphoma** show reciprocal translocations between chromosome 8 (with translocation breakpoints at or near the *c-myc* gene) and chromosomes 2, 14, or 22 (with translocation breakpoints at or near one of the immunoglobulin genes). Each Burkitt lymphoma patient exhibits unique breakpoints in his or her *c-myc* and immunoglobulin gene DNA sequences; however, all lymphoma cells within that patient contain identical translocation breakpoints. This demonstrates that all cancer cells in each case of Burkitt lymphoma arise from a single cell, and this cell passes on its genetic aberrations to its progeny.

Another demonstration that cancer cells are clonal is their pattern of X-chromosome inactivation. As explained earlier in the text (see Chapter 5), female humans are mosaic, with some cells containing an inactivated paternal X chromosome and other cells containing an inactivated maternal X chromosome. X-chromosome inactivation occurs early in development and takes place at random. All cancer cells within a tumor, both primary and metastatic, within one female individual, contain the same inactivated X chromosome. This supports the concept that all the cancer cells in that patient arose from a common ancestral cell.

ESSENTIAL POINT

Cancers are clonal, meaning that all cells within a tumor originate from a single cell that contained a number of mutations. ■

The Cancer Stem Cell Hypothesis

A concept that is related to the clonal origin of cancer cells is that of the cancer stem cell. Many scientists now believe that most of the cells within tumors do not proliferate. Those that do proliferate and give rise to all the cells within the tumor are known as **cancer stem cells.** Stem cells are undifferentiated cells that have the capacity for self-renewal—a process in which the stem cell divides unevenly, creating one daughter cell that goes on to differentiate into a mature cell type and one that remains a stem cell. The cancer stem cell hypothesis contrasts the random or stochastic model. This model predicts that every cell within a tumor has the potential to form a new tumor.

Although scientists still actively debate the existence of cancer stem cells, evidence is accumulating that cancer stem cells do exist, at least in some tumors. Cancer stem cells have been identified in leukemias as well as in solid tumors of the brain, breast, colon, ovary, pancreas, and prostate. It is still not clear what fraction of any tumor is composed of cancer stem cells. For example, human acute myeloid leukemias contain less than 1 cancer stem cell in 10,000. In contrast, some solid tumors may contain as many as 40 percent cancer stem cells.

Scientists are also not sure about the origins of cancer stem cells. It is possible that they may arise from normal adult stem cells within a tissue, or they may be created from more differentiated somatic cells that acquire properties similar to stem cells after accumulating numerous mutations and changes to chromatin structure.

Cancer as a Multistep Process, Requiring Multiple Mutations

Although we know that cancer is a genetic disease initiated by mutations that lead to uncontrolled cell proliferation and metastasis, a single mutation is not sufficient to transform a normal cell into a tumor-forming, malignant cell. If it were sufficient, then cancer would be far more prevalent than it is. In humans, mutations occur spontaneously at a rate of about 10^{-6} mutations per gene, per cell division, mainly due to the intrinsic error rates of DNA replication. Because there are approximately 10^{16} cell divisions in a human body during a lifetime, a person might suffer up to 10^{10} mutations per gene somewhere in the body, during his or her lifetime. However, only about one person in three will suffer from cancer.

The phenomenon of age-related cancer is another indication that cancer develops from the accumulation of several mutagenic events in a single cell. The incidence of most cancers rises exponentially with age. If a single mutation were sufficient to convert a normal cell to a malignant one, then cancer incidence would appear to be independent of age. The age-related incidence of cancer suggests that many independent mutations, occurring randomly and with a low probability, are necessary before a cell is transformed into a malignant cancer cell. Another indication that cancer is a multistep process is the delay that occurs between exposure to **carcinogens** (cancer-causing agents) and the appearance of the cancer. For example, an incubation period of five to eight years separated exposure of people to the radiation of the atomic explosions at Hiroshima and Nagasaki and the onset of leukemias.

The multistep nature of cancer development is supported by the observation that cancers often develop in progressive steps, beginning with mildly aberrant cells and progressing to cells that are increasingly tumorigenic and malignant.

Each step in **tumorigenesis** (the development of a malignant tumor) appears to be the result of two or more genetic alterations that release the cells progressively from the controls that normally operate on proliferation and malignancy. This observation suggests that the progressive genetic alterations that create a cancer cell confer selective advantages to the cell and are propagated through cell divisions during the creation of tumors.

The progressive nature of cancer is illustrated by the development of colorectal cancer. Colorectal cancers are

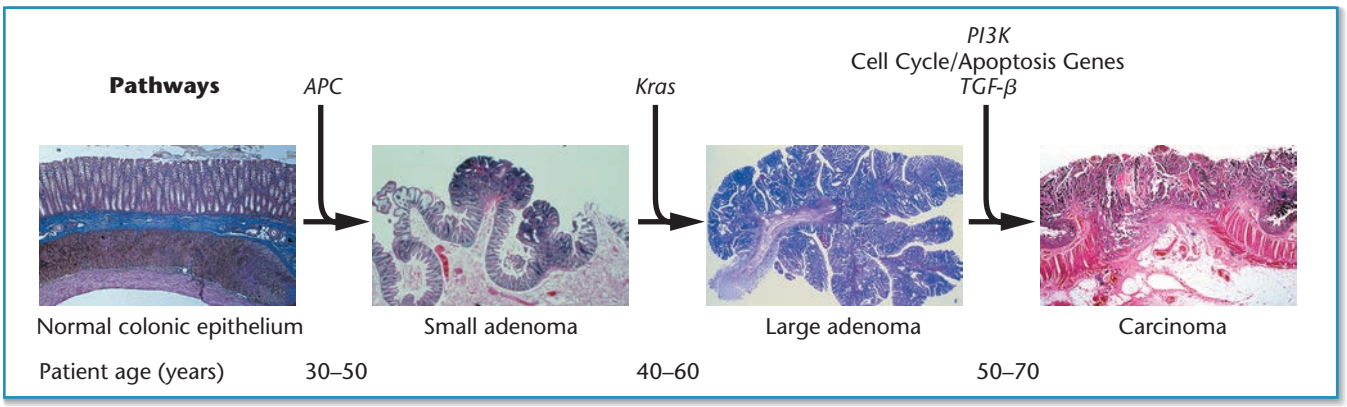


FIGURE 16–2 Steps in the development of colorectal cancers. Some of the genes that acquire driver mutations and cause the progressive development of colorectal cancer are shown at the top of the figure. These driver mutations accumulate over time and can take 40 years or more to result in the formation of a malignant tumor.

known to proceed through several clinical stages that are characterized by the stepwise accumulation of genetic defects in several genes (**Figure 16–2**). The first step is the conversion of a normal epithelial cell into a small cluster of cells known as an **adenoma** or **polyp.** This step requires inactivating mutations in the **adenomatous polyposis coli** (***APC***) gene, a gene that encodes a protein involved in the normal differentiation of intestinal cells. The *APC* gene is a tumor-suppressor gene, which will be discussed later in the chapter. The resulting adenoma grows slowly and is considered benign.

The second step in the development of colorectal cancer is the acquisition of a second genetic alteration in one of the cells within the small adenoma. This is usually a mutation in the ***Kras*** gene, a gene whose product is normally involved with regulating cell growth. The mutations in *Kras* that contribute to colorectal cancer cause the Kras protein to become constitutively active, resulting in unregulated cell division. The cell containing the *APC* and *Kras* mutations grows and expands to form a larger intermediate adenoma of approximately 1 cm in diameter—in a process known as **clonal expansion.** The cells of the original small adenoma (containing the *APC* mutation) are now vastly outnumbered by cells containing the two mutations.

The third step, which transforms a large adenoma into a malignant tumor (**carcinoma**), requires several more waves of clonal expansions triggered by the acquisition of defects in several genes, including *p53*, *PI3K*, and *TGF-β*. The products of these genes control several important aspects of normal cell growth and division, such as apoptosis, growth signaling, and cell-cycle regulation—all of which we will discuss in more detail later in the chapter. The resulting carcinoma is able to further grow and invade the underlying tissues of the colon. A few cells within the carcinoma acquire one or more new mutations that allow them to break free of the tumor, migrate to other parts of the body, and form metastases.

Driver Mutations and Passenger Mutations

Scientists are now applying some of the recent advances in DNA sequencing in order to identify all of the somatic mutations that occur during the development of a cancer cell. These studies compare the DNA sequences of genomes from cancer cells and normal cells derived from the same patient. Data from these studies are revealing that tens of thousands of somatic mutations can be present in cancer cells. Solid tumors such as those of the colon or breast may contain as many as 70 mutated genes. Some other cancers, such as lung cancer and melanomas, may contain several hundred mutations. Researchers believe that only a handful of mutations in each tumor—called **driver mutations**—give a growth advantage to a tumor cell. The remainder of the mutations may be acquired over time, perhaps as a result of the increased levels of DNA damage that accumulate in cancer cells, but these mutations have no direct contribution to the cancer phenotype. These are known as **passenger mutations.** The total number of driver mutations that occur in any particular cancer is small—between 2 and 8.

As we will discover in subsequent sections of this chapter, the genes that acquire driver mutations that lead to cancer (called oncogenes and tumor-suppressor genes) are those that control a large number of essential cellular functions. We will now investigate these fundamental processes, the genes that control them, and how mutations in these genes may lead to cancer.

ESSENTIAL POINT

The development of cancer is a multistep process, requiring mutations in several cancer-related genes. ■

16.2 Cancer Cells Contain Genetic Defects Affecting Genomic Stability, DNA Repair, and Chromatin Modifications

Many researchers believe that the fundamental defect in cancer cells is a derangement of the cells' normal ability to repair DNA damage. This loss of genomic integrity leads to a general increase in the mutation rate for every gene in the genome, including those whose products control aspects of cell proliferation, programmed cell death, and metastasis. The high level of genomic instability seen in cancer cells is known as the **mutator phenotype.** In addition, recent research has revealed that cancer cells contain aberrations in the types and locations of chromatin modifications, particularly DNA and histone methylation patterns.

Genomic Instability and Defective DNA Repair

Genomic instability in cancer cells is characterized by the presence of gross defects such as translocations, aneuploidy, chromosome loss, DNA amplification, and chromosome deletions. Often cancer cells show specific chromosomal defects that are used to diagnose the type and stage of the cancer. For example, leukemic white blood cells from patients with **chronic myelogenous leukemia (CML)** bear a specific translocation, in which the *C-ABL* gene on chromosome 9 is translocated into the *BCR* gene on chromosome 22. This translocation creates a structure known as the **Philadelphia chromosome** (Figure 16–3). The *BCR-ABL* fusion gene codes for a chimeric BCR-ABL protein. The normal ABL protein is a **protein kinase** that acts within signal transduction pathways, transferring growth factor signals from the external environment to the nucleus. The BCR-ABL protein is an abnormal signal transduction molecule in CML cells, which stimulates these cells to proliferate even in the absence of external growth signals.

In keeping with the concept of the cancer mutator phenotype, a number of inherited cancers are caused by defects in genes that control DNA repair. For example, xeroderma pigmentosum (XP) is a rare hereditary disorder that is characterized by extreme sensitivity to ultraviolet light and other carcinogens. Patients with XP often develop skin cancer. Cells from patients with XP are defective in nucleotide excision repair, with mutations appearing in any one of seven genes whose products are necessary to carry out DNA repair. XP cells are impaired in their ability to repair DNA lesions such as thymine dimers induced by UV light. The relationship between XP and genes controlling nucleotide excision repair is also described earlier in the text (see Chapter 14).

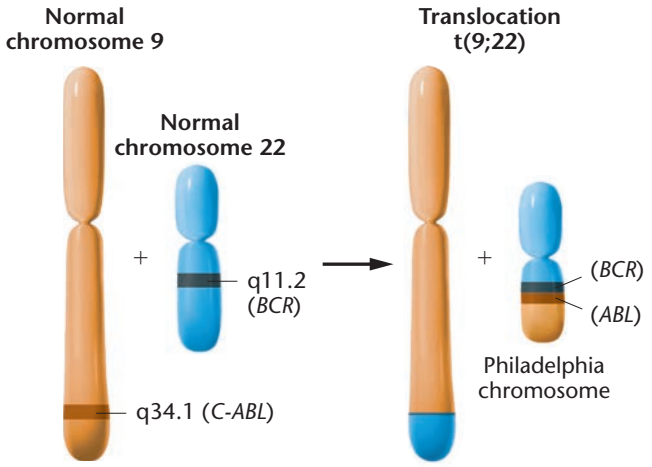


FIGURE 16–3 A reciprocal translocation involving the long arms of chromosomes 9 and 22 results in the formation of a characteristic chromosome, the Philadelphia chromosome, which is associated with chronic myelogenous leukemia (CML). The t(9;22) translocation results in the fusion of the *C-ABL* proto-oncogene on chromosome 9 with the *BCR* gene on chromosome 22. The fusion protein is a powerful hybrid molecule that allows cells to escape control of the cell cycle, contributing to the development of CML.

Another example is hereditary nonpolyposis colorectal cancer (HNPCC), which is caused by mutations in genes controlling DNA repair. HNPCC is an autosomal dominant syndrome, affecting about one in every 200 to 1000 people. Patients affected by HNPCC have an increased risk of developing colon, ovary, uterine, and kidney cancers. Cells from patients with HNPCC show higher than normal mutation rates and genomic instability. At least eight genes are associated with HNPCC, and four of these genes control aspects of DNA mismatch repair. Inactivation of any of these four genes—*MSH2, MSH6, MLH1,* and *MLH3*—causes a rapid accumulation of genome-wide mutations and the subsequent development of cancers.

The observation that hereditary defects in genes controlling nucleotide excision repair and DNA mismatch repair lead to high rates of cancer lends support to the idea that the mutator phenotype is a significant contributor to the development of cancer.

Chromatin Modifications and Cancer Epigenetics

The field of cancer epigenetics is providing new perspectives on the genetics of cancer. **Epigenetics** is the study of factors that affect gene expression but that do not alter the nucleotide sequence of DNA. DNA methylation and histone modifications such as acetylation and phosphorylation are examples of epigenetic modifications. The effects of chromatin modifications and epigenetic

factors on gene expression and hereditary disease are discussed in more detail later in the text (see Special Topic Chapter 1—Epigenetics).

Cancer cells contain altered DNA methylation patterns. Overall, there is much less DNA methylation in cancer cells than in normal cells. At the same time, the promoters of some genes are hypermethylated in cancer cells. These changes are thought to result in the release of transcription repression over the bulk of genes that would be silent in normal cells—including cancer-causing genes—while at the same time repressing transcription of genes that would regulate normal cellular functions such as DNA repair and cell-cycle control.

Histone modifications are also disrupted in cancer cells. Genes that encode histone acetylases, deacetylases, methyltransferases, and demethylases are often mutated or aberrantly expressed in cancer cells. The large numbers of epigenetic abnormalities in tumors have prompted some scientists to speculate that there may be more epigenetic defects in cancer cells than there are gene mutations. In addition, because epigenetic modifications are reversible, it may be possible to treat cancers using epigenetic-based therapies.

ESSENTIAL POINT

Cancer cells show high rates of mutation, chromosomal abnormalities, genomic instability, and abnormal patterns of chromatin modifications. ■

NOW SOLVE THIS

16–1 In chronic myelogenous leukemia (CML), leukemic blood cells can be distinguished from other cells of the body by the presence of a functional BCR-ABL hybrid protein. Explain how this characteristic provides an opportunity to develop a therapeutic approach to a treatment for CML.

■ **HINT:** *This problem asks you to imagine a therapy that is based on the unique genetic characteristics of CML leukemic cells. The key to its solution is to remember that the BCR-ABL fusion protein is found only in CML white blood cells and that this unusual protein has a specific function thought to directly contribute to the development of CML. To help you answer this problem, you may wish to learn more about the cancer drug Gleevec* (see http://www.cancer.gov/cancertopics/druginfo/imatinibmesylate).

16.3 Cancer Cells Contain Genetic Defects Affecting Cell-Cycle Regulation

One of the fundamental aberrations in all cancer cells is a loss of control over cell proliferation. Although some cells, such as epidermal cells of the skin or blood-forming cells in the bone marrow, continue to grow and divide throughout an organism's lifetime, most cells in adult multicellular organisms remain in a nondividing, quiescent, and differentiated state. **Differentiated cells** are those that are specialized for specific functions, such as photoreceptor cells of the retina or muscle cells of the heart. Normal regulation over cell proliferation involves a large number of gene products that control steps in the cell cycle, programmed cell death, and the response of cells to external growth signals. In cancer cells, many of the genes that control these functions are mutated or aberrantly expressed, leading to uncontrolled cell proliferation.

In this section, we will review steps in the cell cycle, some of the genes that control the cell cycle, and how these genes, when mutated, lead to cancer.

The Cell Cycle and Signal Transduction

As we learned earlier in the text (see Chapter 2), the cellular events that occur in sequence from one cell division to the next comprise the **cell cycle.**

Cells in the G0 phase of the cell cycle can often be stimulated to reenter the cell cycle by external growth signals. These signals are delivered to the cell by molecules such as growth factors and hormones that bind to cell-surface receptors, which then relay the signal from the plasma membrane to the cytoplasm. The process of transmitting growth signals from the external environment to the cell nucleus is known as **signal transduction.** Ultimately, signal transduction initiates a program of gene expression that propels the cell out of G0 back into the cell cycle. Cancer cells often have defects in signal transduction pathways. Sometimes, abnormal signal transduction molecules send continuous growth signals to the nucleus even in the absence of external growth signals. An example of abnormal signal transduction due to mutations in the *ras* gene is described in Section 16.4. In addition, malignant cells may not respond to external signals from surrounding cells—signals that would normally inhibit cell proliferation within a mature tissue.

Cell-Cycle Control and Checkpoints

In normal cells, progress through the cell cycle is tightly regulated, and each step must be completed before the next step can begin. There are at least three distinct points in the cell cycle at which the cell monitors external signals and internal equilibrium before proceeding to the next stage. These are the **G1/S,** the **G2/M,** and **M checkpoints.** At the G1/S checkpoint, the cell monitors its size and determines whether its DNA has been damaged. If the cell has not achieved an adequate size, or if the DNA has been damaged, further progress through the cell cycle is halted until these conditions are corrected. If cell size and DNA integrity are normal, the G1/S checkpoint is traversed, and the cell

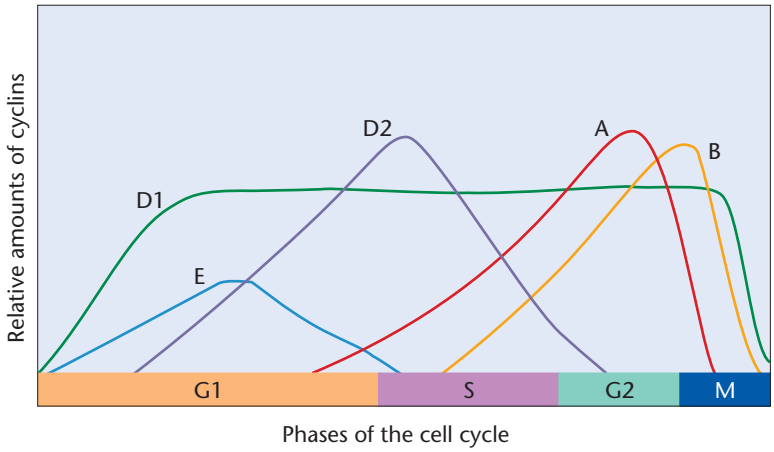


FIGURE 16–4 Relative expression times and amounts of cyclins during the cell cycle.

proceeds to S phase. The second important checkpoint is the G2/M checkpoint, where physiological conditions in the cell are monitored prior to mitosis. If DNA replication or repair of any DNA damage has not been completed, the cell cycle arrests until these processes are complete. The third major checkpoint occurs during mitosis and is called the M checkpoint. At this checkpoint, both the successful formation of the spindle-fiber system and the attachment of spindle fibers to the kinetochores associated with the centromeres are monitored. If spindle fibers are not properly formed or attachment is inadequate, mitosis is arrested.

In addition to regulating the cell cycle at checkpoints, the cell controls progress through the cell cycle by means of two classes of proteins: **cyclins** and **cyclin-dependent kinases** (**CDKs**). The cell accumulates and destroys cyclin proteins in a precise pattern during the cell cycle (**Figure 16–4**). When a cyclin is present, it binds to a specific CDK, triggering activity of the CDK/cyclin complex. The CDK/cyclin complex then selectively phosphorylates and activates other proteins that in turn bring about the changes necessary to advance the cell through the cell cycle. For example, in G1 phase, CDK4/cyclin D complexes activate proteins that stimulate transcription of genes whose products (such as DNA polymerase δ and DNA ligase) are required for DNA replication during S phase. Another CDK/cyclin complex, CDK1/cyclin B, phosphorylates a number of proteins that bring about the events of early mitosis, such as nuclear membrane breakdown, chromosome condensation, and cytoskeletal reorganization. Mitosis can only be completed, however, when cyclin B is degraded and the protein phosphorylations characteristic of M phase are reversed. Although a large number of different protein kinases exist in cells, only a few are involved in cell-cycle regulation.

Mutation or misexpression of any of the genes controlling the cell cycle can contribute to the development of cancer. For example, if genes that control the G1/S or G2/M checkpoints are mutated, the cell may continue to grow and divide without repairing DNA damage. As these cells continue to divide, they accumulate mutations in genes whose products control cell proliferation or metastasis. Similarly, if genes that control progress through the cell cycle, such as those that encode the cyclins, are expressed at the wrong time or at incorrect levels, the cell may grow and divide continuously and may be unable to exit the cell cycle into G0. The result in both cases is that the cell loses control over proliferation and is on its way to becoming cancerous.

Control of Apoptosis

As already described, if DNA replication, repair, or chromosome assembly is defective, normal cells halt their progress through the cell cycle until the condition is corrected. This reduces the number of mutations and chromosomal abnormalities that accumulate in normal proliferating cells. However, if DNA or chromosomal damage is so severe that repair is impossible, the cell may initiate a second line of defense—a process called **apoptosis,** or **programmed cell death.** Apoptosis is a genetically controlled process whereby the cell commits suicide. Besides its role in preventing cancer, apoptosis is also initiated during normal multicellular development in order to eliminate certain cells that do not contribute to the final adult organism. The steps in apoptosis are the same for damaged cells and for cells being eliminated during development: nuclear DNA becomes fragmented, internal cellular structures are disrupted, and the cell dissolves into small spherical structures known as apoptotic bodies. In the final step, the apoptotic bodies are engulfed by the immune system's phagocytic cells. A series of proteases called **caspases** are responsible for initiating apoptosis and for digesting intracellular components.

By removing damaged cells, programmed cell death reduces the number of mutations that are passed to the next generation, including those in cancer-causing genes. Some of the same genes that control cell-cycle checkpoints can trigger apoptosis. These genes are mutated in many cancers. As a result of the mutation or inactivation of these checkpoint genes, the cell is unable to repair its DNA or undergo apoptosis. This inability leads to the accumulation of even more mutations in genes that control growth, division, and metastasis.

ESSENTIAL POINT

Cancer cells have defects in cell-cycle progression, checkpoint controls, and programmed cell death. ■

16.4 Proto-oncogenes and Tumor-Suppressor Genes Are Altered in Cancer Cells

Two general categories of genes are mutated or misexpressed in cancer cells—the proto-oncogenes and the tumor-suppressor genes (Table 16.1). **Proto-oncogenes** encode transcription factors that stimulate expression of other genes, signal transduction molecules that stimulate cell division, and cell-cycle regulators that move the cell through the cell cycle. Their products are important for normal cell functions, especially cell growth and division. When normal cells become quiescent and cease division, they repress the expression of most proto-oncogenes or modify the activities of their products. In cancer cells, one or more proto-oncogenes are altered in such a way that the activities of their products cannot be regulated in a normal fashion. This is sometimes due to mutations that result in an abnormal protein product. In other cases, proto-oncogenes may be overexpressed or expressed at an incorrect time due to mutations within gene-regulatory regions such as enhancer elements or due to alterations in chromatin structure that affect gene expression. If a proto-oncogene is continually in an "on" state, its product may constantly stimulate the cell to divide. When a proto-oncogene is mutated or abnormally expressed and contributes to the development of cancer, it is known as an **oncogene**—a cancer-causing gene. Oncogenes are proto-oncogenes that have experienced a gain-of-function alteration. As a result, only one allele of a proto-oncogene needs to be mutated or misexpressed in order to contribute to cancer. Hence, oncogenes confer a dominant cancer phenotype.

Tumor-suppressor genes are genes whose products normally regulate cell-cycle checkpoints or initiate the process of apoptosis. In normal cells, proteins encoded by tumor-suppressor genes halt progress through the cell cycle in response to DNA damage or growth-suppression signals from the extracellular environment. When tumor-suppressor genes are mutated or inactivated, cells are unable to respond normally to cell-cycle checkpoints, or are unable to undergo programmed cell death if DNA damage is extensive. This leads to the accumulation of more mutations and the development of cancer. When both alleles of a tumor-suppressor gene are inactivated through mutation or epigenetic modifications, and other changes in the cell keep it growing and dividing, cells may become tumorigenic.

The following are examples of proto-oncogenes and tumor-suppressor genes that contribute to cancer when mutated or abnormally expressed. Approximately 400 oncogenes and tumor-suppressor genes are now known, and more will likely be discovered as cancer research continues.

The *ras* Proto-oncogenes

Some of the most frequently mutated genes in human tumors are those in the ***ras* gene family.** These genes are mutated in more than 30 percent of human tumors. The *ras* gene family encodes signal transduction molecules that are associated with the cell membrane and regulate cell growth and division. Ras proteins normally transmit signals from the cell membrane to the nucleus, stimulating the cell to divide in response to external growth factors. Ras proteins alternate between an inactive (switched off) and an active (switched on) state by binding either guanosine diphosphate (GDP) or guanosine triphosphate (GTP). Mutations that convert the *ras* proto-oncogene to an oncogene prevent the Ras protein from hydrolyzing GTP to GDP and hence freeze the Ras protein into its "on" conformation, constantly stimulating the cell to divide.

TABLE 16.1 Some Proto-oncogenes and Tumor-Suppressor Genes

Proto-oncogene	Normal Function	Alteration in Cancer	Associated Cancers
c-myc	Transcription factor, regulates cell cycle, differentiation, apoptosis	Translocation, amplification, point mutations	Lymphomas, leukemias, lung cancer, many types
c-kit	Tyrosine kinase, signal transduction	Mutation	Sarcomas
RARα	Hormone-dependent transcription factor, differentiation	Chromosomal translocations with *PML* gene, fusion product	Acute promyelocytic leukemia
Cyclins	Bind to CDKs, regulate cell cycle	Gene amplification, overexpression	Lung, esophagus, many types
Tumor-Suppressor	**Normal Function**	**Alteration in Cancer**	**Associated Cancers**
RB1	Cell-cycle checkpoints, binds E2F	Mutation, deletion, inactivation by viral oncogene products	Retinoblastoma, osteosarcoma, many types
APC	Cell–cell interaction	Mutation	Colorectal cancers, brain, thyroid
p53	Transcription regulation	Mutation, deletion, viruses	Many types
BRCA1, BRCA2	DNA repair	Point mutations	Breast, ovarian, prostate cancers

The *p53* Tumor-Suppressor Gene

The most frequently mutated gene in human cancers—mutated in more than 50 percent of all cancers —is the ***p53* gene.** This gene encodes a transcription factor that represses or stimulates transcription of more than 50 different genes.

Normally, the p53 protein is continuously synthesized but is rapidly degraded and therefore is present in cells at low levels. Several types of cellular stress events bring about rapid increases in the nuclear levels of activated p53 protein. These include chemical damage to DNA, double-stranded breaks in DNA induced by ionizing radiation, and the presence of DNA-repair intermediates generated by exposure of cells to ultraviolet light.

The p53 protein initiates several different responses to DNA damage including cell-cycle arrest followed by DNA repair and apoptosis if DNA cannot be repaired. These responses are accomplished by p53 acting as a transcription factor that stimulates or represses the expression of genes involved in each response.

In normal cells, p53 can arrest the cell cycle at the G1/S and G2/M checkpoints, as well as retard the progression of the cell through S phase. It accomplishes this by inhibiting cyclin/CDK complexes and regulating the transcription of other genes involved in these phases of the cell cycle.

The p53 protein can also instruct a damaged cell to commit suicide by apoptosis. It does so by activating the transcription of genes whose products control this process. In cancer cells that lack sufficient p53, these gene products are not synthesized and apoptosis may not occur.

Cells lacking functional p53 are unable to arrest at cell-cycle checkpoints or to enter apoptosis in response to DNA damage. As a result, they move unchecked through the cell cycle, regardless of the condition of the cell's DNA. Cells lacking p53 have high mutation rates and accumulate the types of mutations that lead to cancer. Because of the importance of the *p53* gene to genomic integrity, it is often referred to as the "guardian of the genome."

The *RB1* Tumor-Suppressor Gene

The loss or mutation of the ***RB1*** **(retinoblastoma 1)** tumor-suppressor gene contributes to the development of many cancers, including those of the breast, bone, lung, and bladder. The *RB1* gene was originally identified as a result of studies on **retinoblastoma,** an inherited disorder in which tumors develop in the eyes of young children. Retinoblastoma occurs with a frequency of about 1 in 15,000 individuals. In the familial form of the disease, individuals inherit one mutated allele of the *RB1* gene and have an 85 percent chance of developing retinoblastomas as well as an increased chance of developing other cancers. All somatic cells of patients with hereditary retinoblastoma contain one mutated allele of the *RB1* gene. However, it is only when the second normal allele of the *RB1* gene is lost or mutated in certain retinal cells that retinoblastoma develops. In individuals who do not have this hereditary condition, retinoblastoma is extremely rare, as it requires at least two separate somatic mutations in a retinal cell in order to inactivate both copies of the *RB1* gene.

The **retinoblastoma protein (pRB)** is a tumor-suppressor protein that controls the G1/S cell-cycle checkpoint. The pRB protein is found in the nuclei of all cell types at all stages of the cell cycle. However, its activity varies throughout the cell cycle, depending on its phosphorylation state. When cells are in the G0 phase of the cell cycle, the pRB protein is nonphosphorylated and binds to transcription factors such as E2F, inactivating them (**Figure 16–5**). When the cell is stimulated by growth factors, it enters G1 and approaches S phase. Throughout the G1 phase, the pRB protein becomes phosphorylated by the CDK4/cyclin D1 complex. Phosphorylated pRB releases its bound regulatory proteins. When E2F and other regulators are released by pRB, they are free to induce the expression of over 30 genes whose products are required for the transition from G1 into S phase. After cells traverse S, G2, and M phases, pRB reverts to a nonphosphorylated state, binds to regulatory proteins such as E2F, and keeps them sequestered until required for the next cell cycle. In normal

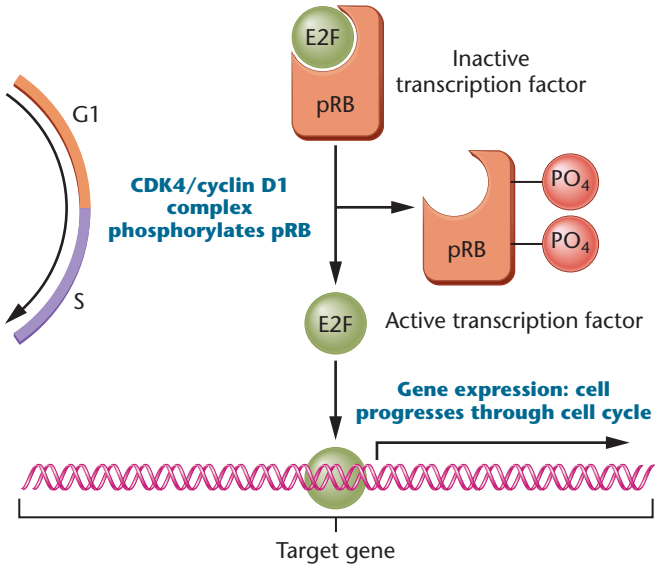


FIGURE 16–5 During G0 and early G1, pRB interacts with and inactivates transcription factor E2F. As the cell moves from G1 to S phase, a CDK4/cyclin D1 complex forms and adds phosphate groups to pRB. As pRB becomes phosphorylated, E2F is released and becomes transcriptionally active, allowing the cell to pass through S phase. Phosphorylation of pRB is transitory; as CDK/cyclin complexes are degraded and the cell moves through the cell cycle to early G1, pRB phosphorylation declines, allowing pRB to reassociate with E2F.

quiescent cells, the presence of the pRB protein prevents passage into S phase. In many cancer cells, including retinoblastoma cells, both copies of the *RB1* gene are defective, inactive, or absent, and progression through the cell cycle is not regulated.

NOW SOLVE THIS

16–2 People with a genetic condition known as Li–Fraumeni syndrome inherit one mutant copy of the *p53* gene. These people have a high risk of developing a number of different cancers, such as breast cancer, leukemia, bone cancer, adrenocortical tumors, and brain tumors. Explain how mutations in one cancer-related gene can give rise to such a diverse range of tumors.

HINT: *This problem involves an understanding of how tumor-suppressor genes regulate cell growth and behavior. The key to its solution is to consider which cellular functions are regulated by the p53 protein and how the absence of p53 could affect each of these functions. Also, read about loss of heterozygosity in Section 16.6.*

16.5 Cancer Cells Metastasize and Invade Other Tissues

As discussed at the beginning of this chapter, uncontrolled growth alone is insufficient to create a malignant and life-threatening cancer. Cancer cells must also become malignant, acquiring the ability to disengage from the original tumor site, to enter the blood or lymphatic system, to invade surrounding tissues, and to develop into secondary tumors. In order to leave the site of the primary tumor and invade other tissues, tumor cells must dissociate from the primary tumor and secrete proteases that digest components of the **extracellular matrix** and **basal lamina,** which normally surround and separate the body's tissues. The extracellular matrix and basal lamina are composed of proteins and carbohydrates. They surround and separate body tissues, form the scaffold for tissue growth, and inhibit the migration of cells.

The ability to invade the extracellular matrix is also a property of some normal cell types. For example, implantation of the embryo in the uterine wall during pregnancy requires cell migration across the extracellular matrix. In addition, white blood cells reach sites of infection by penetrating capillary walls. The mechanisms of invasion are probably similar in these normal cells and in cancer cells. The difference is that, in normal cells, the invasive ability is tightly regulated, whereas in tumor cells, this regulation has been lost.

Once cancer cells have disengaged from the primary tumor and traversed tissue barriers, they enter the blood or lymphatic system and may become lodged in microvessels of other tissues. At this point the cells may undergo a second round of invasion to enter the new tissue and grow into new (metastatic) tumors. Only a small percentage of circulating cancer cells—about 0.01 percent—survive to establish metastatic tumors. Other important features of metastatic cells are increased cell motility, the capacity to stimulate new blood vessel formation, and the ability to escape detection by the host's immune system.

Metastasis is controlled by a large number of gene products, including cell-adhesion molecules, cytoskeleton regulators, and proteolytic enzymes. For example, epithelial tumors have a lower than normal level of the **E-cadherin glycoprotein,** which is responsible for cell–cell adhesion in normal tissues. Also, proteolytic enzymes such as **metalloproteinases** are present at higher than normal levels in many highly malignant tumors. For example, breast cancer cells that metastasize to bone abnormally express the metalloproteinase gene *MMP1*. Those that spread to the lungs overexpress the *MMP1* and *MMP2* genes. It has been shown that the level of aggressiveness of a tumor correlates positively with the levels of proteolytic enzymes expressed by the tumor. In addition, malignant cells are not susceptible to the normal controls conferred by regulatory molecules such as **tissue inhibitors of metalloproteinases** (**TIMPs**).

ESSENTIAL POINT

The ability of cancer cells to metastasize requires defects in gene products that control a number of functions such as cell adhesion, proteolysis, and tissue invasion.

16.6 Predisposition to Some Cancers Can Be Inherited

Although the vast majority of human cancers are sporadic, a small fraction (approximately 5 percent) have a hereditary or familial component. Some of these inherited forms of cancer are listed in Table 16.2.

Most inherited cancer-susceptibility alleles occur in tumor-suppressor genes, and though transmitted in a Mendelian dominant fashion, are not sufficient in themselves to trigger development of a cancer. At least one other somatic mutation in the other copy of the gene must occur in order to drive a cell toward tumorigenesis. In addition, mutations in still other genes are usually necessary to fully express the cancer phenotype. As mentioned earlier, inherited mutations in the *RB1* gene predispose individuals to developing various cancers. Although the normal somatic cells of these patients are heterozygous for the *RB1* mutation, cells within their tumors contain mutations in both copies of the gene. The phenomenon whereby the second, wild-type,

TABLE 16.2 Some Inherited Predispositions to Cancer

Tumor Predisposition Syndrome	Gene Affected
Early-onset familial breast cancer	*BRCA1*
Familial adenomatous polyposis	*APC*
Familial melanoma	*CDKN2*
Gorlin syndrome	*PTCH1*
Hereditary nonpolyposis colon cancer	*MSH2, 6*
Li-Fraumeni syndrome	*p53*
Multiple endocrine neoplasia, type 1	*MEN1*
Multiple endocrine neoplasia, type 2	*RET*
Neurofibromatosis, type 1	*NF1*
Neurofibromatosis, type 2	*NF2*
Retinoblastoma	*RB1*
Von Hippel–Lindau syndrome	*VHL*
Wilms tumor	*WT1*

allele is mutated in a tumor is known as **loss of heterozygosity.** Although loss of heterozygosity is an essential first step in expression of these inherited cancers, further mutations in other proto-oncogenes, tumor-suppressor genes, or chromatin-modifying genes are necessary for the tumor cells to become fully malignant.

The development of hereditary colon cancer illustrates how inherited mutations in one allele of a gene contribute only one step in the multistep pathway leading to malignancy. In Section 16.1, we described how colorectal cancers develop through the accumulation of mutations in several genes, leading to a stepwise clonal expansion of cells and the development of carcinomas. Although the vast majority of colorectal cancers are sporadic, about 1 percent of cases result from a genetic predisposition to cancer known as **familial adenomatous polyposis** (**FAP**). In FAP, individuals inherit one mutant copy of the *APC* (adenomatous polyposis) gene located on the long arm of chromosome 5. Mutations include deletions, frameshift, and point mutations. The normal function of the *APC* gene product is to act as a tumor suppressor controlling growth and differentiation. The presence of a heterozygous *APC* mutation causes the epithelial cells of the colon to partially escape cell-cycle control, and the cells divide to form small clusters of cells called **polyps** or adenomas. People who are heterozygous for this condition develop hundreds to thousands of colon and rectal polyps early in life. Although it is not necessary for the second allele of the *APC* gene to be mutated in polyps at this stage, in the majority of cases, the second *APC* allele becomes mutant in a later stage of cancer development. The remaining steps in development of colorectal carcinoma follow the same order as that shown in Figure 16–2.

ESSENTIAL POINT

Inherited mutations in cancer-susceptibility genes are not sufficient to trigger cancer. Other somatic mutations in proto-oncogenes or tumor-suppressor genes are necessary for the development of hereditary cancers. ■

NOW SOLVE THIS

16–3 Although tobacco smoking is responsible for a large number of human cancers, not all smokers develop cancer. Similarly, some people who inherit mutations in the tumor-suppressor genes *p53* or *RB1* never develop cancer. Explain these observations.

■ **HINT:** *This problem asks you to consider the reasons why only some people develop cancer as a result of environmental factors or mutations in tumor-suppressor genes. The key to its solution is to consider the steps involved in the development of cancer and the number of abnormal functions in cancer cells. Also, consider how genetics may affect DNA repair functions.*

16.7 Viruses and Environmental Agents Contribute to Human Cancers

It is thought that, worldwide, about 15 percent of human cancers are associated with viruses, making virus infection the second largest risk factor for cancer, next to tobacco smoking. The most significant contributors to virus-induced human cancers are listed in Table 16.3. Like other risk factors for cancer, including hereditary predisposition to certain cancers, virus infection alone is not sufficient to trigger human cancers. Other factors, including DNA damage or the accumulation of mutations in one or more of a cell's oncogenes and tumor-suppressor genes, are required to move a cell down the multistep pathway to cancer.

In addition to viruses, environmental agents also contribute to the development of cancer. Any substance or event that damages DNA has the potential to be carcinogenic.

TABLE 16.3 Human Viruses Associated with Cancers

Virus		Associated Cancers
DNA Viruses		
Epstein-Barr virus	EBV	Burkitt lymphoma, nasopharyngeal carcinoma, Hodgkin lymphoma
Hepatitis B virus	HBV	Hepatocellular carcinoma
Hepatitis C virus	HCV	Hepatocellular carcinoma, non-Hodgkin lymphoma
Human papilloma viruses 16, 18	HPV16, 18	Cervical cancer, anogenital cancers, oral cancers
Retroviruses		
Human T-cell lymphotropic virus type 1	HTLV-1	Adult T-cell leukemia and lymphoma
Human immunodeficiency virus type-1	HIV-1	Immune suppression, leading to cancers

Our environment, both natural and human-made, contains abundant carcinogens. These include chemicals, radiation, and chronic infections. Perhaps the most significant carcinogen in our environment is tobacco smoke, which contains at least 60 chemicals that interact with DNA and cause mutations. Epidemiologists estimate that about 30 percent of human cancer deaths are associated with cigarette smoking. Smokers have a 20-fold increased risk of developing lung cancer, which kills more than one million people, worldwide, each year.

Diet is often implicated in the development of cancer. Consumption of red meat and animal fat is associated with some cancers, such as colon, prostate, and breast cancer. The mechanisms by which these substances may contribute to carcinogenesis may involve stimulation of cell division through hormones or creation of carcinogenic chemicals during cooking. Alcohol may cause inflammation of the liver and contribute to liver cancer.

Although most people perceive the human-made, industrial environment to be a highly significant contributor to cancer, it may account for only a small percentage of total cancers, and only in special situations. Some of the most mutagenic agents, and hence potentially the most carcinogenic, are natural substances and natural processes. For example, **aflatoxin,** a component of a mold that grows on peanuts and corn, is one of the most carcinogenic chemicals known. Most chemical carcinogens, such as **nitrosamines,** are components of synthetic substances and are found in some preserved meats; however, many are naturally occurring. For example, natural pesticides and antibiotics found in plants may be carcinogenic, and the human body itself creates alkylating agents in the acidic environment of the gut. Nevertheless, these observations do not diminish the serious cancer risks to specific populations who are exposed to human-made carcinogens such as synthetic pesticides or asbestos.

DNA lesions brought about by natural radiation (X rays, ultraviolet light), dietary substances, and substances in the external environment contribute the majority of environmentally caused mutations that lead to cancer. In addition, normal metabolism creates oxidative end products that can damage DNA, proteins, and lipids. It is estimated that the human body suffers about 10,000 damaging DNA lesions per day due to the actions of oxygen free radicals. DNA repair enzymes deal successfully with most of this damage; however, some damage may lead to permanent mutations. The process of DNA replication itself is mutagenic. Hence, substances such as growth factors or hormones that stimulate cell division are ultimately mutagenic and perhaps carcinogenic. Chronic inflammation due to infection also stimulates tissue repair and cell division, resulting in DNA lesions accumulating during replication. These mutations may persist, particularly if cell-cycle checkpoints are compromised due to mutations or inactivation of tumor-suppressor genes such as *p53* or *RB1*.

GENETICS, TECHNOLOGY, AND SOCIETY

Breast Cancer: The Double-Edged Sword of Genetic Testing

The prospect of using genetics to prevent and cure a wide range of diseases is exciting. However, in our enthusiasm, we often forget that these new technologies still have significant limitations and profound ethical complexities. The story of genetic testing for breast cancer illustrates how we must temper our high expectations with respect for uncertainty.

Breast cancer is the most common cancer among women. A woman's lifetime risk of developing breast cancer is about 12 percent. Each year, more than 200,000 new cases are diagnosed in the United States. Breast cancer is not limited to women; about 1400 men are also diagnosed with the disease each year.

Approximately 5 to 10 percent of breast cancers are familial, a category defined by the early onset of the disease and the appearance of several cases of breast or ovarian cancer among near blood relatives. In 1994, two genes were identified that show linkage to familial breast cancers: *BRCA1* and *BRCA2*. These two genes encode tumor-suppressor proteins that are involved in repairing damaged DNA. Women who inherit germ-line mutations in *BRCA1* have an approximately 60 percent chance of developing breast cancer and a 39 percent chance of developing ovarian cancer. Those who inherit mutations in *BRCA2* have an approximately 45 percent chance of developing breast cancer and a 15 percent chance of developing ovarian cancer. In men, mutations in these two genes lead to increased risks of both breast and prostate cancers.

BRCA1 and *BRCA2* genetic tests are available, and these detect over 2000 different mutations that are known to occur within the coding regions of these genes. Many patients at risk for familial breast cancer opt to undergo genetic testing. These patients feel that test results could motivate them to take steps to prevent breast or ovarian cancers, guide them in childbearing decisions, and provide information concerning the risk to close relatives. But all these potential benefits are fraught with uncertainties.

A woman whose *BRCA* test results are negative may feel relieved and assume that she is not subject to familial breast cancer. However, her risk of developing breast cancer is still 12 percent (the population risk), and she should continue to monitor herself for the disease. Also, a negative *BRCA* genetic test does not eliminate the possibility that she carries an inherited mutation in another gene that increases breast cancer risk or that her *BRCA1* or *BRCA2* gene mutations exist in regions of the genes that are inaccessible to current genetic tests.

A woman whose test results are positive faces difficult choices. Her treatment options consist of close monitoring, prophylactic mastectomy or oophorectomy (removal of breasts and ovaries, respectively), or taking prophylactic drugs such as tamoxifen. Prophylactic surgery reduces her risk but does not eliminate it, as cancers can still occur in tissues that remain after surgery. Drugs such as tamoxifen are helpful but have serious side effects. Genetic tests also affect the patient's entire family. People often experience fear, anxiety, and guilt on learning that they are carriers of a genetic disease. Confidentiality is also a major concern. Patients fear that their genetic test results may be leaked to insurance companies or employers, jeopardizing their prospects for jobs or affordable health and life insurance.

The unanswered scientific and ethical questions about *BRCA1* and *BRCA2* genetic testing are many and important. As we develop genetic tests for more and more diseases over the next few decades, our struggle with these kinds of questions will continue.

Your Turn

Take time, individually or in groups, to answer the following questions. Investigate the references and links, to help you understand some of the issues that surround genetic testing for breast cancer.

1. New genomics research is rapidly identifying genes linked to human diseases, including breast cancer. How many genes are now thought to be involved in familial breast cancer? Are genetic tests available to detect mutations in these genes?

Search for recent scientific data on breast cancer susceptibility genes by using the PubMed Web site (http://www.ncbi.nlm.nih.gov/pubmed), as described in the Exploring Genomics feature in Chapter 2.

2. Certain ethnic groups have a higher than average prevalence of mutations in *BRCA1* and *BRCA2*. Recently, Israel proposed a national screening program to detect *BRCA1* and *BRCA2* mutations among all women in the Ashkenazi Jewish population. What would be the scientific and ethical pros and cons of conducting such a wide population screen, rather than restricting genetic testing to people from high-risk families?

Read about this topic in the New York Times *article at* www.nytimes.com/2013/11/27/health/in-israel-a-push-to-screen-for-cancer-gene-leaves-many-conflicted.html

3. If your family was at risk for familial breast cancer, would you opt to take the *BRCA1* and *BRCA2* genetic tests? What actions would you take if you received a positive test result? How would you feel about such a result?

Two helpful sources are: (a) Surbone, A. 2001. Ethical implications of genetic testing for breast cancer susceptibility. *Crit. Rev. in Onc./Hem.* 40: 149–157. *(b) BRCA1 and BRCA2: Cancer Risk and Genetic Testing. National Cancer Institute Fact Sheet.* http://www.cancer.gov/cancertopics/factsheet/Risk/BRCA.

CASE STUDY | Screening for cancer can save lives

A woman who was a heavy smoker visited her doctor for a cervical smear test to screen for cancer. When the cytology results came, a large number of abnormal precancerous cells were found on her cervix. These were successfully removed under a local anesthetic, using a technique called the large loop excision of the transformation zone (LLETZ), which cut away the precancerous area. A follow-up cervical smear test six months later showed only normal cells.

1. What might have happened if the woman had not bothered to go for the screening?
2. What can the woman do to reduce the risk of cervical and other cancers in the future?
3. What other risk factors are closely linked to cervical cancer?

INSIGHTS AND SOLUTIONS

1. In disorders such as retinoblastoma, a mutation in one allele of the *RB1* gene can be inherited from the germ line, causing an autosomal dominant predisposition to the development of eye tumors. To develop tumors, a somatic mutation in the second copy of the *RB1* gene is necessary, indicating that the mutation itself acts as a recessive trait. Given that the first mutation can be inherited, in what ways can a second mutational event occur?

Solution: In considering how this second mutation arises, we must look at several types of mutational events, including changes in nucleotide sequence and events that involve whole chromosomes or chromosome parts. Retinoblastoma results when both copies of the *RB1* locus are lost or inactivated. With this in mind, you must first list the phenomena that can result in a mutational loss or the inactivation of a gene.

(continued)

Insights and Solutions—continued

One way the second *RB1* mutation can occur is by a nucleotide alteration that converts the remaining normal *RB1* allele to a mutant form. This alteration can occur through a nucleotide substitution or through a frameshift mutation caused by the insertion or deletion of nucleotides during replication. A second mechanism involves the loss of the chromosome carrying the normal allele. This event would take place during mitosis, resulting in chromosome 13 monosomy and leaving the mutant copy of the gene as the only *RB1* allele. This mechanism does not necessarily involve loss of the entire chromosome; deletion of the long arm (*RB1* is on 13q) or an interstitial deletion involving the *RB1* locus and some surrounding material would have the same result. Alternatively, a chromosome aberration involving loss of the normal copy of the *RB1* gene might be followed by duplication of the chromosome carrying the mutant allele. Two copies of chromosome 13 would be restored to the cell, but the normal *RB1* allele would not be present. Finally, a recombination event followed by chromosome segregation could produce a homozygous combination of mutant *RB1* alleles.

2. Proto-oncogenes can be converted to oncogenes in a number of different ways. In some cases, the proto-oncogene itself becomes amplified up to hundreds of times in a cancer cell. An example is the *cyclin D1* gene, which is amplified in some cancers. In other cases, the proto-oncogene may be mutated in a limited number of specific ways, leading to alterations in the gene product's structure. The *ras* gene is an example of a proto-oncogene that becomes oncogenic after suffering point mutations in specific regions of the gene. Explain why these two proto-oncogenes (*cyclin D1* and *ras*) undergo such different alterations in order to convert them into oncogenes.

Solution: The first step in solving this question is to understand the normal functions of these proto-oncogenes and to think about how either amplification or mutation would affect each of these functions.

The cyclin D1 protein regulates progression of the cell cycle from G1 into S phase, by binding to CDK4 and activating this kinase. The cyclin D1/CDK4 complex phosphorylates a number of proteins including pRB, which in turn activate other proteins in a cascade that results in transcription of genes whose products are necessary for DNA replication in S phase. The simplest way to increase the activity of cyclin D1 would be to increase the number of cyclin D1 molecules available for binding to the cell's endogenous CDK4 molecules. This can be accomplished by several mechanisms, including amplification of the *cyclin D1* gene. In contrast, a point mutation in the *cyclin D1* gene would most likely interfere with the ability of the cyclin D1 protein to bind to CDK4; hence, mutations within the gene would probably repress cell-cycle progression rather than stimulate it.

The *ras* gene product is a signal transduction protein that operates as an on/off switch in response to external stimulation by growth factors. It does so by binding either GTP (the "on" state) or GDP (the "off" state). Oncogenic mutations in the *ras* gene occur in specific regions that alter the ability of the Ras protein to exchange GDP for GTP. Oncogenic Ras proteins are locked in the "on" conformation, bound to GTP. In this way, they constantly stimulate the cell to divide. An amplification of the *ras* gene would simply provide more molecules of normal Ras protein, which would still be capable of on/off regulation. Hence, simple amplification of *ras* would less likely be oncogenic.

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we focused on cancer as a genetic disease. In particular, we discussed the relationship between cancer, the cell cycle, and mutations in proto-oncogenes and tumor-suppressor genes. Based on your knowledge of these topics, answer several fundamental questions:
   (a) How do we know that malignant tumors arise from a single cell that contains mutations?
   (b) How do we know that cancer development requires more than one mutation?
   (c) How do we know that cancer cells contain defects in DNA repair?

CONCEPT QUESTION

2. Review the Chapter Concepts list on page 321. These concepts relate to the multiple ways in which genetic alterations lead to the development of cancers. The sixth concept states that DNA methylation and histone modifications contribute to the genetic alterations leading to cancer. Write a short essay describing how these changes in cancer cells contribute to the development of cancers. ■
3. What is the relationship between signal transduction and cellular proliferation?
4. How do normal cells and cancer cells differ in terms of cell-cycle regulation?
5. Describe kinases and cyclins. How do they interact to cause cells to move through the cell cycle?
6. What is the role of the retinoblastoma protein in cell-cycle regulation? Is the retinoblastoma gene a tumor-suppressor gene or an oncogene?
7. Can cancer be inherited or infectious?
8. What is apoptosis, and under what circumstances do cells undergo this process?
9. Define tumor-suppressor genes. Why is a mutation in a single copy of a tumor-suppressor gene expected to behave as a recessive gene?
10. A genetic variant of the retinoblastoma protein, called PSM-RB (phosphorylation site mutated RB), is not able to be phosphorylated by the action of CDK4/cyclin D1 complex. Explain why PSM-RB is said to have a constitutive growth-suppressing action on the cell cycle.
11. Part of the Ras protein is associated with the plasma membrane, and part extends into the cytoplasm. How does the Ras protein transmit a signal from outside the cell into the cytoplasm? What happens in cases where the *ras* gene is mutated?
12. If a cell suffers damage to its DNA while in S phase, how can this damage be repaired before the cell enters mitosis?

13. Distinguish between oncogenes and proto-oncogenes. In what ways can proto-oncogenes be converted to oncogenes?
14. Of the two classes of genes associated with cancer, tumor-suppressor genes and oncogenes, mutations in which group can be considered gain-of-function mutations? In which group are the loss-of-function mutations? Explain.
15. How do translocations such as the Philadelphia chromosome contribute to cancer?
16. Given that cancers can be environmentally induced and that some environmental factors are the result of lifestyle choices such as smoking, sun exposure, and diet, what percentage of the money spent on cancer research do you think should be devoted to research and education on preventing cancer rather than on finding cancer cures?
17. What are the most significant environmental agents that contribute to human cancers?
18. Explain the role of *p53* protein in protecting normal cells against cancer. With respect to this protein and its function, explain how a normal cell turns cancerous.
19. What is loss of heterozygosity and how does this process contribute to the development of cancers?
20. Mention the causative agents of DNA lesions in the human body that can lead to cancer.
21. Radiotherapy (treatment with ionizing radiation) is one of the most effective current cancer treatments. It works by damaging DNA and other cellular components. In which ways could radiotherapy control or cure cancer, and why does radiotherapy often have significant side effects?
22. Genetic tests that detect mutations in the *BRCA1* and *BRCA2* oncogenes are widely available. These tests reveal a number of mutations in these genes—mutations that have been linked to familial breast cancer. Assume that a young woman in a suspected breast cancer family takes the *BRCA1* and *BRCA2* genetic tests and receives negative results. That is, she does not test positive for the mutant alleles of *BRCA1* or *BRCA2*. Can she consider herself free of risk for breast cancer?
23. Explain the connection between DNA methylation and cancer.
24. While all cancer cells are proliferative, only some become malignant. Explain this statement.
25. As part of a cancer research project, you have discovered a gene that is mutated in many metastatic tumors. After determining the DNA sequence of this gene, you compare the sequence with those of other genes in the human genome sequence database. Your gene appears to code for an amino acid sequence that resembles sequences found in some serine proteases. Conjecture how your new gene might contribute to the development of highly invasive cancers.
26. A study by Bose and colleagues (1998. *Blood* 92: 3362–3367) and a previous study by Biernaux and others (1996. *Bone Marrow Transplant* 17: (Suppl. 3) S45–S47) showed that *BCR-ABL* fusion gene transcripts can be detected in 25 to 30 percent of healthy adults who do not develop chronic myelogenous leukemia (CML). Explain how these individuals can carry a fusion gene that is transcriptionally active and yet do not develop CML.
27. Those who inherit a mutant allele of the *RB1* gene are at risk for developing a bone cancer called osteosarcoma. You suspect that in these cases, osteosarcoma requires a mutation in the second *RB1* allele, and you have cultured some osteosarcoma cells and obtained a cDNA clone of a normal human *RB1* gene. A colleague sends you a research paper revealing that a strain of cancer-prone mice develops malignant tumors when injected with osteosarcoma cells, and you obtain these mice. Using these three resources, what experiments would you perform to determine (a) whether osteosarcoma cells carry two *RB1* mutations, (b) whether osteosarcoma cells produce any pRB protein, and (c) if the addition of a normal *RB1* gene will change the cancer-causing potential of osteosarcoma cells?
28. The following table shows neutral polymorphisms found in control families (those with no increased frequency of breast and ovarian cancer). Examine the data in the table and answer the following questions:
    (a) What is meant by a neutral polymorphism?
    (b) What is the significance of this table in the context of examining a family or population for *BRCA1* mutations that predispose an individual to cancer?
    (c) Is the PM2 polymorphism likely to result in a neutral missense mutation or a silent mutation?
    (d) Answer part (c) for the PM3 polymorphism.

Neutral Polymorphisms in *BRCA1*

			Frequency in Control Chromosomes*			
Name	Codon Location	Base in Codon†	A	C	G	T
PM1	317	2	152	0	10	0
PM6	878	2	0	55	0	100
PM7	1190	2	109	0	53	0
PM2	1443	3	0	115	0	58
PM3	1619	1	116	0	52	0

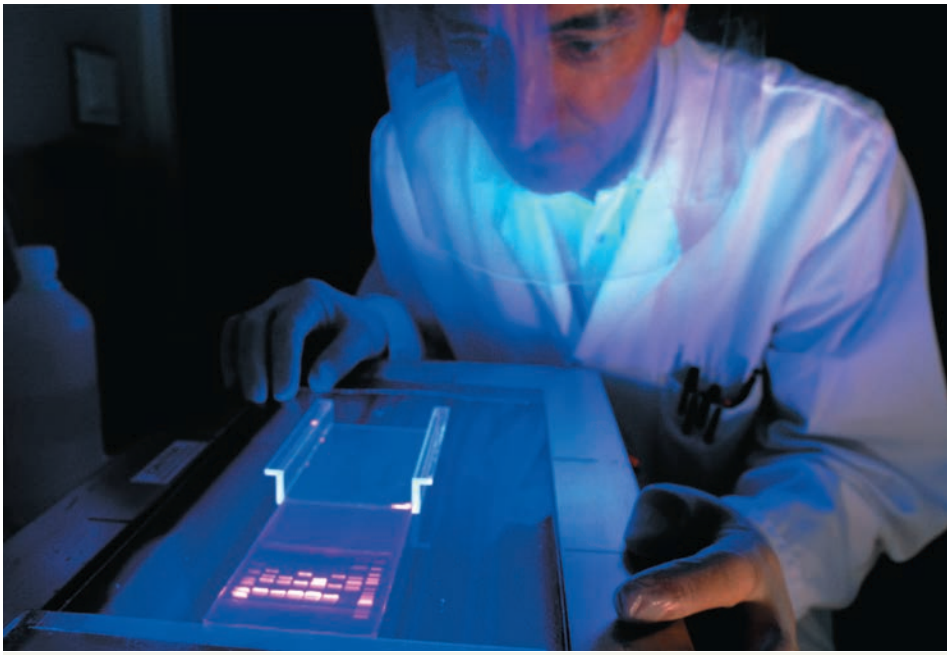
*The number of chromosomes with a particular base at the indicated polymorphic site (A, C, G, or T) is shown.

†Position 1, 2, or 3 of the codon.

17 Recombinant DNA Technology

CHAPTER CONCEPTS

- Recombinant DNA technology creates combinations of DNA sequences from different sources.
- A common application of recombinant DNA technology is to clone a DNA segment of interest.
- Specific DNA segments are inserted into vectors to create recombinant DNA molecules that are transferred into eukaryotic or prokaryotic host cells, where the recombinant DNA replicates as the host cells divide.
- DNA libraries are collections of cloned DNA and were historically used to isolate specific genes.
- DNA segments can be quickly amplified millions of times using the polymerase chain reaction (PCR).
- DNA, RNA, and proteins can be analyzed using a range of molecular techniques.
- Sequencing reveals the nucleotide composition of DNA, and major improvements in sequencing technologies have rapidly advanced many areas of modern genetics research, particularly genomics.
- Gene knockout methods and transgenic animals have become invaluable for studying gene function *in vivo*.
- Recombinant DNA technology has revolutionized our ability to investigate the genomes of diverse species and has led to the modern revolution in genomics.



A researcher examines an agarose gel containing separated DNA fragments stained with the DNA-binding dye ethidium bromide and visualized under ultraviolet light.

Researchers of the mid- to late 1970s developed various techniques to create, replicate, and analyze **recombinant DNA** molecules—DNA created by joining together pieces of DNA from different sources. The methods used to copy or **clone** DNA, called **recombinant DNA technology** and often known as "gene splicing" in the early days, marked a major advance in research in molecular biology and genetics, allowing scientists to isolate and study specific DNA sequences. For their contributions to the development of this technology, Daniel Nathans, Hamilton Smith, and Werner Arber were awarded the 1978 Nobel Prize in Physiology or Medicine.

The power of recombinant DNA technology is astonishing, enabling geneticists to identify and isolate a single gene or DNA segment of interest from a genome. Through cloning, large quantities of identical copies of this specific DNA molecule can be produced. These identical copies, or clones, can then be manipulated for numerous purposes, including conducting research on the structure and organization of the DNA, studying gene expression, studying protein products to understand their structure and function, and producing important commercial products from the protein encoded by a gene. The fundamental techniques involved in recombinant DNA technology subsequently led to the field of genomics, enabling scientists to sequence and analyze entire genomes. Note that some of the topics discussed in this chapter are explored in greater depth later in the text (see Special Topic Chapters 3—DNA Forensics, 5—Genetically Modified Foods, and 6—Gene Therapy). In this chapter, we review basic methods of recombinant DNA technology used to isolate, replicate, and analyze DNA.

17.1 Recombinant DNA Technology Began with Two Key Tools: Restriction Enzymes and DNA Cloning Vectors

Although natural genetic processes such as crossing over produce recombined DNA molecules, the term *recombinant DNA* is generally reserved for molecules produced by artificially joining DNA obtained from different sources. We begin our discussion of recombinant DNA technology by considering two important tools used to construct and amplify recombinant DNA molecules: DNA-cutting enzymes called **restriction enzymes** and **DNA cloning vectors**. The use of restriction enzymes and cloning vectors was largely responsible for advancing the field of molecular biology because a wide range of laboratory techniques are based on recombinant DNA technology.

Restriction Enzymes Cut DNA at Specific Recognition Sequences

Restriction enzymes are produced by bacteria as a defense mechanism against infection by bacteriophage. They restrict or prevent viral infection by degrading the DNA of invading viruses. More than 3500 restriction enzymes have been identified, and over 250 are commercially produced and available for use by researchers. A restriction enzyme recognizes and binds to DNA at a specific nucleotide sequence called a **recognition sequence** or **restriction site** (Figure 17–1). The enzyme then cuts both strands of the DNA within that sequence by cleaving the phosphodiester backbone of DNA. Scientists commonly refer to this as "digestion" of DNA. The usefulness of restriction enzymes in cloning derives from their ability to accurately and reproducibly cut genomic DNA into fragments. Restriction enzymes represent sophisticated molecular scissors for cutting DNA into fragments of desired sizes. Restriction sites are present randomly in the genome. The actual fragment sizes produced by digestion with a given restriction enzyme vary based on the number and location of sites throughout a DNA sample.

Recognition sequences exhibit a form of symmetry described as a *palindrome*: the nucleotide sequence reads the same on both strands of the DNA when read in the 5′ to 3′ direction. Each restriction enzyme recognizes its particular recognition sequence and cuts the DNA in a characteristic cleavage pattern. The most common recognition sequences are four or six nucleotides long, but some contain eight or more nucleotides. Enzymes such as *Eco*RI and *Hin*dIII make offset cuts in the DNA strands, thus producing fragments with single-stranded overhanging ends called *cohesive ends* (or "sticky" ends), while others such as *Alu*I and *Bal*I cut both strands at the same nucleotide pair, producing DNA fragments with double-stranded ends called *blunt-end* fragments. Four common restriction enzymes, their restriction sites, and source microbes are shown in Figure 17—1.

One of the first restriction enzymes to be identified was isolated from *Escherichia coli* strain R and was designated *Eco*RI. DNA fragments produced by *Eco*RI digestion (Figure 17–2) have cohesive ends because they can base-pair with complementary single-stranded ends on other DNA fragments cut using *Eco*RI. When mixed together, single-stranded ends of DNA fragments from different sources

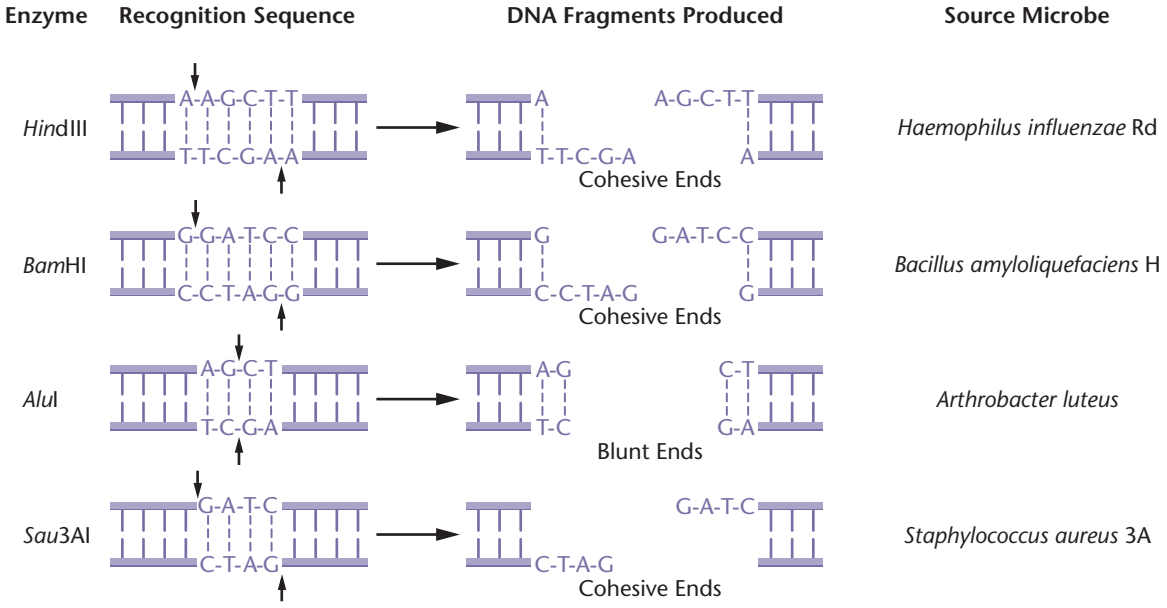


FIGURE 17–1 Common restriction enzymes, with their recognition sequence, DNA cutting patterns, and sources. Arrows indicate the location in the DNA cut by each enzyme.

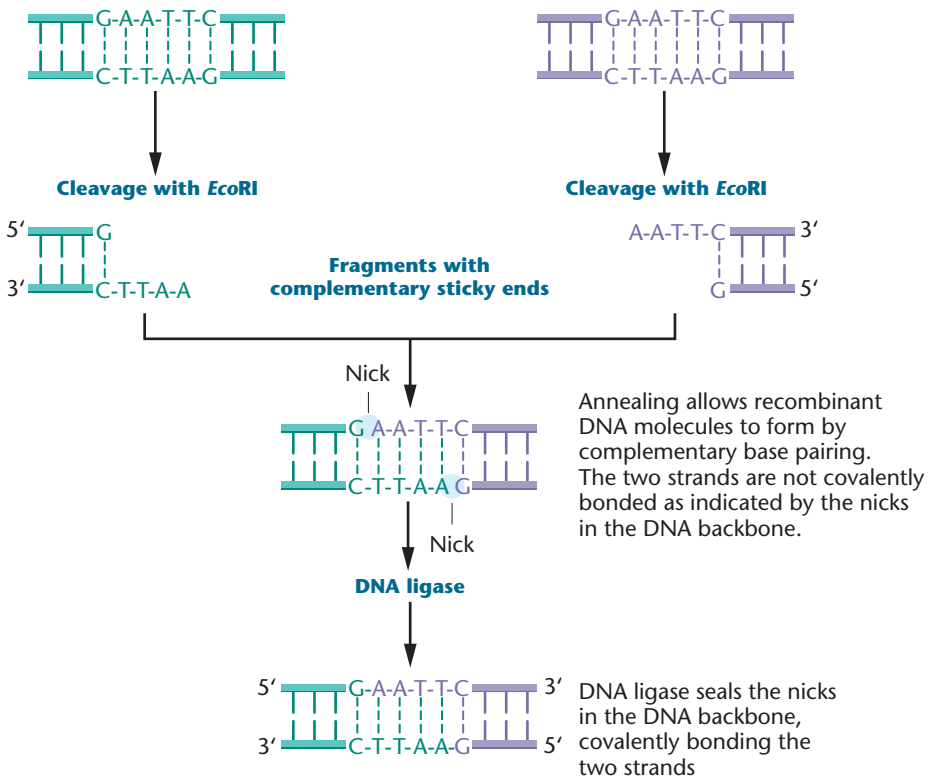


FIGURE 17–2 DNA from different sources is cleaved with *Eco*RI and mixed to allow annealing. The enzyme DNA ligase forms phosphodiester bonds between these fragments to create an intact recombinant DNA molecule.

cut with the same restriction enzyme can **anneal,** or stick together, by hydrogen bonding of complementary base pairs in single-stranded ends. Addition of the enzyme **DNA ligase**—recall the role of DNA ligase in DNA replication as discussed earlier in the text (see Chapter 10)—to DNA fragments will seal the phosphodiester backbone of DNA to covalently join the fragments together to form recombinant DNA molecules (Figure 17–2).

Scientists often use restriction enzymes that create cohesive ends since the overhanging ends make cloning less technically challenging. Blunt-end ligation is more technically challenging because it is not facilitated by hydrogen bonding, but a scientist can ligate fragments digested by different blunt-end generating enzymes.

ESSENTIAL POINT

Recombinant DNA technology was made possible by the discovery of proteins called restriction enzymes, which cut DNA at specific sequences, producing fragments that can be joined with other DNA fragments to form recombinant DNA molecules. ■

DNA Vectors Accept and Replicate DNA Molecules to Be Cloned

Scientists recognized that DNA fragments produced by restriction-enzyme digestion could be copied or cloned if they had a technique for replicating the fragments. The second key tool that allowed DNA cloning was the development of **cloning vectors,** DNA molecules that accept DNA fragments and replicate these fragments when vectors are introduced into host cells.

Many different vectors are available for cloning. Vectors differ in terms of the host cells they can enter and replicate in and in the size of DNA fragment inserts they can carry, but most DNA vectors have several key properties.

- A vector contains several restriction sites that allow insertion of the DNA fragments to be cloned.
- Vectors must be capable of replicating in host cells to allow for independent replication of the vector DNA and any DNA fragment it carries.
- To distinguish host cells that have taken up vectors from host cells that have not, the vector contains a **selectable marker gene** (usually an antibiotic resistance gene or a gene that encodes a protein which produces a visible product, such as color or fluorescent light).
- Most vectors incorporate specific sequences that allow for sequencing inserted DNA.

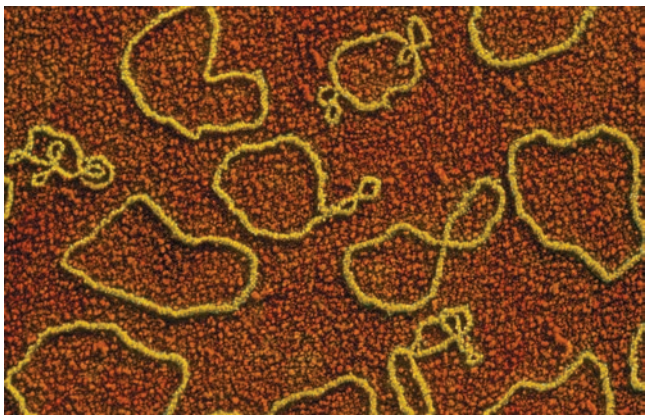
Bacterial Plasmid Vectors

Genetically modified bacterial **plasmids** were the first vectors developed, and they are still widely used for cloning. Plasmid cloning vectors were derived from naturally occurring plasmids. Recall from Chapter 8 that plasmids are extrachromosomal, double-stranded DNA molecules that

replicate independently from the chromosomes within bacterial cells [**Figure 17–3(a)**]. Plasmids have been extensively modified by genetic engineering to serve as cloning vectors. Many commercially prepared plasmids are readily available with a range of useful features [**Figure 17–3(b)**]. Plasmids are introduced into bacteria by the process of **transformation** (see Chapter 8). Two main techniques are widely used for bacterial transformation. One approach involves treating cells with calcium ions and using a brief heat shock to pulse DNA into cells. The other technique, called **electroporation,** uses a brief, but high-intensity, pulse of electricity to move DNA into bacterial cells. Only one or a few plasmids generally enter a bacterial host cell by transformation.

Because plasmids have an *origin of replication* (*ori*) that allows for plasmid replication, many plasmids can increase their copy number to produce several hundred copies in a single host cell. These plasmids greatly enhance the number of DNA clones that can be produced. Plasmid vectors have also been genetically engineered to contain a number of restriction sites for commonly used restriction enzymes in a region called the *multiple cloning site*. Multiple cloning sites allow scientists to clone a range of different fragments generated by many commonly used restriction enzymes.

(a)



(b)

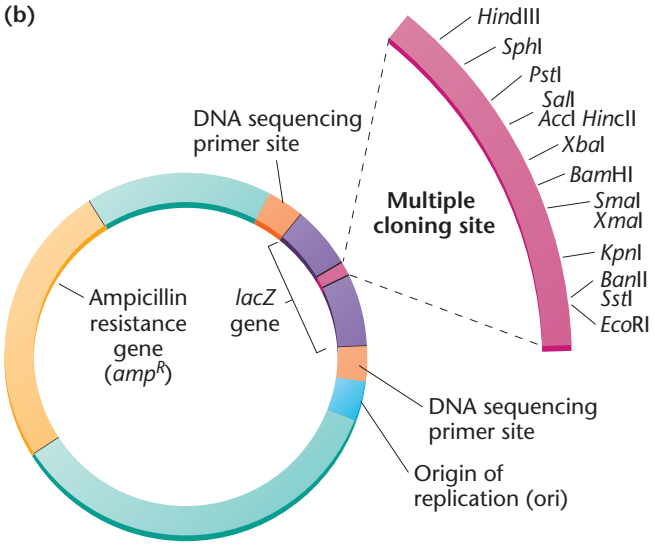


FIGURE 17–3 (a) A color-enhanced electron micrograph of plasmids isolated from *E. coli.* (b) A diagram of a typical DNA cloning plasmid.

Cloning DNA with a plasmid generally begins by cutting both the plasmid DNA and the DNA to be cloned with the same restriction enzyme (**Figure 17–4**). Typically, the plasmid is cut once within the multiple cloning site to produce a linear vector. DNA restriction fragments from the DNA to be cloned are added to the linearized vector in the presence of DNA ligase. Sticky ends of DNA fragments anneal, joining the DNA to be cloned and the plasmid. DNA ligase is then used to create phosphodiester bonds to seal nicks in the DNA backbone, thus producing recombinant DNA, which is then introduced into bacterial host cells by transformation. Once inside the cell, plasmids replicate quickly to produce multiple copies.

However, when cloning DNA using plasmids, not all plasmids will incorporate DNA to be cloned. For example, a plasmid cut with a particular restriction enzyme can close back on itself (self-ligation) if cut ends of the plasmid rejoin. Obviously then, such nonrecombinant plasmids are not desired. Also, during transformation, not all host cells will take up plasmids. Therefore it is important that bacterial cells containing recombinant DNA can be readily identified in a cloning experiment. One way this is accomplished is through the use of selectable marker genes described earlier. Genes that provide resistance to antibiotics such as ampicillin and genes such as the *lacZ* gene are very effective selectable marker genes. **Figure 17–5** provides an example of how these genes can be used to select for and identify bacteria containing recombinant plasmids. This process is referred to as **"blue-white" screening** for a reason that will soon become obvious. In blue-white screening a plasmid is used that contains the *lacZ* gene incorporated into the multiple cloning site. The *lacZ* gene encodes the enzyme β-galactosidase, which, as you learned earlier in the text (see Chapter 15), is used to cleave the disaccharide lactose into its component monosaccharides glucose and galactose. Blue-white screening takes advantage of the enzymatic activity of β-galactosidase.

Using this approach, one can easily identify transformed bacterial cells containing recombinant or nonrecombinant plasmids. If a DNA fragment is inserted anywhere in the multiple cloning site, the *lacZ* gene is disrupted and will not produce functional copies of β-galactosidase. Transformed bacteria in this experiment are plated on agar plates that contain an antibiotic—ampicillin in this case. Nontransformed bacteria cannot grow well on these plates because they do not have the amp^R gene and so the ampicillin kills these cells. These agar plates also contain a substance called X-gal (technically 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside). X-gal is similar to lactose in structure. It is a substrate for β-galactosidase, and when it is

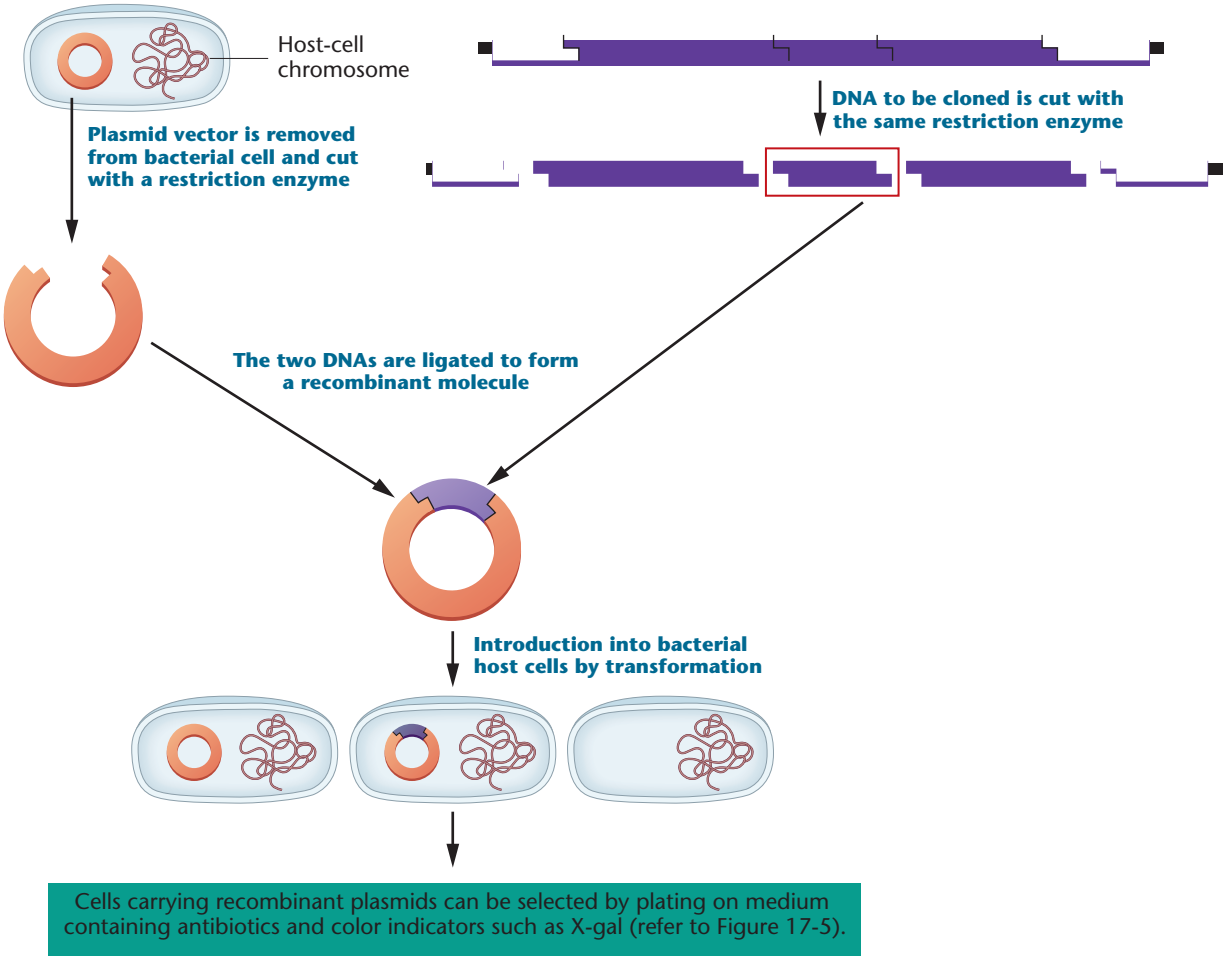


FIGURE 17–4 Cloning with a plasmid vector involves cutting both plasmid and the DNA to be cloned with the same restriction enzyme. The DNA to be cloned is ligated into the vector and transferred to a bacterial host for replication. Bacterial cells carrying plasmids with DNA inserts can be identified by selection and then isolated. The cloned DNA is then recovered from the bacterial host for further analysis.

cleaved by β-galactosidase it turns blue. As a result, bacterial cells carrying nonrecombinant plasmids (those that have self-ligated and thus do not contain inserted DNA) have a functional *lacZ* gene and produce β-galactosidase, which cleaves X-gal in the medium, and these cells turn blue. However, recombinant bacteria with plasmids containing an inserted DNA fragment will form white colonies when they grow on X-gal medium because the plasmids in these cells are not producing functional β-galactosidase (Figure 17–5). Bacteria in these white colonies are clones of each other—genetically identical cells with copies of recombinant plasmids. White colonies can be transferred to flasks of bacterial culture broth and grown in large quantities, after which it is relatively easy to isolate and purify recombinant plasmids from these cells.

Plasmids are still the workhorses for many applications of recombinant DNA technology, but they have a major limitation: because they are small, they can only accept inserted pieces of DNA up to about 25 kilobases (kb) in size, and most plasmids can often only accept substantially smaller pieces. Therefore, as recombinant DNA technology has developed and it has become desirable to clone large pieces of DNA, other vectors have been developed primarily for their ability to accept larger pieces of DNA and because they can be used with other types of host cells beside bacteria.

Other Types of Cloning Vectors

Phage vector systems were among the earliest vectors used in addition to plasmids. These included genetically modified strains of bacteriophage λ. Phage vectors were popular for quite some time because they can carry inserts up to 45 kb, more than twice as long as DNA inserts in most plasmid vectors. DNA fragments are ligated into the phage vector to produce recombinant λ vectors that are subsequently packaged into phage protein heads *in vitro* and introduced into bacterial host cells growing on petri plates. Inside the bacteria, the vectors replicate and form many copies of infective phage, each of which carries a DNA insert. As they reproduce, they lyse their bacterial host cells, forming the clear spots known as plaques (described in Chapter 8),

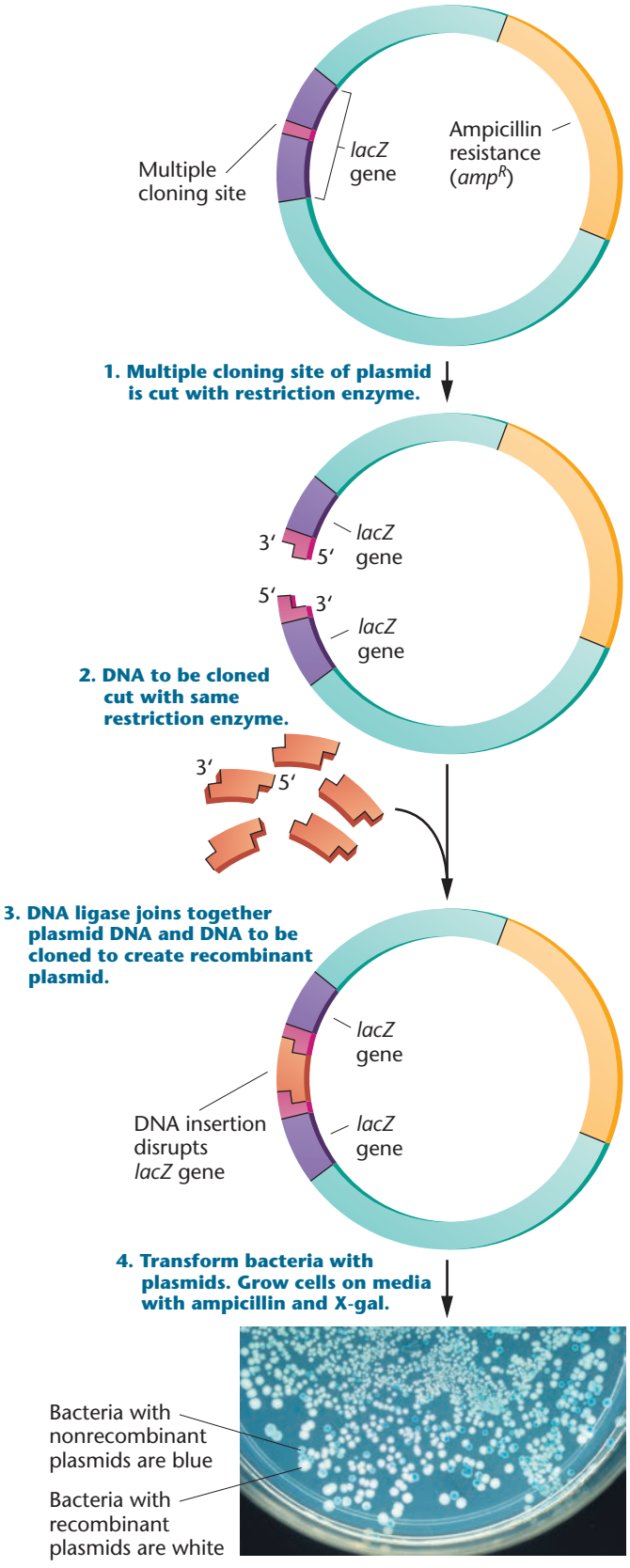


FIGURE 17–5 In blue-white screening, DNA inserted into multiple cloning site of a plasmid disrupts the *lacZ* gene so that bacteria containing recombinant DNA are unable to metabolize X-gal, resulting in white colonies that allow direct identification of bacterial colonies carrying cloned DNA inserts. Photo of a petri dish showing the growth of bacterial cells after uptake of recombinant plasmids. Cells in blue colonies contain vectors without cloned DNA inserts, whereas cells in white colonies contain vectors carrying DNA inserts.

from which phage can be isolated and the cloned DNA can be recovered.

Bacterial artificial chromosomes (**BACs**) and **yeast artificial chromosomes** (**YACs**) are two other examples of vectors that can be used to clone large fragments of DNA. For example, the mapping and analysis of large eukaryotic genomes such as the human genome required cloning vectors that could carry very large DNA fragments such as segments of an entire chromosome. BACs are essentially very large but low copy number (typically one or two copies/bacterial cell) plasmids that can accept DNA inserts in the 100- to 300-kb range. Like natural chromosomes, a YAC has telomeres at each end, origins of replication, and a centromere. Yeast chromosomes range in size from 230 kb to over 1900 kb, making it possible to clone DNA inserts from 100 to 1000 kb in YACs.

Unlike the vectors described so far, **expression vectors** are designed to ensure mRNA expression of a cloned gene with the purpose of producing many copies of the gene's encoded protein in a host cell. Expression vectors are available for prokaryotic and eukaryotic host cells and contain the appropriate sequences to initiate both transcription and translation of the cloned gene. For many research applications that involve studies of protein structure and function, producing a recombinant protein in bacteria (or other host cells) and purifying the protein is a routine approach, although it is not always easy to properly express a protein that maintains its biological function. The biotechnology industry also relies heavily on expression vectors to produce commercially valuable protein products from cloned genes, a topic we will discuss later in the text (see Chapter 19).

Introducing genes into plants is a common application that can be done in many ways, and we will discuss aspects of genetic engineering of food plants later in the text (see Special Topic Chapter 5—Genetically Modified Foods). One widely used approach to insert genes into plant cells involves the soil bacterium *Rhizobium radiobacter,* which infects plant cells and produces tumors (called crown galls) in many species of plants. Formerly *Agrobacterium tumefaciens*, this bacterium was renamed based on genomic analysis.

Rhizobium contains a plasmid called the **Ti plasmid** (tumor-inducing). Restriction sites in Ti plasmids can be used to insert foreign DNA, and recombinant vectors are introduced into *Rhizobium* by transformation. Tumor-inducing genes from Ti plasmids are removed from the vector so that the recombinant vector does not result in tumor production. *Rhizobium* containing recombinant DNA is mixed with plant cells (not all types of plant cells can be infected by *Rhizobium*).

Once inside the cell, the plasmid is integrated into a chromosome of the host cell. Plant cells carrying a recombinant Ti plasmid can be grown in tissue culture. The presence of certain compounds in the culture medium in which plant cells are grown stimulates the formation of roots and shoots, and eventually a mature plant carrying a foreign gene.

ESSENTIAL POINT

Vectors replicate autonomously in host cells and facilitate the cloning and manipulation of newly created recombinant DNA molecules. ■

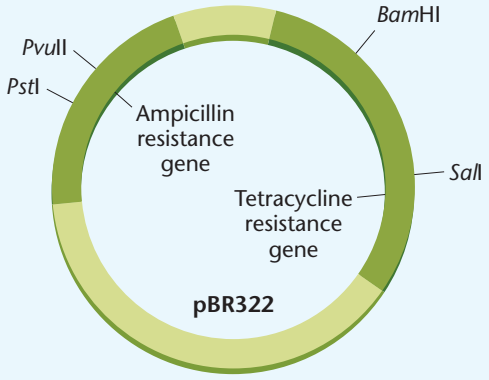
NOW SOLVE THIS

17–1 An ampicillin-resistant, tetracycline-resistant plasmid, pBR322, is cleaved with *Pst*I, which cleaves within the ampicillin resistance gene. The cut plasmid is ligated with *Pst*I-digested *Drosophila* DNA to prepare a genomic library, and the mixture is used to transform *E. coli* K12.

(a) Which antibiotic should be added to the medium to select cells that have incorporated a plasmid?

(b) If recombinant cells were plated on medium containing ampicillin or tetracycline and medium with both antibiotics, on which plates would you expect to see growth of bacteria containing plasmids with *Drosophila* DNA inserts?

(c) How can you explain the presence of colonies that are resistant to both antibiotics?



■ **HINT:** *This problem involves an understanding of antibiotic selectable marker genes in plasmids and antibiotic DNA selection for identifying bacteria transformed with recombinant plasmid DNA. The key to its solution is to recognize that inserting foreign DNA into the plasmid vector disrupts one of the antibiotic resistance genes in the plasmid.*

17.2 DNA Libraries Are Collections of Cloned Sequences

Only relatively small DNA segments—representing just a single gene or even a portion of a gene—are produced by cloning DNA into vectors, particularly plasmids. In the cloning discussions we have had so far, we have described *how* DNA can be inserted into vectors and cloned—a relatively straightforward process—but we have not discussed how one knows what particular DNA sequence they have cloned. Simply cutting DNA and inserting into vectors does not tell you what gene or sequences have been cloned.

During the first several decades of DNA cloning, scientists created **DNA libraries,** which represent a collection of cloned DNA. Depending on how a library is constructed, it may contain genes and noncoding regions of DNA. Generally, there are two main types of libraries, genomic DNA libraries and complementary DNA (cDNA) libraries.

Genomic Libraries

Ideally, a **genomic library** consists of many overlapping fragments of the genome, with at least one copy of every DNA sequence in an organism's genome, which in summary span the entire genome. In making a genomic library, DNA is extracted from cells or tissues and cut randomly with restriction enzymes, and the resulting fragments are inserted into vectors using techniques that we discussed in the previous section. Since some vectors (such as plasmids) can carry only a few thousand base pairs of inserted DNA, selecting the vector so that the library contains the whole genome in the smallest number of clones is an important consideration. Because genomic DNA is the foreign DNA introduced into vectors, genomic libraries contain coding and noncoding segments of DNA such as introns, and vectors in the library may contain more than one gene or only a portion of a gene.

As you will learn later in the text (see Chapter 18), **whole-genome shotgun cloning** approaches (see Figure 18–1) and new sequencing methodologies are readily replacing traditional genomic DNA libraries because they effectively allow one to sequence an entire genomic DNA sample without the need for inserting DNA fragments into vectors and cloning them in host cells. Later in the text (see Chapter 18), we will also consider how DNA sequence analysis using bioinformatics allows one to identify protein-coding and noncoding sequences in cloned DNA.

Complementary DNA (cDNA) Libraries

Complementary DNA (cDNA) libraries offer certain advantages over genomic libraries and continue to be a useful methodology for gene cloning. This is primarily because a cDNA library contains DNA copies which are made from mRNA molecules isolated from cultured cells or a tissue sample. cDNA is complementary to the nucleotide sequence of the mRNA, and so unlike a genomic library, which contains all of the DNA in a genome—gene coding and noncoding sequences—a cDNA library contains only expressed genes. As a result, cDNA libraries

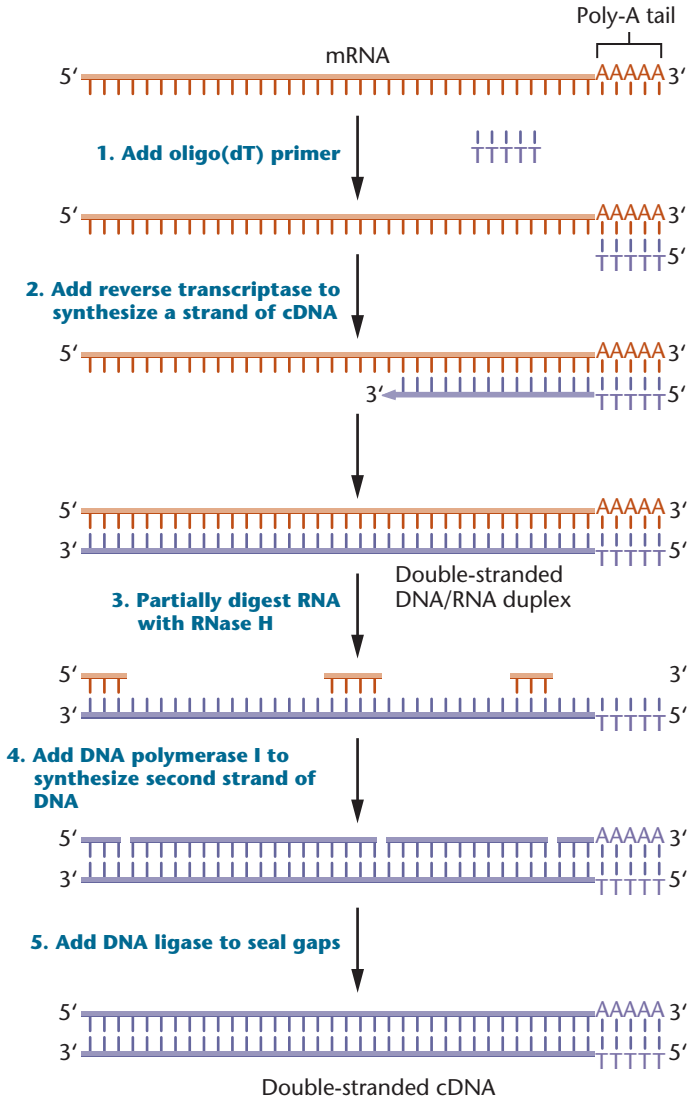


FIGURE 17–6 Producing cDNA from mRNA. Because most eukaryotic mRNAs have a poly-A tail at the 3′ end, a short oligo(dT) molecule annealed to this tail serves as a primer for the enzyme reverse transcriptase. The enzyme reverse transcriptase uses the mRNA as a template to synthesize a complementary DNA strand (cDNA) and forms an mRNA/cDNA double-stranded duplex. The mRNA is digested with the enzyme RNAse H, producing gaps in the RNA strand. The 3′ ends of the remaining RNA serve as primers for DNA polymerase I, which synthesizes a second DNA strand. The result is a double-stranded cDNA molecule that can be cloned into a suitable vector.

have been particularly useful for identifying and studying genes expressed in certain cells or tissues under certain conditions: for example, during development, cell death, cancer, and other biological processes. One can also use these libraries to compare expressed genes from normal tissues and diseased tissues. For instance, this approach has been used to identify genes involved in cancer formation, such as those genes that contribute to progression from a normal cell to a cancer cell and genes involved in cancer cell metastasis (spreading).

Preparation of a cDNA library is shown in Figure 17–6. These libraries provide a snapshot of the genes that were transcriptionally active in a tissue at a particular time because the relative amount of cDNA in a particular library is equivalent to the amount of starting mRNA isolated from the tissue and used to make the library. Because cDNA libraries provide a catalog of all the genes active in a cell at a specific time, they have been very valuable tools for scientists isolating and studying genes in particular tissues.

Specific Genes Can Be Recovered from a Library by Screening

Genomic and cDNA libraries often consist of several hundred thousand different DNA clones, much like a large book library may have many books but only a few of interest to your studies in genetics. So how can libraries be used to locate a specific gene of interest in a library? To find a specific gene, we need to identify and isolate only the clone or clones containing that gene. We must also determine whether a given clone contains all or only part of the gene we are studying. Several methods allow us to sort through a library and isolate specific genes of interest, and this approach is called **library screening.** The choice of method often depends on available information about the gene being sought. The process for screening a library with a probe is shown in Figure 17–7.

Often, probes are used to screen a library to recover clones of a specific gene. A **probe** is any DNA or RNA sequence that is complementary to some part of a cloned sequence present in the library—the target gene or sequence to be identified. A probe must be labeled or tagged in different ways so that it can be identified. When a probe is used in a **hybridization** reaction, the probe binds to any complementary DNA sequences present in one or more clones. Probes can be labeled with radioactive isotopes, or increasingly these days probes are labeled with nonradioactive compounds that undergo chemical or color reactions to indicate the location of a specific clone in a library.

Probes are derived from a variety of sources—often related genes isolated from another species can be used if enough of the DNA sequence is conserved. For example, genes from rats, mice, or even *Drosophila* that have conserved sequence similarity to human genes can be used as probes to identify human genes during library screening.

ESSENTIAL POINT

DNA libraries are collections of cloned DNA that can be screened to identify and isolate specific sequences of interest. ■

As we have discussed here, libraries enable scientists to clone DNA and then identify individual genes in the library.

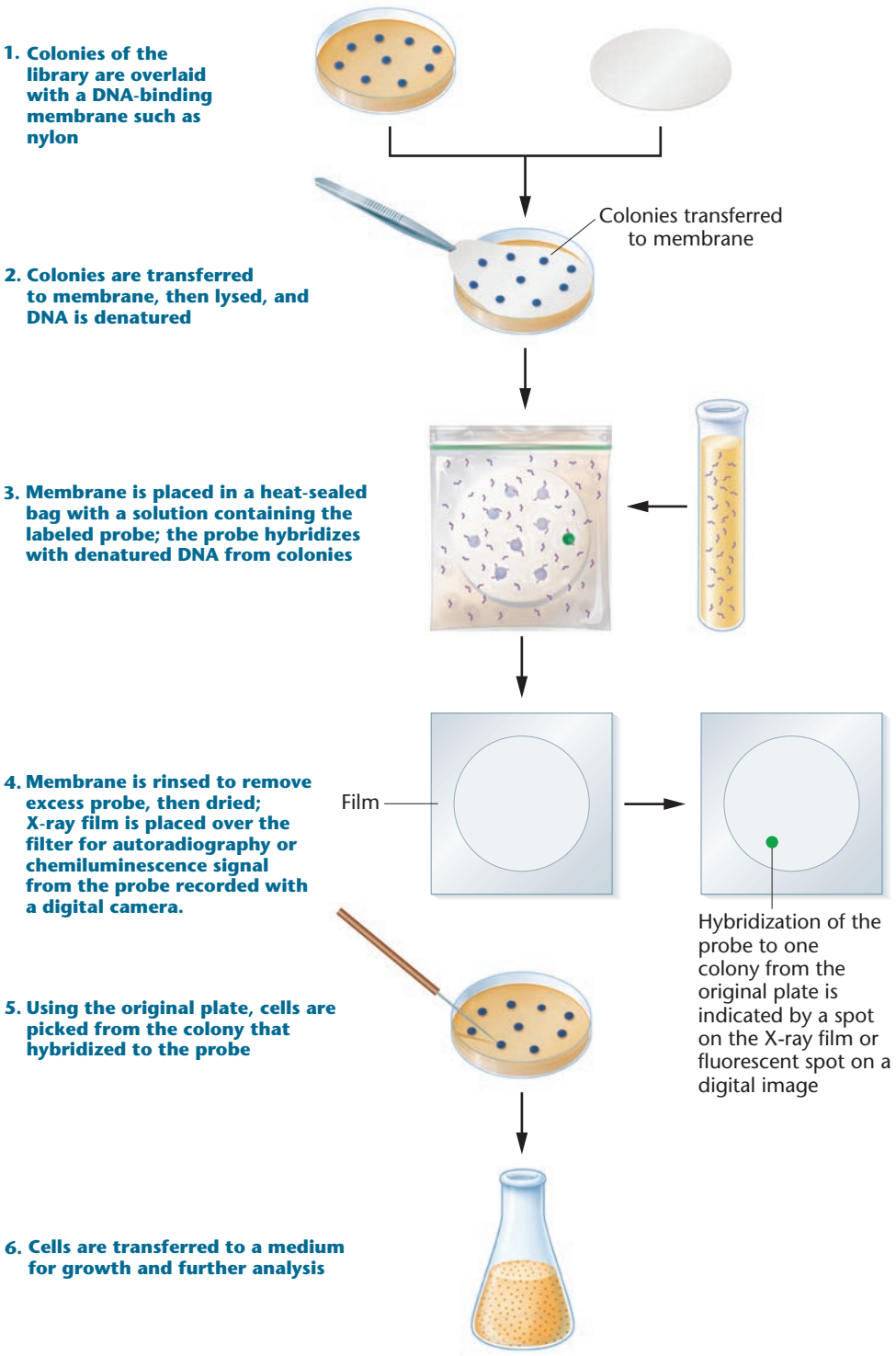


FIGURE 17–7 Screening a library to recover a specific gene. The library, present in bacteria on petri plates, is overlaid with a DNA-binding membrane, and colonies are transferred to the membrane. Colonies on the membrane are lysed, and the DNA is denatured to single strands. The membrane is placed in a hybridization bag along with buffer and a labeled single-stranded DNA probe. During incubation, the probe forms a double-stranded hybrid with any complementary sequences on the membrane. The membrane is removed from the bag and washed to remove excess probe. Hybrids are detected by placing a piece of X-ray film over the membrane and exposing it for a short time or by chemiluminescence detection and image capture using a digital camera. Colonies containing the cloned DNA that hybridized to the probe are identified from the orientation of the spots. Cells are picked from this colony for growth and further analysis.

Cloning DNA from libraries is still a technique with valuable applications. However, as you will learn later in the text (see Chapter 18), the basic methods of recombinant DNA technology were the foundation for the development of powerful techniques for whole-genome cloning and sequencing, which led to the **genomics** era of modern genetics and molecular biology. Genomic techniques, in which entire genomes are being sequenced without creating libraries, are replacing many traditional recombinant DNA approaches that cloned or identified individual or a few genes at a time.

17.3 The Polymerase Chain Reaction Is a Powerful Technique for Copying DNA

Cloning DNA using vectors and host cells is labor intensive and time consuming. In 1986, another technique, called the **polymerase chain reaction** (**PCR**), became available to the scientific community. This advance revolutionized recombinant DNA methodology and further accelerated the pace of biological research. The significance of this method was underscored by the awarding of the 1993 Nobel Prize in Chemistry to Kary Mullis, who developed the technique.

PCR is a rapid method of DNA cloning that extends the power of recombinant DNA research and in many cases eliminates the need to use host cells for cloning. PCR is also a method of choice for many applications, whether in molecular biology, human genetics, evolution, development, conservation, or forensics.

By copying a specific DNA sequence through a series of *in vitro* reactions, PCR can amplify target DNA sequences that are initially present in very small quantities in a population of other DNA molecules. When using PCR to clone DNA, double-stranded target DNA to be amplified is placed in a tube with DNA polymerase, Mg^{2+} (as an important cofactor for DNA polymerase), and the four deoxyribonucleoside triphosphates. In addition, some information about the nucleotide sequence of the target DNA is required. This sequence information is used to synthesize two oligonucleotide **primers:** short (typically about 20 nt long) single-stranded DNA sequences, one complementary to the 5′ end of one strand of target DNA to be amplified and another primer complementary to the opposing strand of target DNA at its 3′ end. When added to a sample of double-stranded DNA that has been denatured into single strands, the primers bind to complementary nucleotides flanking the sequence to be cloned. DNA polymerase can then extend the 3′ end of each primer to synthesize second strands of the target DNA. Therefore, one complete reaction process, called a **cycle,** doubles the number of DNA molecules in the reaction (**Figure 17–8**). Repetition of the process produces large numbers of copied DNA very quickly. If desired, the PCR products can be cloned into plasmid vectors for further use.

Most routine PCR applications involve a series of three reaction steps in a cycle. These three steps are as follows:

1. **Denaturation:** The double-stranded DNA to be amplified is *denatured* into single strands by heating to 92–95°C for about 1 minute. The DNA can come from many sources, including genomic DNA, mummified remains, fossils, or forensic samples such as blood, semen, or hair.
2. **Hybridization/Annealing:** The temperature of the reaction is lowered to a temperature between 45°C and 65°C, which causes primer binding, also called hybridization or annealing, to the denatured, single-stranded DNA. The primers serve as starting points for DNA polymerase to synthesize new DNA strands complementary to the target DNA. Factors such as primer length, base composition of primers (GC-rich primers are more thermally stable than AT-rich primers), and whether or not all bases in a primer are complementary to bases in the target sequence are among primary considerations when selecting a hybridization temperature for an experiment.
3. **Extension:** The reaction temperature is adjusted to between 65°C and 75°C, and DNA polymerase uses the primers as a starting point to synthesize new DNA strands by adding nucleotides to the ends of the primers in a 5′ to 3′ direction.

PCR is a chain reaction because the number of new DNA strands is doubled in each cycle, and the new strands, along with the old strands, serve as templates in the next cycle. Each cycle takes 2 to 5 minutes and can be repeated immediately, so that in less than 3 hours, 25 to 30 cycles result in over a million-fold increase in the amount of DNA (Figure 17–8). This process is automated by instruments called *thermocyclers,* or simply PCR machines, that can be programmed to carry out a predetermined number of cycles.

A key requirement for PCR is the type of DNA polymerase used in PCR reactions. Multiple PCR cycles involve repetitive heating and cooling of samples, which eventually lead to heat denaturation and loss of activity of most proteins. PCR reactions rely on thermostable forms of DNA polymerase capable of withstanding multiple heating and cooling cycles without significant loss of activity. PCR became a major tool when DNA polymerase was isolated from *Thermus aquaticus,* a bacterium living in the hot springs of Yellowstone National Park. Called *Taq Polymerase,* this enzyme is capable of tolerating extreme temperature changes and was the first thermostable polymerase used for PCR.

PCR-based DNA cloning has several advantages over library cloning approaches. PCR is rapid and can be carried out in a few hours, rather than the days required for making and screening DNA libraries. PCR is also very sensitive and amplifies specific DNA sequences from vanishingly small DNA samples, including the DNA in a single cell. This

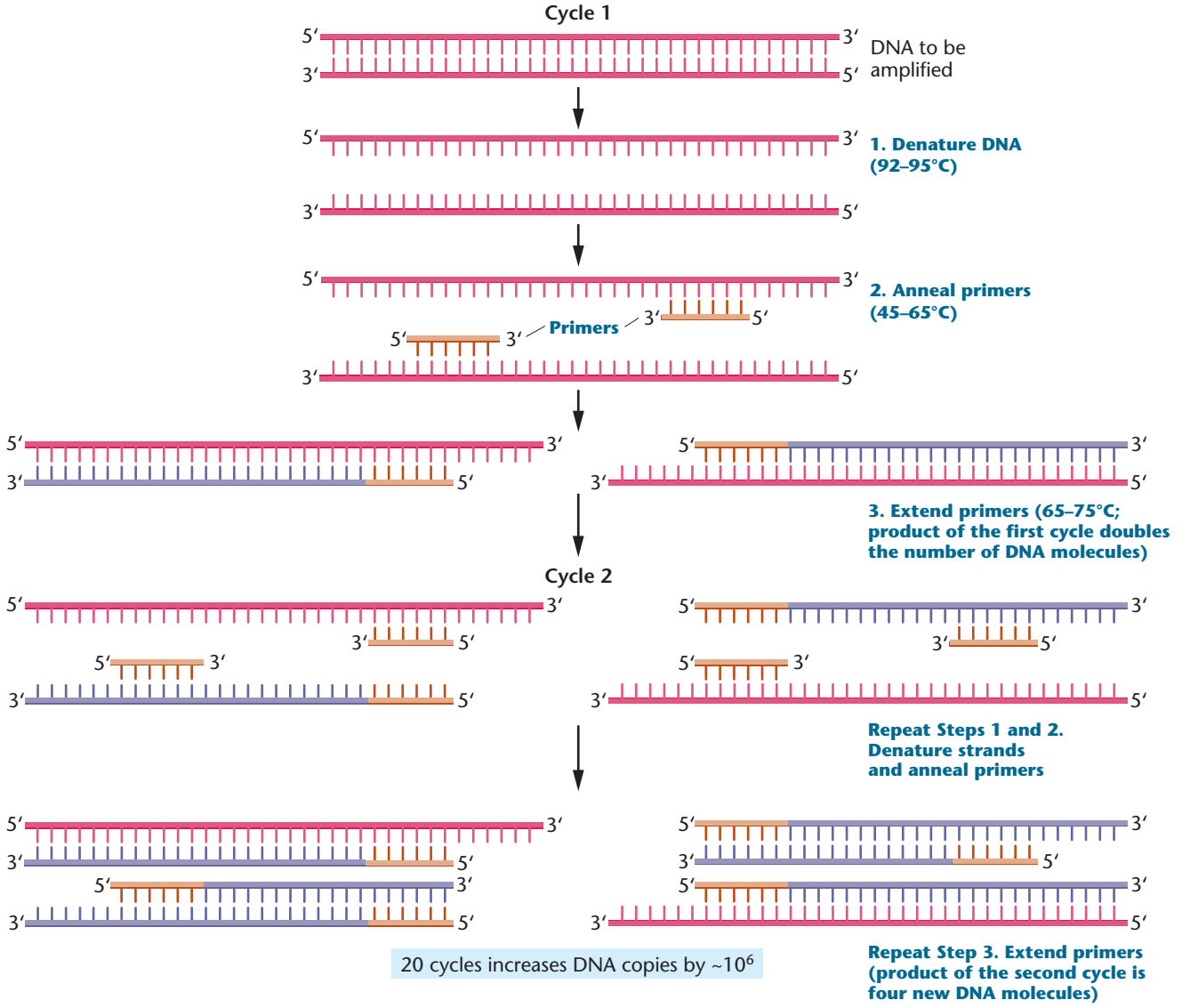


FIGURE 17–8 In the polymerase chain reaction (PCR), the target DNA is denatured into single strands; each strand is then annealed to short, complementary primers. DNA polymerase extends the primers in the 5′ to 3′ direction, using the single-stranded DNA as a template. The result after one round of replication is a doubling of DNA molecules to create two newly synthesized double-stranded DNA molecules. Repeated cycles of PCR can quickly amplify the original DNA sequence more than a millionfold. *Note:* shown here is a relatively short sequence of DNA being amplified. Typically, much longer segments of DNA are used for PCR, and the primers bind somewhere within the DNA molecule and not so close to the end of the actual molecule.

feature of PCR is invaluable in several kinds of applications, including genetic testing, forensics, and molecular paleontology. With carefully designed primers, DNA samples that have been partially degraded, contaminated with other materials, or embedded in a matrix (such as amber) can be recovered and amplified, when conventional cloning would be difficult or impossible. A wide variety of PCR-based techniques involve different variations of the basic technique described here.

Limitations of PCR

Although PCR is a valuable technique, it does have limitations: some information about the nucleotide sequence of the target DNA must be known in order to synthesize primers. In addition, even minor contamination of the sample with DNA from other sources can cause problems. For example, cells shed from a researcher's skin can contaminate samples gathered from a crime scene, making it difficult to obtain accurate results. PCR reactions must always be performed with carefully designed and appropriate controls. Also, PCR typically cannot amplify particularly long segments of DNA. DNA polymerase in a PCR reaction only extends primers for relatively short distances and does not continue processively until it reaches the other end of long template strands of DNA. Because of this characteristic, With deletions to text noted above, top

of the page should begin with Applications of PCR scientists often use PCR to amplify pieces of DNA that are several hundred to several thousand nucleotides in length, which is fine for most routine applications.

Applications of PCR

The PCR has been one of the most widely used techniques in genetics and molecular biology for over 20 years. PCR and its variations have many other applications as well. In short, PCR is one of the most versatile techniques in modern genetics. As you will learn later in the text (see Chapter 19), gene-specific primers provide a way of using PCR for screening mutations involved in genetic disorders, allowing the location and nature of a mutation to be determined quickly. Primers can be designed to distinguish between target sequences that differ by only a single nucleotide. This makes it possible to synthesize allele-specific probes for genetic testing; thus PCR is important for diagnosing genetic disorders. PCR is also a key diagnostic methodology for detecting bacteria and viruses (such as hepatitis or HIV) in humans, and pathogenic bacteria such as *E. coli* and *Staphylococcus aureus* in contaminated food.

PCR techniques are particularly advantageous when studying samples from single cells, fossils, or a crime scene, where a single hair or even a saliva-moistened postage stamp is the source of the DNA. Later in the text (see Special Topic Chapter 3—DNA Forensics), we will discuss how PCR is used in human identification, including remains identification, and in forensic applications. Using PCR, researchers can also explore uncharacterized DNA regions adjacent to known regions and even sequence DNA.

Reverse transcription PCR (**RT-PCR**) is a powerful methodology for studying gene expression, that is, mRNA production by cells or tissues. In RT-PCR, RNA is isolated from cells or tissues to be studied, and reverse transcriptase is used to generate double-stranded cDNA molecules, as described earlier when we discussed preparation of cDNA libraries. This reaction is followed by PCR to amplify cDNA with a set of primers specific for the gene of interest. Amplified cDNA fragments are then separated and visualized on an agarose gel. Because the amount of amplified cDNA in RT-PCR is based on the relative number of mRNA molecules in the starting reaction, RT-PCR can be used to evaluate relative levels of gene expression in different samples. The amplified cDNA can be inserted into plasmid vectors, which are replicated to produce a cDNA library. RT-PCR is more sensitive than conventional cDNA preparation and is a powerful tool for identifying mRNAs that may be present in only one or two copies per cell.

Finally, in discussing PCR approaches, one of the most valuable modern PCR techniques involves a method called **quantitative real-time PCR** (**qPCR**) or simply real-time PCR. This approach makes it possible to determine the amount of PCR product made during an experiment, which enables researchers to quantify amplification reactions as they occur in "real time" without having to run a gel.

ESSENTIAL POINT

PCR allows DNA to be amplified, or copied, without cloning and is a rapid and sensitive method with wide-ranging applications. ■

NOW SOLVE THIS

17–2 You have just created the world's first genomic library from the African okapi, a relative of the giraffe. No genes from this genome have been previously isolated or described. You wish to isolate the gene encoding the oxygen-transporting protein β-globin from the okapi library. This gene has been isolated from humans, and its nucleotide sequence and amino acid sequence are available in databases. Using the information available about the human β-globin gene, what two strategies can you use to isolate this gene from the okapi library?

■ **HINT:** *This problem asks you to design PCR primers to amplify the β-globin gene from a species whose genome you just sequenced. The key to its solution is to remember that you have at your disposal sequence data for the human β-globin gene and consider that PCR experiments require the use of primers that bind to complementary bases in the DNA to be amplified.*

For more practice, see Problems 15 and 16.

17.4 Molecular Techniques for Analyzing DNA

In addition to cloning and PCR methods, a wide range of molecular techniques are available to geneticists, molecular biologists, and almost anyone who does research involving DNA and RNA, particularly those who study the structure, expression, and regulation of genes. In the following sections, we consider some of the most commonly used molecular methods that provide information about the organization and function of cloned sequences. Throughout later sections of the text you will see these and other techniques discussed in the context of certain applications in modern genetics.

Restriction Mapping

Historically, one of the first steps in characterizing a DNA clone was the construction of a **restriction map.**

A restriction map establishes the number of, order of, and distances between restriction-enzyme cleavage sites along a cloned segment of DNA, thus providing information about the length of the cloned insert and the location of restriction-enzyme cleavage sites within the clone. The data the maps provide can be used to reclone fragments of a gene or compare its internal organization with that of other cloned sequences.

Before DNA sequencing and bioinformatics became popular, restriction maps were created experimentally by cutting DNA with different restriction enzymes and separating DNA fragments by gel electrophoresis (refer to Figure 9–20 and see chapter opening photo). The digestion pattern of fragments generated can then be interpreted to determine the location of restriction sites for different enzymes.

Because of advances in DNA sequencing and the use of bioinformatics, restriction maps are now created by simply using software to identify restriction-enzyme cutting sites in sequenced DNA. The Exploring Genomics exercise in this chapter involves a Web site, Webcutter, which is commonly used for generating restriction maps. Restriction maps were an important way of characterizing cloned DNA and could be constructed in the absence of any other information about the DNA, including whether or not it encodes a gene or has other functions. In the Human Genome Project, restriction maps of the human genome were important for digesting the genome into pieces that could be sequenced.

Nucleic Acid Blotting

Several of the techniques described in this chapter rely on hybridization between complementary nucleic acid (DNA or RNA) molecules. One of the most widely used methods for detecting such hybrids is called Southern blotting (after Edwin Southern, who devised it). The **Southern blot** method can be used to identify which clones in a library contain a given DNA sequence and to characterize the size of the fragments. Southern blots can also be used to identify fragments carrying specific genes in genomic DNA digested with a restriction enzyme.

Southern blotting has two components: separation of DNA fragments by gel electrophoresis and hybridization of the fragments using labeled probes (**Figure 17–9**). Gel electrophoresis can be used to characterize the number of fragments produced by restriction digestion of relatively small pieces of DNA and to estimate their molecular weights. However, restriction-enzyme digestion of large genomes—such as the human genome, with more than 3 billion nucleotides—will produce so many different fragments that they will run together on a gel to produce a continuous smear. The identification of specific fragments in these cases is accomplished in the next step: hybridization characterizes the DNA sequences present in the fragments.

To produce Figure 17–10, researchers cut samples of genomic DNA with several restriction enzymes. The agarose gel electrophoresis pattern of fragments obtained for each restriction enzyme is shown in **Figure 17–10(a)**. A Southern blot of this gel is illustrated in **Figure 17–10(b)**. The probe hybridized to complementary sequences, identifying fragments of interest.

Southern blotting led to the development of other blotting approaches. RNA blotting was subsequently called **Northern blot analysis** or simply **Northern blotting,** and following a naming scheme that correlates with the directionality of a compass, a related blotting technique involving proteins is known as **Western blotting.** Western blotting is a widely used technique for analyzing proteins. Thus part of the historical significance of Southern blotting is that it led to the development of other blotting methods that are key tools for studying nucleic acids and proteins.

Prior to the development of RT-PCR and real-time PCR, Northern blotting was a common approach used to study gene expression. To determine whether a gene is actively being expressed in a given cell or tissue type, Northern blotting probes for the presence of mRNA complementary to a cloned gene. To do this, mRNA is extracted from a specific cell or tissue type and separated by gel electrophoresis. The RNA is then transferred to a membrane, as in Southern blotting, and the membrane is exposed to a labeled single-stranded DNA or RNA probe derived from a cloned copy of the gene. If mRNA complementary to the DNA probe is present, the complementary sequences will hybridize and be detected as a band on the film. Northern blots provide information about the expression of specific genes and are used to study patterns of gene expression in embryonic tissues, cancer, and genetic disorders. Northern blots also detect alternatively spliced mRNAs (multiple types of transcripts derived from a single gene) and can be used to derive other information about transcribed mRNAs such as the size of a gene's mRNA transcripts, measuring band density and the amount of mRNA expressed by a gene. Northern blots are occasionally still used to study RNA expression, but because PCR-based techniques are faster and more sensitive than blotting methods, techniques such as RT-PCR are often the preferred approach, particularly for measuring changes in gene expression.

Finally, as noted earlier in the text (see Chapter 9), **fluorescence *in situ* hybridization,** or **FISH,** is a powerful tool that involves hybridizing a probe directly to a chromosome or RNA without blotting (see Figure 9–18 and Figures 20–8 and 20–9). FISH can be carried out with isolated chromosomes on a slide or directly

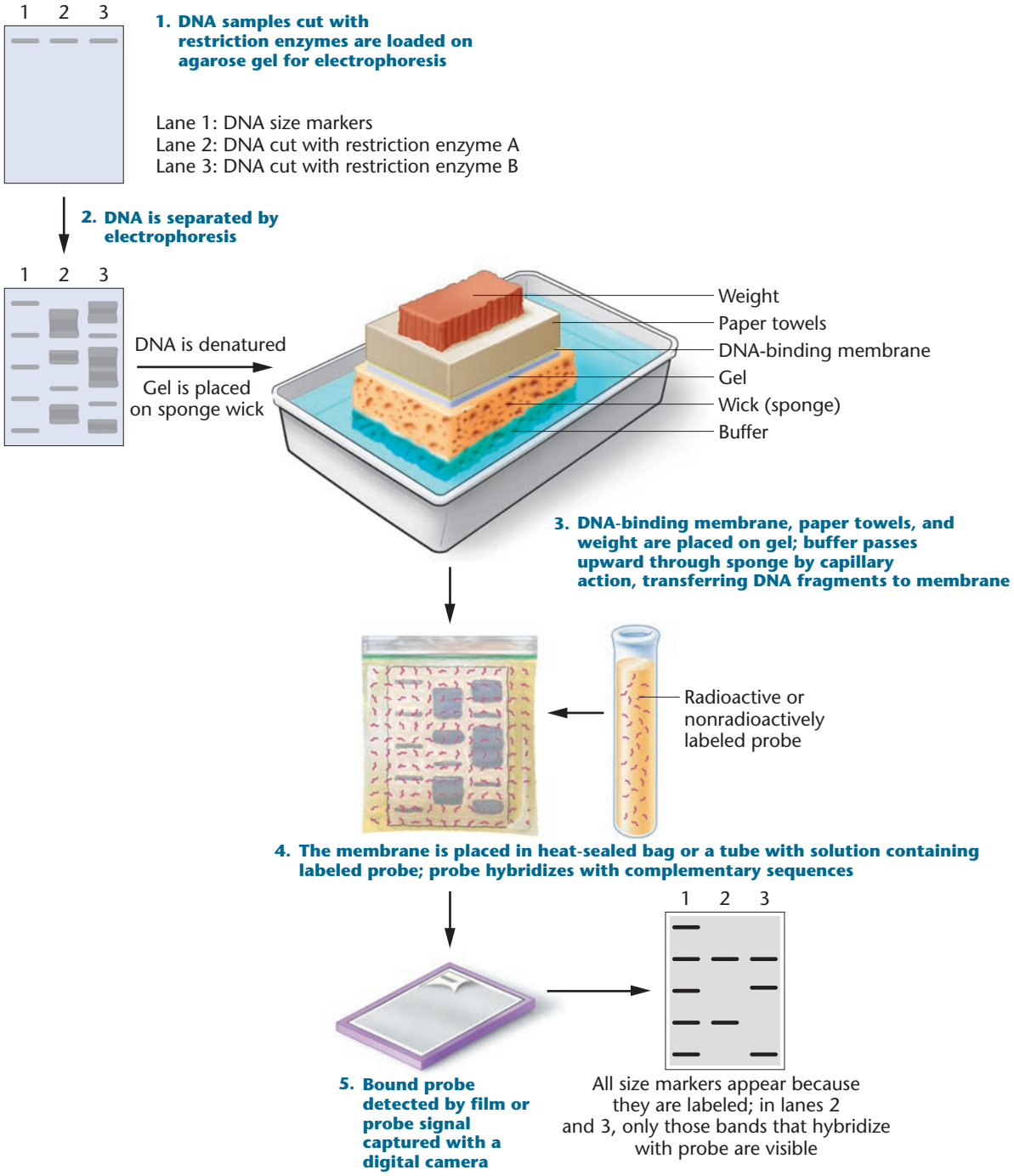


FIGURE 17–9 In the Southern blotting technique, samples of the DNA to be probed are cut with restriction enzymes and the fragments are separated by gel electrophoresis. The pattern of fragments is visualized and photographed under ultraviolet illumination. The gel is placed in an alkaline solution to denature DNA into single-strands then it is placed on a sponge wick that is in contact with a buffer solution and covered with a DNA-binding membrane. Layers of paper towels or blotting paper are placed on top of the membrane and held in place with a weight. Capillary action draws the buffer through the gel, transferring the DNA fragments from the gel to the membrane. Single-stranded DNA fragments on the membrane is hybridized to a labeled DNA probe. The membrane is washed to remove excess probe and overlaid with a piece of X-ray film for autoradiography or chemiluminescence from the probe detected with a digital camera. The hybridized fragments show up as bands on the X-ray film.

in situ in tissue sections or entire organisms, particularly when embryos are used for various studies in developmental genetics (**Figure 17–11**). For example, in developmental studies one can identify which cell types in an embryo express different genes during specific stages of development.

Variations of the FISH technique are also used to produce **spectral karyotypes** in which individual chromosomes can

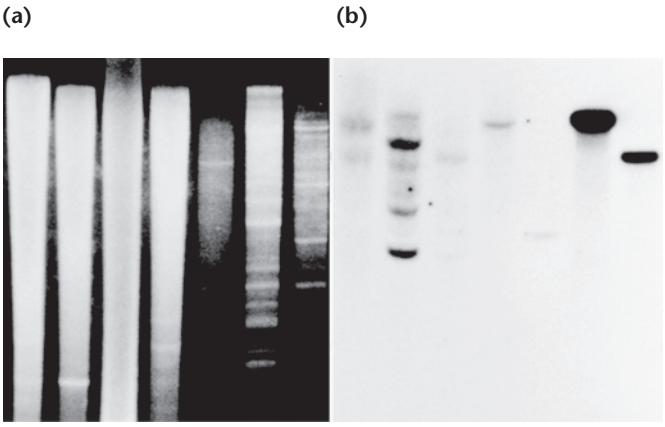


FIGURE 17–10 (a) Agarose gel stained with ethidium bromide to show DNA fragments. (b) Exposed X-ray film of a Southern blot prepared from the gel in part (a). Only those bands containing DNA sequences complementary to the probe show hybridization.

be detected using probes labeled with dyes that will fluoresce at different wavelengths (see Chapter 6 opening photograph and Figure 16–1).

ESSENTIAL POINT

DNA and RNA can be analyzed through a variety of methods that involve hybridization techniques. ■

17.5 DNA Sequencing Is the Ultimate Way to Characterize DNA at the Molecular Level

In a sense, cloned DNA, from a single gene to an entire genome, is completely characterized at the molecular level only when its nucleotide sequence is known. The ability to sequence DNA has greatly enhanced our understanding of genome organization and increased our knowledge of gene structure, function, and mechanisms of regulation.

Historically, the most commonly used method of DNA sequencing was developed by Fred Sanger and his colleagues and is known as **dideoxynucleotide chain-termination sequencing** or simply **Sanger sequencing.** In this technique, a double-stranded DNA molecule whose sequence is to be determined is converted to single strands that are used as a template for synthesizing a series of complementary strands. The DNA to be sequenced is mixed with a primer that is complementary to the target DNA or vector, along with DNA polymerase, and the four deoxyribonucleotide triphosphates (dATP, dCTP, dGTP, and dTTP) are added to each tube.

The key to the Sanger technique is the addition of a small amount of one modified deoxyribonucleotide (**Figure 17–12**), called a **dideoxynucleotide** (abbreviated ddNTP). Notice that dideoxynucleotides have a 3′ hydrogen instead of a 3′ hydroxyl group. Dideoxynucleotides are called chain-termination nucleotides because they lack the 3′ oxygen

FIGURE 17–11 *In situ* hybridization of a zebrafish embryo 48 hours after fertilization showing expression of *atp2a1* mRNA, which encodes a muscle-specific calcium pump. The probe revealing *atp2a1* expression produces dark blue staining. Notice that this staining is restricted to muscle cells surrounding the developing spinal cord of the embryo.

required to form a phosphodiester bond with another nucleotide. Thus when ddNTPs are included in a reaction as DNA synthesis takes place, the polymerase occasionally inserts a dideoxynucleotide instead of a deoxyribonucleotide into a growing DNA strand. Since the dideoxynucleotide has no 3′-OH group, it cannot form a 3′ bond with another nucleotide, and DNA synthesis terminates because DNA polymerase cannot add new nucleotides to a ddNTP. The Sanger reaction takes advantage of this key modification.

For example, in Figure 17–12, notice that the shortest fragment generated is a sequence that has added ddCTP to the 3′ end of the primer and the chain has terminated. Over time as the reaction proceeds, eventually a ddNTP will be inserted at every location in the newly synthesized DNA so that each strand synthesized differs in length by one nucleotide and is terminated by a ddNTP. This allows for separation of these DNA fragments by gel electrophoresis, which can then be used to determine the sequence.

When the Sanger technique was first developed, four separate reaction tubes, each with a different single ddNTP (e.g., ddATP, ddCTP, ddGTP, and ddTTP), were used. These reactions typically used either a radioactively labeled primer or a radioactively labeled ddNTP for analysis of the sequence following polyacrylamide gel electrophoresis and autoradiography. Historically, this approach involved large polyacrylamide gels in which each reaction was loaded on a separate lane of the gel and ladder-like banding patterns revealed by autoradiography were read to determine the sequence. This original approach could typically read several hundred bases per reaction. *Read length*—that is, the amount of sequence that can be generated in a single individual reaction and the total amount of DNA sequence generated in a sequence *run*, which is effectively read length times the number of reactions an instrument can run during a given period of time—has become a hot area for innovation in sequencing technology.

In the past 20 years, modifications of the Sanger technique led to technologies that allowed sequencing reactions to occur in a single tube. As shown in Figure 17–12, each of the four ddNTPs is labeled with a different-colored fluorescent

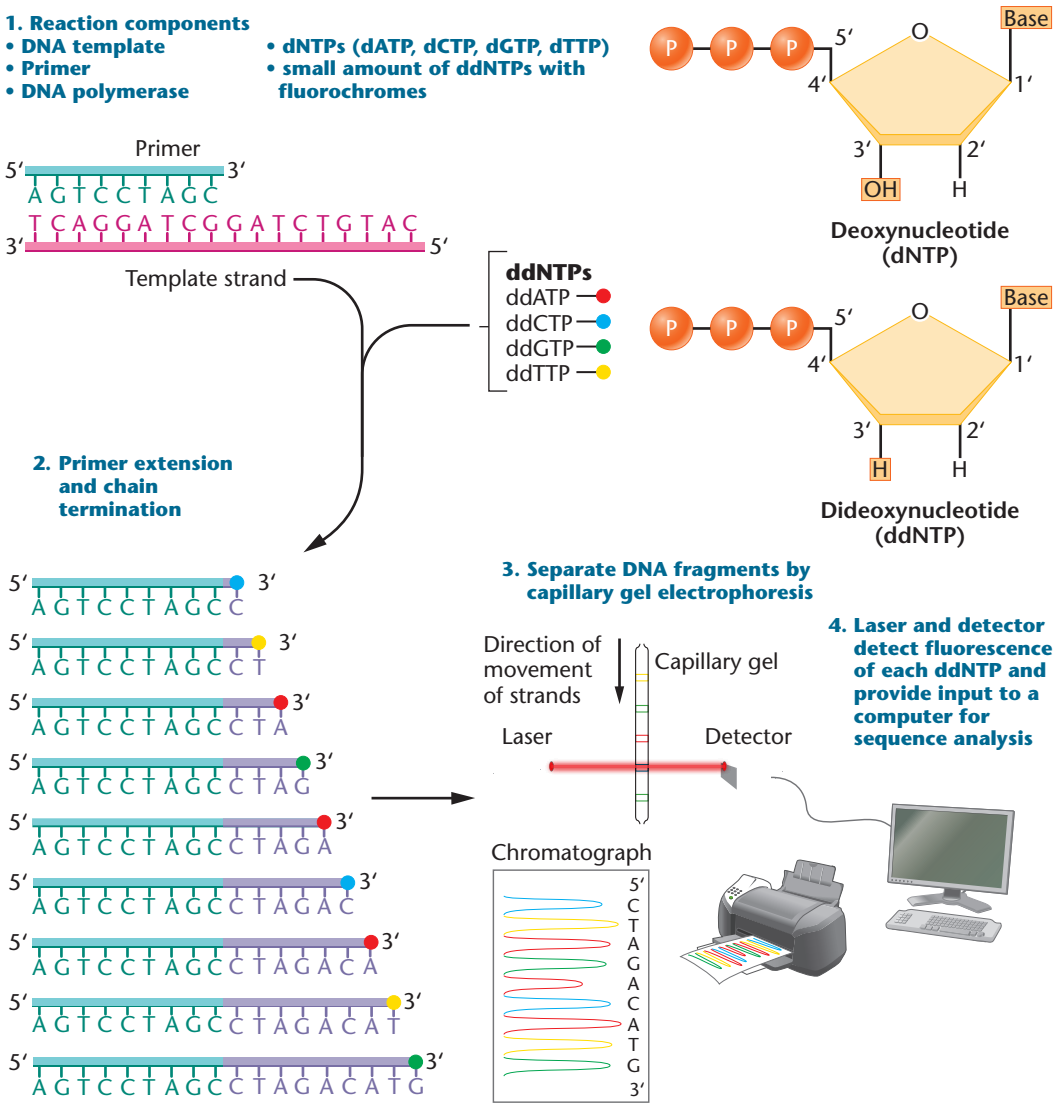


FIGURE 17–12 Computer-automated DNA sequencing using the chain-termination (Sanger) method. (1) A primer is annealed to a sequence adjacent to the DNA being sequenced (usually near the multiple cloning site of a cloning vector). (2) A reaction mixture is added to the primer–template combination. This includes DNA polymerase, the four dNTPs, and small molar amounts of dideoxynucleotides (ddNTPs) labeled with fluorescent dyes. All four ddNTPs are added to the same tube, and during primer extension, all possible lengths of chains are produced. During primer extension, the polymerase occasionally (randomly) inserts a ddNTP instead of a dNTP, terminating the synthesis of the chain because the ddNTP does not have the OH group needed to attach the next nucleotide. Over the course of the reaction, all possible termination sites will have a ddNTP inserted. The products of the reaction are added to a single lane on a capillary gel, and the bands are read by a detector and imaging system. This process is computer automated, and robotic machines sequence several hundred thousand nucleotides in a 24-hour period and then store and analyze the data automatically. The sequence is obtained by extension of the primer and is read from the newly synthesized strand. Thus in this case, the sequence obtained begins with 5′-CTAGACATG-3′.

dye. These reactions were carried out in PCR-like fashion using cycling reactions that permit greater read and run capabilities. The reaction products were separated through a single, ultrathin-diameter polyacrylamide tube gel called a capillary gel (capillary gel electrophoresis). As DNA fragments move through the gel, they are scanned with a laser. The laser stimulates fluorescent dyes on each DNA fragment, which then emit different wavelengths of light for each ddNTP. Emitted light is captured by a detector that amplifies and feeds this information into a computer to convert the light patterns into a DNA sequence that is technically called an electropherogram or chromatograph. The data are represented as a series of colored peaks, each corresponding to one nucleotide in the sequence.

Since the early 1990s, DNA sequencing has largely been performed through computer-automated Sanger-reaction-based technology. Such systems generate relatively large amounts of sequence DNA. Computer-automated sequences can achieve read lengths of approximately 1000 bp with about 99.999 percent accuracy for about $0.50 per kb. Automated DNA sequencers often contain multiple capillary gels (as many as 96) that are several feet long and can process several thousand bases of sequences so that many of these

instruments make it possible to generate over 2 million bp of sequences in a day! These systems became essential for the rapidly accelerating progress of the Human Genome Project.

Sequencing Technologies Have Progressed Rapidly

Since Fred Sanger was awarded part of the 1980 Nobel Prize in Chemistry (which he shared with Walter Gilbert and Paul Berg) for sequencing technology, DNA sequencing technologies have undergone an incredible evolution to dramatically improve sequencing capabilities. New innovations in sequencing technology are developing quickly. Sanger sequencing approaches (particularly those involving computer-automated instruments such as capillary electrophoresis) still have their place in everyday routine applications that require sequencing, such as sequencing a relatively short piece of DNA amplified by PCR.

When it comes to sequencing entire genomes, however, Sanger sequencing technologies are outdated. The costs of Sanger sequencing are relatively high compared to newer technologies, and Sanger sequencing output, even with computer-automated DNA sequencing, is simply not high enough to support the growing demand for genomic data. This demand is being driven in large part by personalized genomics (see Chapter 19) and the desire to reveal the genetic basis of human diseases, which will require tens of thousands of individual genome sequences. As we will discuss later in the text (see Chapter 19), a race was on to develop sequencing technologies for the complete and accurate sequencing of an individual human genome for $1000. Several sequencing companies claim that they have technologies that can sequence entire genomes for $1000 but scientists do not agree on what level of accuracy and other factors should be expected at this price. In mid 2015, current information from the National Human Genome Research Institute estimated a full cost of ~$4200 to sequence a human genome in less than one day. Nonetheless, there is every indication that technology will allow for rapid, accurate, and routine sequencing of genomes for $1000 or less relatively soon thanks to the development of **next-generation sequencing (NGS) technologies.**

NGS technologies dispense with the Sanger technique and capillary electrophoresis methods (*first-generation sequencing*) in favor of sophisticated, parallel formats (simultaneous reaction formats) that synthesize DNA from tens of thousands of identical strands simultaneously and then use state-of-the art fluorescence imaging techniques to detect new synthesized strands and average sequence data across many molecules being sequenced. NGS technologies are providing an unprecedented capacity for generating massive amounts of DNA sequence data rapidly and at dramatically reduced costs per base.

The desire for NGS within the research community and challenges such as the $1000 genome have led to an intense race among many companies eager to produce NGS methods. As a result, there are at least five NGS technologies. Next-generation sequencing started around 2005. Some of the first instruments were capable of producing as much data as 50 capillary electrophoresis systems and are up to 200 times faster and cheaper than conventional Sanger approaches. NGS instruments generally produce short read lengths of ~200–400 bp, and then these snippets of sequence are stitched together using software to produce a coherent, complete genome.

Shortly after NGS methods were commercialized, companies were announcing progress on **third-generation sequencing (TGS)**. TGS methods are based on strategies that sequence a *single molecule* of single-stranded DNA, and at least four different approaches are being explored.

Recently, TGS was used to sequence the genomes of five strains of *Vibrio cholerae* involved in a cholera outbreak in Haiti in less than an hour. This genetic determination of the *Vibrio* strains resulted in rapid action to successfully treat the outbreak with antibiotics to which these bacteria were not resistant. Most TGS technologies still have somewhat high error rates for sequencing accuracy—about 15 percent errors in sequence generated. At around $750,000, these technologies also remain very expensive.

The genomics research community has embraced NGS and third-generation sequencing technologies. Which approaches will eventually emerge as the sequencing methods of choice for the long-term is unclear, but what is clear is that the landscape of sequencing capabilities has dramatically changed for the better and never before have scientists had the ability to generate so much sequence data so quickly. Through 2006, new sequencing technologies were cutting sequencing costs in half about every two years. And keep an eye on the $1000 genome: there is every reason to believe sequencing technology will get us there soon. Our overview of DNA sequencing in this chapter is a great introduction to a detailed discussion of genomics and many related topics found later in the text (see Chapter 18).

ESSENTIAL POINT

DNA sequencing technologies are changing rapidly. Next-generation and third-generation methods produce large amounts of fairly accurate sequence data in a short time. ■

17.6 Creating Knockout and Transgenic Organisms for Studying Gene Function

Recombinant DNA technology has also made it possible to directly manipulate genes in organisms in ways that allow scientists to learn about gene function *in vivo*. These approaches enable scientists to create genetically engineered plants and animals for research and for commercial applications. We conclude this chapter by briefly

discussing gene knockout technology and creating transgenic animals as examples of *gene-targeting* methods that have revolutionized research in genetics. In selected chapters of the book, the Modern Approaches to Understanding Gene Function boxes have highlighted examples of specific research projects involving gene-targeting approaches.

Gene Targeting and Knockout Animal Models

The concept behind *gene targeting* is to manipulate a specific allele, locus, or base sequence to learn about the functions of a gene of interest. In the 1980s, scientists devised a gene-targeting technique for creating **gene knockout** (often abbreviated as KO) organisms, specifically mice. The pioneers of knockout technology, Dr. Mario Capecchi of the University of Utah and colleagues Oliver Smithies of the University of North Carolina, Chapel Hill, and Sir Martin Evans of Cardiff University, UK, received the 2007 Nobel Prize in Physiology or Medicine for developing this technique.

The fundamental purpose of creating a knockout is to disrupt or eliminate a specific gene or genes of interest and then ask, "what happens?" If physical, behavioral, and biochemical changes or other metabolic phenotypes or functions are observed in the KO animal, then one can begin to see that the gene of interest has some functional role or roles in the observed phenotypes. Thus one of the most valuable reasons for creating a KO is to learn about gene function. The KO techniques developed in mice led to similar technologies for making KOs in zebrafish, rats, pigs, fruit flies, and many other organisms including plants.

Knockout mice have revolutionized research in genetics, molecular biology, and biomedical research in many ways. Scientists have used knockout methods to create thousands of knockout organisms that have advanced our understanding of gene function, created animal models for many human diseases, and enabled scientists to make transgenic animals (which we will discuss below). Applications of KO technology have also provided the foundation for gene-targeting approaches in gene therapy that we discuss later in the text (see Special Topic Chapter 6—Gene Therapy).

Generally, generating a KO mouse or a transgenic mouse is a very labor-intensive project that can take several years of experiments and crosses and a significant budget to complete. However, once a KO mouse is made, assuming it is fertile, a colony of mice can be maintained; often KO mice are shared around the world so that other researchers can work with them. Many companies will produce KO mice for researchers. It is also possible to make *double-knockout animals (DKOs)* and even *triple-knockout animals (TKOs)*. This approach is typically used when scientists want to study the functional effects of disrupting two or three genes thought to be involved in a related mechanism or pathway.

A KO animal can be made in several ways (**Figure 17–13**). Here we outline a strategy for making KO mice, but the same

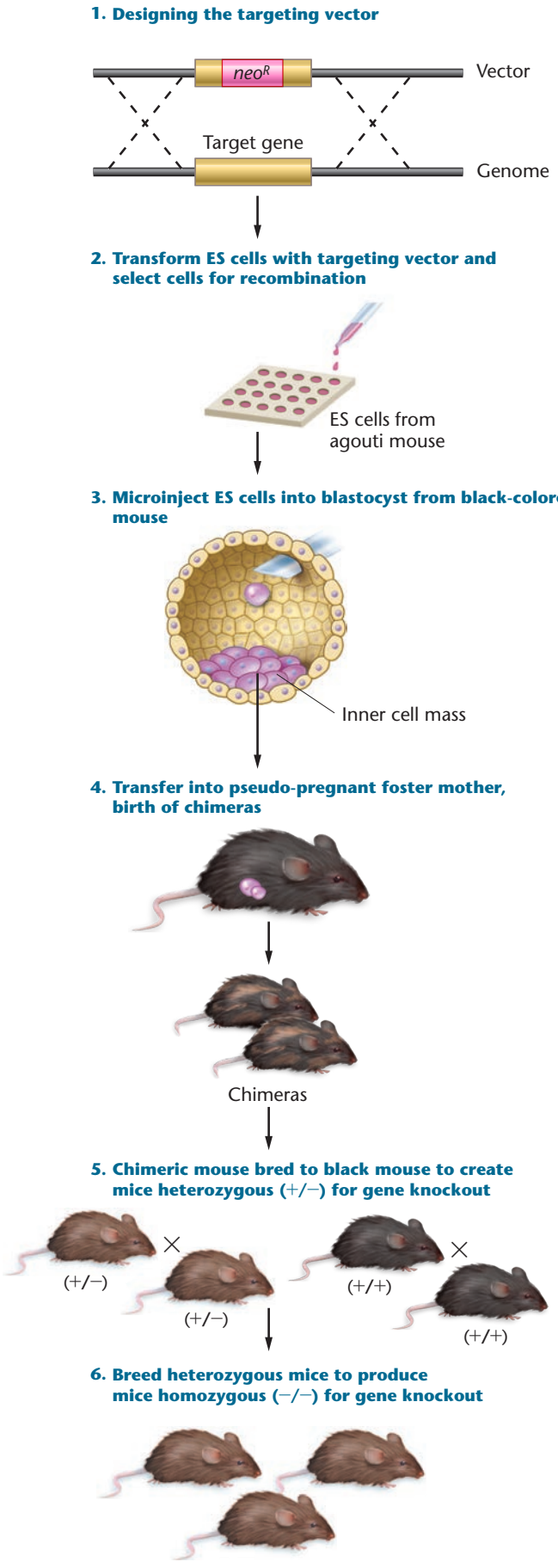


FIGURE 17–13 A basic strategy for producing a knockout mouse.

basic methods apply when making most KO animals. The DNA sequence for the gene of interest to be targeted for KO must be known. Scientists also need to know some sequence information about noncoding sequences that flank the gene at its location in the genome. A *targeting vector* is then constructed. The purpose of the targeting vector is to create a segment of DNA that can be introduced into cells. It then undergoes homologous recombination with the gene of interest (the target gene) to disrupt or replace the gene of interest, thereby rendering it nonfunctional. The targeting vector contains a copy of the gene of interest that has been mutated by inserting a large segment of foreign DNA, essentially a large insertion mutation. This foreign DNA will disrupt the reading frame of the target gene so that if the gene is transcribed into mRNA and translated into protein, it will produce a nonfunctional protein.

To help scientists determine whether the targeting vector has been properly introduced into the genome, the insertion sequence typically contains a marker gene. The example shown in Figure 17–13 uses a marker sequence for neomycin resistance (neo^R). Neomycin is an antibiotic that blocks protein synthesis in both prokaryotic and eukaryotic cells, and its role will become apparent momentarily. The gene for green fluorescent protein (GFP), the *lacZ* gene that we discussed earlier in this chapter, and other marker genes are also sometimes used as marker genes. These markers allow for very visual examples of KO or transgenic organisms, and you will see an example of a GFP transgenic animal later in this section.

There are several ways to introduce the targeting vector into cells. One popular approach involves using electroporation to deliver the vector into **embryonic stem (ES) cells** grown in culture. The ES cells are harvested from the inner cell mass of a mouse embryo at the blastocyst stage. Alternatively, the targeting vector is directly injected into the blastocyst with the hopes that it will enter ES cells in the inner cell mass. Sometimes it is possible to make KOs by isolating newly fertilized eggs from a female mouse (or female of the desired animal species) and microinjecting the targeting vector DNA directly into the nucleus of the egg or into one of the pronuclei of a fertilized egg [**Figure 17–14(a)**].

Randomly, in a very small percentage of ES cells that have taken in the targeting vector, the actions of the endogenous enzyme recombinase will catalyze homologous recombination between the targeting vector and the sequence for the gene of interest. In the few recombinant ES cells that will be created, the targeting vector will usually replace the original gene on only one of two chromosomes.

Recombinant ES cells can be selected for by treating cultured ES cells with a reagent that will kill cells that lack the targeting vector. For the example shown in Figure 17–13, neomycin is added to cultured ES cells. Cells containing the targeting vector are resistant to neomycin, but ES cells that are not nonrecombinant die. Recombinant ES cells are then injected into a mouse embryo at the blastocyst stage where they will be incorporated into the inner cell mass of the blastocyst. The blastocyst is then placed into the uterus of a surrogate mother mouse, sometimes called a *pseudopregnant mouse*—a female mouse bred by a sterile male mouse. The pseudopregnant mouse offers a uterus that is

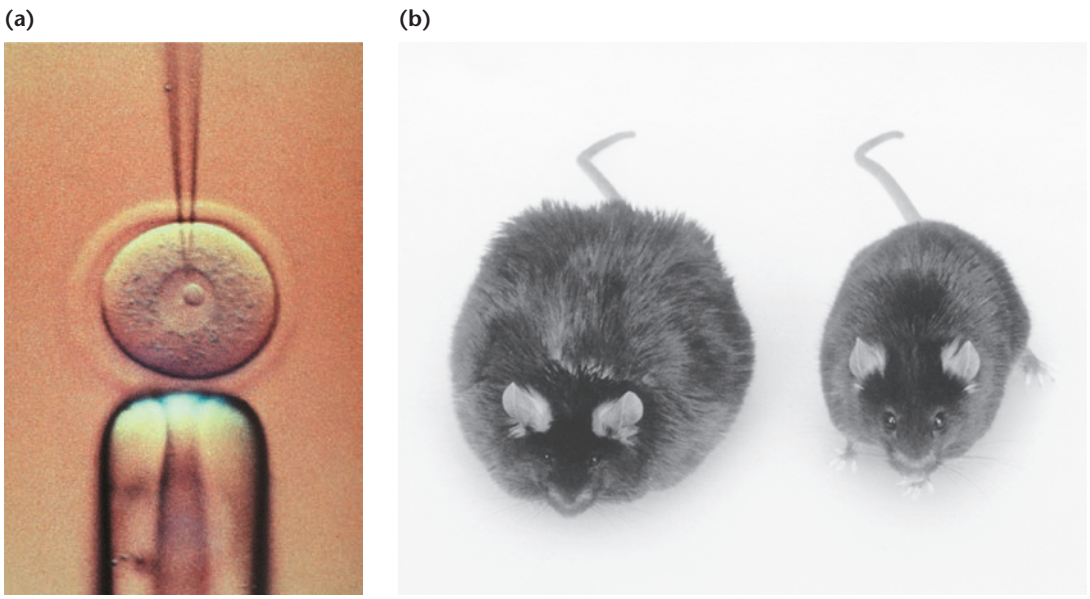


FIGURE 17–14 (a) Microinjecting DNA into a fertilized egg to create a knockout or a transgenic mouse. A fertilized egg is held by a suction or holding pipette (seen below the egg), and a microinjection needle delivers cloned DNA into the nucleus of the egg. (b) On the left is a null mouse (−/−) for both copies of the obese (*ob*) gene, which produces a peptide hormone called leptin. The mouse on the right is wild type (+/+) for the *ob* gene. The *ob* knockout mouse weighs almost five times as much as its wild-type sibling. Naturally occurring mutations in the human *Ob* gene create weight disorders for affected individuals.

receptive to implantation of the blastocyst containing the targeting vector.

From the implanted embryos the surrogate will give birth to mice that are *chimeras*: some cells in their body arise from KO stem cells, and others arise from stem cells of the donor blastocyst. Researchers screen for mice that contain the targeting vector by obtaining DNA from a sample of tail tissue, purifying the DNA, and performing PCR to verify that the targeting vector sequence is present in the animal's genome. As long as the targeting vector DNA is present in germ cells, the sequence will be inherited in all of the offspring generated by these mice, but typically most F_1 generation KO mice produced this way are heterozygous (+/–) for the gene of interest and the targeting vector and not homozygous for the KO. Sibling matings of F_1 animals can then be used to generate homozygous KO animals, referred to as *null mice* and given a –/– designation because they lack wild-type copies of the targeted gene of interest. As mentioned at the beginning of this section, KO animal models serve invaluable roles for learning about gene function, and they continue to be essential for biomedical research on disease genes [see **Figure 17–14(b)**].

Despite all of the work that goes into trying to produce a KO organism, sometimes viable offspring are never born. The KO results in *embryonic lethality*. Knocking out a gene that is important during embryonic development may kill the embryo before it has a chance to fully develop. Typically, researchers will examine embryos from the pseudopregnant mouse to see if they can determine at what stage of embryonic development the embryo is dying. This examination often reveals specific organ defects that can also be informative about the function of the KO gene.

If null mice for a particular gene of interest cannot be derived by traditional KO approaches, a **conditional knockout** can often provide a way to study such a gene. Conditional knockouts allow one to control when a target gene is disrupted. This study can be done at a particular time in the animal's development. For example, if a target gene displays embryonic lethality, one can use a conditional KO to allow an animal to progress through development and be born before disrupting. Another advantage of conditional KOs is that target genes can also be turned off in a particular tissue or organ instead of the entire animal.

Gene-Editing Methods

Gene-editing methods allow researchers to create changes in a specific sequence to remove, correct or replace a defective gene or parts of a gene. Gene editing is based on using different nucleases to create breaks in the genome in a sequence specific manner. Later in the text (see Special Topic Chapter 6—Gene Therapy), we discuss how gene-editing methods with *transcription activator-like effector nucleases (TALENs)* and *zinc finger nucleases (ZFNs)* can be used for gene therapy. These approaches have also been used for the *in vivo* genetic engineering of mice, rats, and other species for targeted modifications of specific genes. We also mention *CRISPRs* (Clustered Regularly Interspaced Palindromic Repeats) as a powerful new tool for gene-targeting experiments.

Making a Transgenic Animal: The Basics

Transgenic animals (**Figure 17–15**), also sometimes called **knock-in animals,** express or often overexpress a particular gene of interest (the transgene). The method of creating transgenic animals is conceptually simple, and many of the steps are similar to the steps involved in making a KO. Moreover, as is true when making KOs, there are species-specific challenges associated with creating transgenics. Many of the prevailing techniques used to make transgenics were also developed in mice.

Conceptually, introducing a transgene into an organism involves steps similar to making a KO. But instead of trying to disrupt a target gene, a vector with the transgene is created to undergo homologous recombination into the host-cell genome. In some applications, tissue-specific promoter sequences can be used so that the transgene is expressed only in specific tissues. For example, tissue-specific promoters are used in the biotechnology industry to express specific recombinant products in milk for subsequent purification. It is often easier to make a transgenic than a KO because the vector just needs to be incorporated into the host genome somewhere (hopefully in a noncoding region) but often not at a particular locus as is necessary when making a KO.

FIGURE 17–15 Transgenic mice incorporating the GFP gene from jellyfish enable scientists to tag particular genes with GFP to track gene activity, including activity in subsequent generations of mice generated from these transgenics. This procedure can be very valuable for examining the transgenerational effects on gene expression, including epigenetic changes.

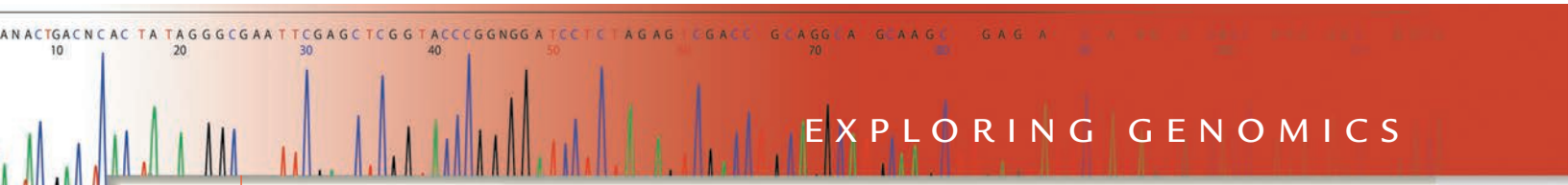
The vector with the transgene can be put into ES cells or injected directly into embryos. Then, in a relatively small percentage of embryos or eggs, the transgenic DNA becomes inserted into the egg cell genome by recombination due to the action of naturally occurring DNA recombinases. After this stage, the rest of the process is similar to making a KO: embryos are placed into pseudopregnant females, and resulting crosses are used to derive mice that are homozygous for the transgene.

In a transgenic, the transgene is often overexpressed in order to study its effects on the appearance and functions of mice. There are many variations and purposes for making transgenics. Transgenic animals overexpressing certain genes, expressing human genes or genes from a different species, and expressing mutant genes are among examples of transgenics that are valuable models for basic and applied research to understand gene function.

As we will consider later in the text (see Chapter 19), transgenic animals and plants are also created in order to produce commercially valuable biotechnology products. Later in the text (see Special Topic Chapter 5—Genetically Modified Foods), you will learn about examples of transgenic food crops.

ESSENTIAL POINT

Gene-targeting methods to create knockout animals and transgenic animals are widely used, valuable approaches for studying gene function *in vivo*. ■



EXPLORING GENOMICS

Manipulating Recombinant DNA: Restriction Mapping and Designing PCR Primers

MasteringGenetics™ Visit the Study Area: Exploring Genomics

As you learned in this chapter, restriction enzymes are sophisticated "scissors" that molecular biologists use to cut DNA, and they are routinely used in genetics and molecular biology laboratories for recombinant DNA experiments. A wide variety of online tools assist scientists working with restriction enzymes and manipulating recombinant DNA for different applications. Here we explore **Webcutter**, a site that makes recombinant DNA experiments much easier.

■ Exercise I – Creating a Restriction Map in Webcutter

Suppose you had cloned and sequenced a gene and you wanted to design a probe approximately 600 bp long that could be used to analyze expression of this gene in different human tissues by Northern blot analysis. Not too long ago, you had primarily two ways to approach this task. You could digest the cloned DNA with whatever restriction enzymes were in your freezer, then run agarose gels and develop restriction maps in the hope of identifying cutting sites that would give you the size fragment you wanted. Or you could scan the sequence with your eyes, looking for restriction sites of interest—a very time-consuming and eye-straining effort! Internet sites such as **Webcutter** take the guesswork out of developing restriction maps and make it relatively easy to design experiments for manipulating recombinant DNA. In this exercise, you will use Webcutter to create a restriction map of human DNA with the enzymes *Eco*RI, *Bam*HI, and *Pst*I.

1. Access **Webcutter** at http://rna.lundberg.gu.se/cutter2/. Go to the Companion Web site for *Concepts of Genetics* and open the Exploring Genomics exercise for this chapter. Copy the sequence of cloned human DNA found there and paste it into the text box in Webcutter.
2. Scroll down to "Please indicate which enzymes to include in the analysis." Click the button indicating "Use only the following enzymes." Select the restriction enzymes *Eco*RI, *Bam*HI, and *Pst*I from the list provided, then click "Analyze sequence." (*Note:* Use the command, control, or shift key to select multiple restriction enzymes.)
3. After examining the results provided by Webcutter, create a table showing the number of cutting sites for each enzyme and the fragment sizes that would be generated by digesting with each enzyme. Draw a restriction map indicating cutting sites for each enzyme with distances between each site and the total size of this piece of human DNA.

■ Exercise II – Designing a Recombinant DNA Experiment

Now that you have created a restriction map of your piece of human DNA, you need to ligate the DNA into a plasmid DNA vector that you can use to make your probe. To do this, you will need to determine which restriction enzymes would best be suited for cutting both the plasmid and the human DNA.

1. Referring back to the Companion Web site and the Exploring Genomics exercise for this chapter, copy the plasmid DNA sequence from Exercise II into the text box in Webcutter and identify cutting sites for the same

enzymes you used in Exercise I. Then answer the following questions:

a. What is the total size of the plasmid DNA analyzed in Webcutter?

b. Which enzyme(s) could be used in a recombinant DNA experiment to ligate the plasmid to the *largest* DNA fragment from the human gene? Briefly explain your answer.

c. What size recombinant DNA molecule will be created by ligating these fragments?

d. Draw a simple diagram showing the cloned DNA inserted into the plasmid and indicate the restriction-enzyme cutting site(s) used to create this recombinant plasmid.

2. As you prepare to carry out this subcloning experiment, you find that the expiration dates on most of your restriction enzymes have long since passed. Rather than run an experiment with old enzymes, you decide to purchase new enzymes. Fortunately, a site called **REBASE®**: **The Restriction Enzyme Database** can help you. Over 300 restriction enzymes are commercially available rather inexpensively, but scientists are always looking for ways to stretch their research budgets as far as possible. REBASE is excellent for locating enzyme suppliers and enzyme specifics, particularly if you need to work with an enzyme that you are unfamiliar with. Visit **REBASE®** at http://rebase.neb.com/rebase/rebase.html to identify companies that sell the restriction enzyme(s) you need for this experiment.

CASE STUDY Should we worry about recombinant DNA technology?

Early in the 1970s, when recombinant DNA research was first developed, scientists realized that there may be unforeseen dangers, and after a self-imposed moratorium on all such research, they developed and implemented a detailed set of safety protocols for the construction, storage, and use of genetically modified organisms. These guidelines then formed the basis of regulations adopted by the federal government. Over time, safer methods were developed, and these stringent guidelines were gradually relaxed or, in many cases, eliminated altogether. Now, however, the specter of bioterrorism has re-focused attention on the potential misuses of recombinant DNA technology. For example, individuals or small groups might use the information in genome databases coupled with recombinant DNA technology to construct or reconstruct agents of disease, such as the smallpox virus or the deadly influenza virus.

1. Do you think that the question of recombinant DNA research regulation by university and corporations should be revisited to monitor possible bioterrorist activity?
2. Should freely available access to genetic databases, including genomes, and gene or protein sequences be continued, or should it be restricted to individuals who have been screened and approved for such access?
3. Forty years after its development, the use of recombinant DNA technology is widespread and is found even in many middle school and high school biology courses. Are there some aspects of gene splicing that might be dangerous in the hands of an amateur?

INSIGHTS AND SOLUTIONS

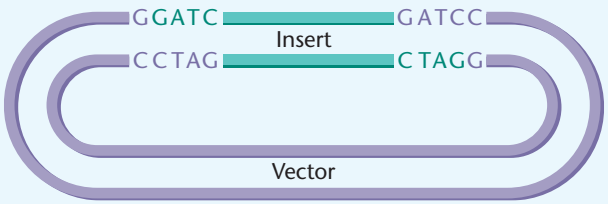
The recognition sequence for the restriction enzyme *Sau*3AI is GATC (see Figure 17–1); in the recognition sequence for the enzyme *Bam*HI—GGATCC—the four internal bases are identical to the *Sau*3AI sequence. The single-stranded ends produced by the two enzymes are identical. Suppose you have a cloning vector that contains a *Bam*HI recognition sequence and you also have foreign DNA that was cut with *Sau*3AI.

(a) Can this DNA be ligated into the *Bam*HI site of the vector, and if so, why?

(b) Can the DNA segment cloned into this sequence be cut from the vector with *Sau*3AI? With *Bam*HI? What potential problems do you see with the use of *Bam*HI?

Solution: (a) DNA cut with *Sau*3AI can be ligated into the vector's *Bam*HI cutting site because the single-stranded ends generated by the two enzymes are identical.

(b) The DNA can be cut from the vector with *Sau*3AI because the recognition sequence for this enzyme (GATC) is maintained on each side of the insert. Recovering the cloned insert with *Bam*HI is more problematic. In the ligated vector, the conserved sequences are GGATC (left) and GATCC (right). The correct base for recognition by *Bam*HI will *follow* the conserved sequence (to produce GGATCC on the left) only about 25 percent of the time, and the correct base will *precede* the conserved sequence (and produce GGATCC on the right) about 25 percent of the time as well. Thus, *Bam*HI will be able to cut the insert from the vector ($0.25 \times 0.25 = 0.0625$), or only about 6 percent, of the time.



Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter we focused on how specific DNA sequences can be copied, identified, characterized, and sequenced. At the same time, we found many opportunities to consider the methods and reasoning underlying these techniques. From the explanations given in the chapter, what answers would you propose to the following fundamental questions?
   (a) In a recombinant DNA cloning experiment, how can we determine whether DNA fragments of interest have been incorporated into plasmids and, once host cells are transformed, which cells contain recombinant DNA?
   (b) When using DNA libraries to clone genes, what combination of techniques are used to identify a particular gene of interest?
   (c) What steps make PCR a chain reaction that can produce millions of copies of a specific DNA molecule in a matter of hours without using host cells?
   (d) How has DNA sequencing technology evolved in response to the emerging needs of genome scientists?

CONCEPT QUESTION

2. Review the Chapter Concepts list on page 338. All of these refer to recombinant DNA methods and applications. Write a short essay or sketch a diagram that provides an overview of how recombinant DNA techniques help geneticists study genes. ■
3. What roles do restriction enzymes, vectors, and host cells play in recombinant DNA studies? What role does DNA ligase perform in a DNA cloning experiment? How does the action of DNA ligase differ from the function of restriction enzymes?
4. The human insulin gene contains a number of sequences that are removed in the processing of the mRNA transcript. In spite of the fact that bacterial cells cannot excise these sequences from mRNA transcripts, explain how a gene like this can be cloned into a bacterial cell and produce insulin.
5. Although many cloning applications involve introducing recombinant DNA into bacterial host cells, many other cell types are also used as hosts for recombinant DNA. Why?
6. You want to perform restriction digestion on a specific segment of DNA containing the nucleotide sequence shown below. Which restriction enzyme will you use for this experiment? Give the complementary sequence of the identified restriction site. (Consult Figure 17–1 for a list of restriction sites.)

   TCTGTGAGAATTCCTAGGTA

7. Restriction sites are palindromic; that is, they read the same in the 5′ to 3′ direction on each strand of DNA. What is the advantage of having restriction sites organized this way?
8. List the advantages and disadvantages of using plasmids as cloning vectors. What advantages do BACs and YACs provide over plasmids as cloning vectors?
9. What are the advantages of using a restriction enzyme whose recognition site is relatively rare? When would you use such enzymes?
10. The introduction of genes into plants is a common practice that has generated not only a host of genetically modified foodstuffs, but also significant worldwide controversy. Interestingly, a tumor-inducing plasmid is often used to produce genetically modified plants. Is the use of a tumor-inducing plasmid the source of such controversy?
11. What is a cDNA library, and for what purpose can it be used?
12. If you performed a PCR experiment starting with only one copy of double-stranded DNA, approximately how many DNA molecules would be present in the reaction tube after 15 cycles of amplification?
13. In a control experiment, a plasmid containing a *Hin*dIII recognition sequence within a kanamycin resistance gene is cut with *Hin*dIII, re-ligated, and used to transform *E. coli* K12 cells. Kanamycin-resistant colonies are selected, and plasmid DNA from these colonies is subjected to electrophoresis. Most of the colonies contain plasmids that produce single bands that migrate at the same rate as the original intact plasmid. A few colonies, however, produce two bands, one of original size and one that migrates much higher in the gel. Diagram the origin of this slow band as a product of ligation.
14. What advantages do cDNA libraries provide over genomic DNA libraries? Describe cloning applications where the use of a genomic library is necessary to provide information that a cDNA library cannot.
15. To create a cDNA library, cDNA can be inserted into vectors and cloned. In the analysis of cDNA clones, it is often difficult to find clones that are full length—that is, many clones are shorter than the mature mRNA molecules from which they are derived. Why is this so?
16. List the steps involved in screening a genomic library. What must be known before starting such a procedure? What are the potential problems with such a procedure, and how can they be overcome or minimized?
17. What is quantitative real-time PCR (qPCR)? Describe what happens during a qPCR reaction and how it is quantified.
18. We usually think of enzymes as being most active at around 37°C, yet in PCR the DNA polymerase is subjected to multiple exposures of relatively high temperatures and seems to function appropriately at 70–75°C. What is special about the DNA polymerizing enzymes typically used in PCR?
19. How do next-generation sequencing (NGS) and third-generation sequencing (TGS) differ from Sanger sequencing?
20. What is the difference between a knockout animal and a transgenic animal?
21. One complication of making a transgenic animal is that the transgene may integrate at random into the coding region, or the regulatory region, of an endogenous gene. What might be the consequences of such random integrations? How might this complicate genetic analysis of the transgene?
22. When disrupting a mouse gene by knockout, why is it desirable to breed mice until offspring homozygous (–/–) for the knockout target gene are obtained?
23. What techniques can scientists use to determine if a particular transgene has been integrated into the genome of an organism?

18 Genomics, Bioinformatics, and Proteomics

CHAPTER CONCEPTS

- Genomics applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze genomes.
- Disciplines in genomics encompass several areas of study, including structural and functional genomics, comparative genomics, and metagenomics, and have led to an "omics" revolution in modern biology.
- Bioinformatics merges information technology with biology and mathematics to store, share, compare, and analyze nucleic acid and protein sequence data.
- The Human Genome Project has greatly advanced our understanding of the organization, size, and function of the human genome.
- Since completion of the Human Genome Project, a new era of genomics studies is providing deeper insights into the human genome.
- Comparative genomics analysis has revealed similarities and differences in genome size and organization.
- Metagenomics is the study of genomes from environmental samples and is valuable for identifying microbial genomes.
- Transcriptome analysis provides insight into patterns of gene expression and gene-regulatory activity of a genome.
- Proteomics focuses on the protein content of cells and on the structures, functions, and interactions of proteins.
- Systems biology approaches attempt to uncover complex interactions among genes, proteins, and other cellular components.

Alignment comparing DNA sequence for the leptin gene from dogs (top) and humans (bottom). Vertical lines and shaded boxes indicate identical bases. *LEP* encodes a hormone that functions to suppress appetite. This type of analysis is a common application of bioinformatics and a good demonstration of comparative genomics.

The term **genome,** meaning the complete set of DNA in a single cell of an organism, was coined at a time when geneticists began to turn from the study of individual genes to a focus on the larger picture. In 1977, as recombinant DNA-based techniques were developed, Fred Sanger and colleagues began the field of **genomics,** the study of genomes, by using a newly developed method of DNA sequencing to sequence the 5400-nucleotide genome of the virus ϕX174. Other viral genomes were sequenced in short order, but even this technology was slow and labor-intensive, limiting its use to small genomes.

During the next three decades, the development of computer-automated DNA sequencing methods made it possible to consider sequencing the larger and more complex genomes of eukaryotes, including the human genome. The development of recombinant DNA technologies coupled with the advent of new, powerful DNA sequencing methods and bioinformatics is responsible for rapidly accelerating the field of genomics.

Genomic technologies have developed so quickly that modern biological research is currently experiencing a genomics revolution. In this chapter, we will examine basic technologies used in genomics and then discuss examples of genome data and different disciplines of genomics. We will also discuss *transcriptome analysis,* the study of genes expressed in a cell or tissue (the "transcriptome"), and *proteomics,* the study of proteins present in a cell or tissue.

The chapter concludes with a brief look at *systems biology,* a new area of contemporary biology that incorporates and integrates genomics, transcriptome analysis, and proteomics data. Later in the text (see Chapter 19) we will continue our discussion of genomics by describing many modern applications of recombinant DNA and genomic technologies. Please note that some of the topics discussed in this chapter are explored in greater depth in later chapters (see Special Topic Chapter 1—Epigenetics, Special Topic Chapter 2—Emerging Roles of RNA, and Special Topic Chapter 4—Genomics and Personalized Medicine).

18.1 Whole-Genome Sequencing Is a Widely Used Method for Sequencing and Assembling Entire Genomes

As discussed earlier in the text (see Chapter 17), recombinant DNA technology made it possible to generate DNA libraries that could be used to identify, clone, and sequence specific genes of interest. But a primary limitation of library screening and even of most polymerase chain reaction (PCR) approaches is that they typically can identify only relatively small numbers of genes at a time. Genomics allows the sequencing of entire genomes. **Structural genomics** focuses on sequencing genomes and analyzing nucleotide sequences to identify genes and other important sequences such as gene-regulatory regions.

The most widely used strategy for sequencing and assembling an entire genome involves variations of a method called **whole-genome sequencing** (**WGS**), also known as **shotgun cloning** or shotgun sequencing. In simple terms, this technique is analogous to you and a friend taking your respective copies of this genetics textbook and randomly ripping the pages into strips about 5 to 7 inches long. Each chapter represents a chromosome, and all of the letters in the entire book are the "genome." Then you and your friend would go through the painstaking task of comparing the pieces of paper to find places that match, overlapping sentences—areas where there are similar sentences on different pieces of paper. Eventually, in theory, many of the strips containing matching sentences would overlap in ways that you could use to reconstruct the pages and assemble the order of the entire text.

Figure 18–1 shows a basic overview of WGS. First, an entire chromosome is cut into short, overlapping fragments, either by mechanically shearing the DNA in various ways (such as excessive heat treatment or sonication in which sonic energy is used to break DNA) or by using restriction enzymes to cleave the DNA at different locations. For simplicity, here we present a basic example of DNA digestion using restriction enzymes. Increasingly, nonenzymatic approaches for shearing DNA are being used. Different restriction enzymes can be used so that chromosomes are cut at different sites; or sometimes, *partial digests* of DNA using the same restriction enzyme are

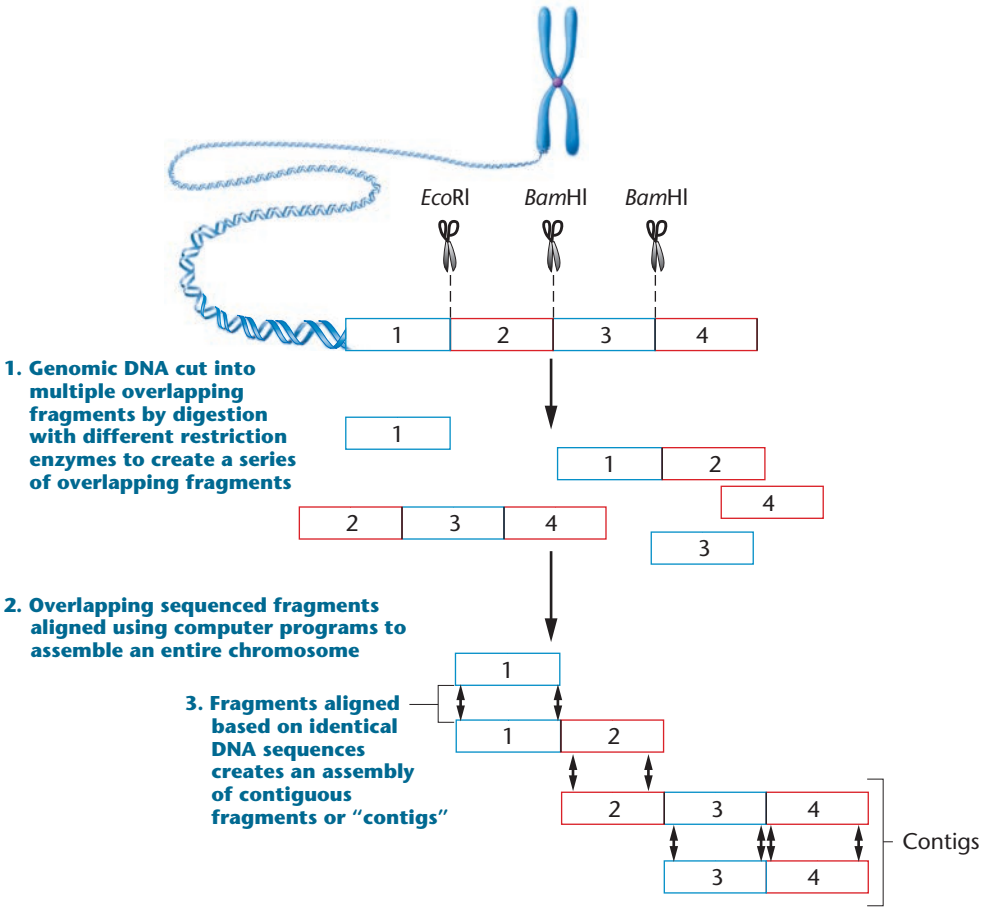


FIGURE 18–1 An overview of whole-genome sequencing (WGS) and assembly. This approach shows one strategy that involves using restriction enzymes to digest genomic DNA into overlapping fragments, which are then sequenced and aligned using bioinformatics to identify overlapping fragments based on sequence identity. Notice that *Eco*RI digestion of the portion of DNA depicted here produces two fragments (contigs 1, 2–4), whereas digestion with *Bam*HI produces three fragments (contigs 1–2, 3, 4).

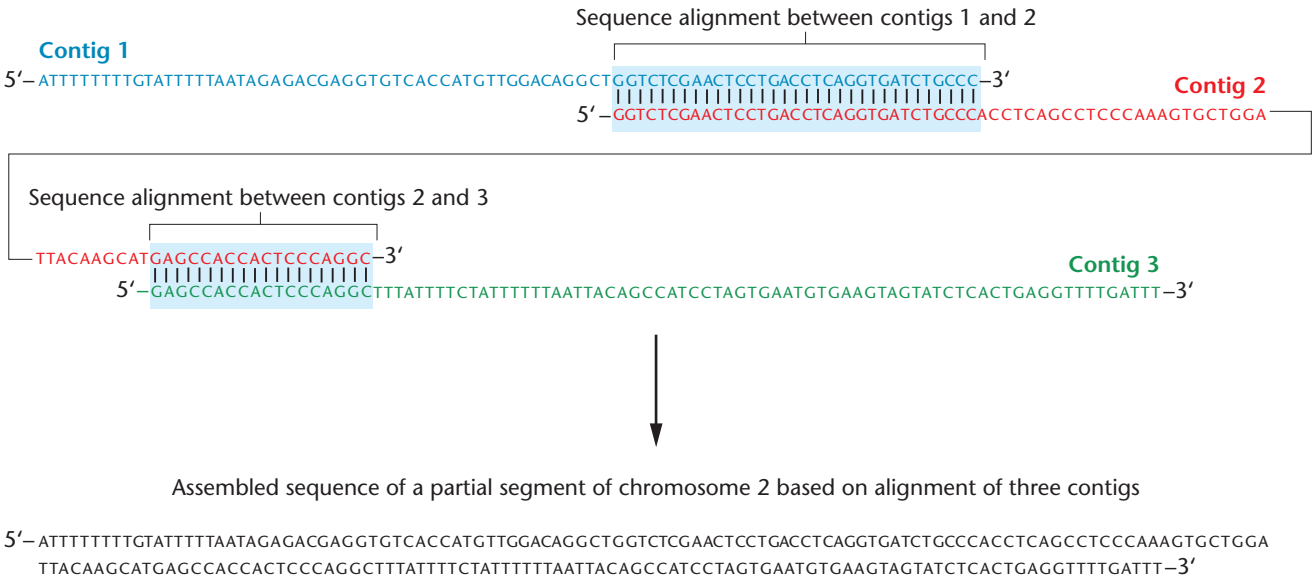


FIGURE 18–2 DNA-sequence alignment of contigs on human chromosome 2. Single-stranded DNA for three different contigs from human chromosome 2 is shown in blue, red, or green. In reality, contig alignment involves fragments that are several thousand bases in length. Alignment of the three contigs allows a portion of chromosome 2 to be assembled. Alignment of all contigs for a particular chromosome would result in assembly of a completely sequenced chromosome.

used. With partial digests, DNA is incubated with restriction enzymes for only a short period of time, so that not every site in a particular sequence is cut to completion by an individual enzyme. Restriction digests of whole chromosomes generate thousands to millions of overlapping DNA fragments. For example, a 6-bp cutter such as *Eco*RI creates about 700,000 fragments when used to digest the human genome!

In the next section, we will discuss the importance of bioinformatics to genomics. One of the earliest bioinformatics applications to be developed for genomic purposes was the use of algorithm-based software programs for creating a DNA-sequence **alignment,** in which similar sequences of bases are lined up for comparison. Alignment identifies overlapping sequences, allowing scientists to reconstruct their order in a chromosome. Because these overlapping fragments are adjoining segments that collectively form one continuous DNA molecule within a chromosome, they are called **contiguous fragments,** or "**contigs**." Figure 18–2 shows an example of contig alignment and assembly for a portion of human chromosome 2. For simplicity, this figure shows relatively short sequences for each contig, which in actuality would be much longer. The figure is also simplified in that, in actual alignments, assembled sequences do not always overlap only at their ends.

The whole-genome shotgun sequencing method was developed by J. Craig Venter and colleagues at The Institute for Genome Research (TIGR). In 1995, TIGR scientists used this approach to sequence the 1.83-million-bp genome of the bacterium *Haemophilus influenzae.* This was the first completed genome sequence from a free-living organism, and it demonstrated "proof of concept" that shotgun sequencing could be used to sequence an entire genome. Even after the genome for *H. influenzae* was sequenced, many scientists were skeptical that a shotgun approach would work on the larger genomes of eukaryotes. Now shotgun approaches are the predominant method for sequencing genomes.

Cutting a genome into contigs is not particularly difficult; however, a primary hurdle that had to be overcome to advance whole-genome sequencing was the question of how to sequence millions or billions of base pairs in a timely and cost-effective way. This was a major challenge for scientists working on the Human Genome Project (Section 18.4). The major technological breakthrough that made genomics possible was the development of computer-automated sequencers.

As we discussed in Chapter 17, next- and third-generation sequencers now enable genome scientists to produce sequence nearly 50,000 times faster than sequencers in 2000 with greater output, improved accuracy, and reduced cost.

ESSENTIAL POINT

Whole-genome shotgun sequencing enables scientists to assemble sequence maps of entire genomes. ■

18.2 DNA Sequence Analysis Relies on Bioinformatics Applications and Genome Databases

Genomics necessitated the rapid development of **bioinformatics,** the use of computer hardware and software and mathematics applications to organize, share, and analyze data related to gene structure, gene sequence and expression, and protein structure and function. However, even before whole-genome sequencing projects had been initiated, a large amount of sequence information from a range of different organisms was accumulating as a result of gene cloning by recombinant DNA techniques. Scientists around the world needed databases that could be used to store, share, and obtain the maximum amount of information from protein and DNA sequences. Thus, bioinformatics software was already being used to compare and analyze DNA sequences and to create private and public databases. Once genomics emerged as a new approach for analyzing DNA, however, bioinformatics became even more important than before. Today, it is a dynamic area of biological research, providing new career opportunities for anyone interested in merging an understanding of biological data with information technology, mathematics, and statistical analysis.

Among the most common applications of bioinformatics are to compare DNA sequences, as in contig alignment; to identify genes in a genomic DNA sequence; to find gene-regulatory regions, such as promoters and enhancers; to identify structural sequences, such as telomeric sequences, in chromosomes; to predict the amino acid sequence of a putative polypeptide encoded by a cloned gene sequence; to analyze protein structure and predict protein functions on the basis of identified domains and motifs; and to deduce evolutionary relationships between genes and organisms on the basis of sequence information.

High-throughput DNA sequencing techniques were developed nearly simultaneously with the expansion of the Internet. As genome data accumulated, many DNA-sequence databases became freely available online. Databases are essential for archiving and sharing data with other researchers and with the public. One of the largest genomic databases, called **GenBank,** is maintained by the National Center for Biotechnology Information (NCBI) in Washington, D.C., and is the largest publicly available database of DNA sequences. GenBank shares and acquires data from databases in Japan and Europe; it contains more than 1.5 trillion bases of sequence data from over 100,000 species; and it doubles in size roughly every 14–18 months! The Human Genome Nomenclature Committee, supported by the NIH, establishes rules for assigning names and symbols to newly cloned human genes. As sequences are identified and genes are named, each sequence deposited into GenBank is provided with an **accession number** that scientists can use to access and retrieve that sequence for analysis.

The NCBI is an invaluable source of public access databases and bioinformatics tools for analyzing genome data. You have already been introduced to NCBI and GenBank through several Exploring Genomics exercises. In Exploring Genomics for this chapter, you will use NCBI and GenBank to compare and align contigs in order to assemble a chromosome segment.

Annotation to Identify Gene Sequences

One of the fundamental challenges of genomics is that, although genome projects generate tremendous amounts of DNA sequence information, these data are of little use until they have been analyzed and interpreted. Genome projects accumulate nucleotide sequences, and then scientists have to make sense of those sequences. Thus, after a genome has been sequenced and compiled, scientists are faced with the task of identifying gene-regulatory sequences and other sequences of interest in the genome so that gene maps can be developed. This process, called **annotation,** relies heavily on bioinformatics, and a wealth of different software tools are available to carry it out.

One initial approach to annotating a sequence is to compare the newly sequenced genomic DNA to the known sequences already stored in various databases. The NCBI provides access to **BLAST** (**Basic Local Alignment Search Tool**), a very popular software application for searching through banks of DNA and protein sequence data. Using BLAST, we can compare a segment of genomic DNA to sequences throughout major databases such as GenBank to identify portions that align with or are the same as existing sequences.

Figure 18–3 shows a representative example of a sequence alignment based on a BLAST search. Here a 280-bp chromosome 12 contig from the rat was used to search a mouse database to determine whether a sequence in the rat contig matched a known gene in mice. Notice that the rat contig (the query sequence in the BLAST search) aligned with base pairs 174,612 to 174,891 of mouse chromosome 8. The accession number for the mouse chromosome sequence, NT_039455.6, is indicated at the top of the figure. BLAST searches calculate a **similarity score**—also called the **identity** value—determined by the sum of identical matches between aligned sequences divided by the total number of bases aligned. Gaps, indicating missing bases in the two sequences, are usually ignored in calculating similarity scores. The aligned rat and mouse sequences were 93 percent similar and showed no gaps in the alignment.

Because this mouse sequence on chromosome 8 is known to contain an insulin receptor gene (encoding a protein that

ref | NT_039455.6 | Mm8_39495_36
Mus musculus chromosome 8 genomic contig, strain C57BL/6J
Features in this part of subject sequence: insulin receptor
Score = 418 bits (226), Expect = 2e⁻114
Identities = 262/280 (93%), Gaps = 0/280 (0%)

```
Query   1       CAGGCCATCCCGAAAGCGAAGATCCCTTGAAGAGGTGGGCAATGTGACAGCCACTACACC  60
                ||||||||||||||||||||||||||||||||||||||||||| |||||||||||||| ||||
Sbjct   174891  CAGGCCATCCCGAAAGCGAAGATCCCTTGAAGAGGTGGGGAATGTGACAGCCACCACACT  174832

Query   61      CACACTTCCAGATTTTCCCAACATCTCCTCCACCATCGCGCCCACAAGCCACGAAGAGCA  120
                |||||||||||||||| |||||| ||||||| ||||| | ||||||||| || || |||||
Sbjct   174831  CACACTTCCAGATTTCCCCAACGTCTCCTCTACCATTGTGCCCACAAGTCAGGAGGAGCA  174772

Query   121     CAGACCATTTGAGAAAGTAGTAAACAAGGAGTCACTTGTCATCTCTGGCCTGAGACACTT  180
                ||| |||||||||||||| || ||||||||||||||||||||||||||||||||||||||
Sbjct   174771  CAGGCCATTTGAGAAAGTGGTGAACAAGGAGTCACTTGTCATCTCTGGCCTGAGACACTT  174712

Query   181     CACTGGGTACCGCATTGAGCTGCAGGCATGCAATCAGGACTCCCCAGAAGAGAGGTGCAG  240
                |||||||||||||||||||||||||||||||||||||||| || |||||||| ||||||||||
Sbjct   174711  CACTGGGTACCGCATTGAGCTGCAGGCATGCAATCAAGATTCCCCAGATGAGAGGTGCAG  174652

Query   241     CGTGGCTGCCTACGTCAGTGCCCGGACCATGCCTGAAGGT  280
                ||||||||||||||||||||||||||||||||||||||||
Sbjct   174651  TGTGGCTGCCTACGTCAGTGCCCGGACCATGCCTGAAGGT  174612
```

FIGURE 18–3 BLAST results showing a 280-base sequence of a chromosome 12 contig from rats (*Rattus norvegicus*, the "query") aligned with a portion of chromosome 8 from mice (*Mus musculus*, the "subject") that contains a partial sequence for the insulin receptor gene. Vertical lines indicate exact matches. The rat contig sequence was used as a query sequence to search a mouse database in GenBank. Notice that the two sequences show 93 percent identity, strong evidence that this rat contig sequence contains a gene for the insulin receptor.

binds the hormone insulin), it is highly likely that the rat contig sequence also contains an insulin receptor gene. We will return to the topic of similarity in Sections 18.3 and 18.6, where we consider how similarity between gene sequences can be used to infer function and to identify evolutionarily related genes through comparative genomics.

Hallmark Characteristics of a Gene Sequence Can Be Recognized during Annotation

A major limitation of this approach to annotation is that it only works if similar gene sequences are already in a database. Fortunately, it is not the only way to identify genes. Whether the genome under study is from a eukaryote or a prokaryote, several hallmark characteristics of protein-coding genes can be searched for using bioinformatics software (**Figure 18–4**). We discussed many of these characteristics of a "typical" gene earlier in the text (see Chapters 12 and 15). For instance, gene-regulatory sequences found upstream of genes are marked by identifiable sequences such as promoters, enhancers, and silencers. Recall from earlier in the text (see Chapter 15) that TATA box, GC box, and CAAT box sequences are often present in the promoter region of eukaryotic genes. Recall also that

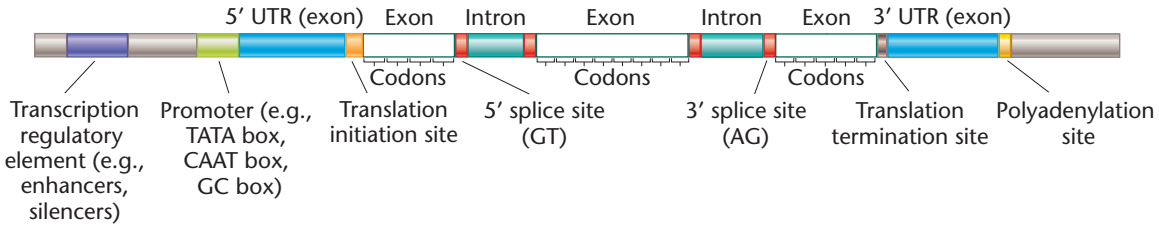


FIGURE 18–4 Characteristics of a protein-coding gene that can be used during annotation to identify a gene in an unknown sequence of genomic DNA. Most eukaryotic genes are organized into coding segments (exons) and noncoding segments (introns). When annotating a genome sequence to determine whether it contains a gene, it is necessary to distinguish between introns and exons, gene-regulatory sequences, such as promoters and enhancers, untranslated regions (UTRs), and gene termination sequences.

splice sites between **exons** and **introns** contain a predictable sequence (most introns begin with CT and end with AG) and such splice-site sequences are important for determining intron and exon boundaries. Interestingly, current estimates indicate that only 6 percent of human genes are transcribed from a single, linear stretch of DNA that does not contain any introns.

Annotation is intended to reveal identifiable features that provide clues to the presence of a protein coding gene. For example, protein-coding genes contain one or more **open reading frames** (**ORFs**), sequences of triplet nucleotides that, after transcription and mRNA splicing, are translated into the amino acid sequence of a protein. ORFs typically begin with an initiation sequence, usually ATG,which transcribes into the AUG start codon of an mRNA molecule, and end with a termination sequence, TAA, TAG, or TGA, which corresponds to the stop codons of UAA, UAG, and UGA in mRNA.

Downstream elements, such as termination sequences and well-defined sequences at the end of a gene, where a polyadenylation sequence signals the addition of a poly-A tail to the 39 end of a mRNA transcript, are also important for annotation (Figure 18–4). Annotation can sometimes be a little bit easier for prokaryotic genes than for eukaryotic genes because there are no introns in prokaryotic genes. Gene-prediction programs are used to annotate sequences. These programs incorporate search elements for many of the criteria mentioned above and have become invaluable applications of bioinformatics.

NOW SOLVE THIS

18–1 In a sequence encompassing 99.4 percent of the euchromatic regions of human chromosome 1, Gregory et al. (Gregory, S.G. et al., *Nature*, 441:315–321, 2006) identified 3141 genes.

(a) How does one identify a gene within a raw sequence of bases in DNA?
(b) What features of a genome are used to verify likely gene assignments?
(c) Given that chromosome 1 contains approximately 8 percent of the human genome, and assuming that there are approximately 20,000 genes, would you consider chromosome 1 to be "gene rich"?

■ **HINT:** *This problem involves a basic understanding of bioinformatics and gene annotation approaches to determine how potential gene sequences can be identified in a stretch of sequenced DNA.*

For more practice, see Problem 11.

ESSENTIAL POINT

Bioinformatics can be used for sequence annotation to identify protein-coding DNA sequencing and noncoding sequences such as regulatory elements. ■

18.3 Genomics Attempts to Identify Potential Functions of Genes and Other Elements in a Genome

As the term suggests, **functional genomics** is the study of gene functions, based on the resulting RNAs or possible proteins they encode, and the functions of other components of the genome, such as gene-regulatory elements. Functional genomics can involve experimental approaches to confirm or refute computational predictions about genome functions (such as the number of protein-coding genes), and it also considers how genes are expressed and the regulation of gene expression.

Predicting Gene and Protein Functions by Sequence Analysis

One approach to assigning functions to genes is to use sequence similarity searches, as described in the previous section. Programs such as BLAST are used to search through databases to find alignments between the newly sequenced genome and genes that have already been identified, either in the same or in different species. You were introduced to this approach for predicting gene function in Figure 18–3, when we demonstrated how sequence similarity to the mouse gene was used to identify a gene in a rat contig as the insulin receptor gene. Inferring gene function from similarity searches is based on a relatively simple idea. If a genome sequence shows statistically significant similarity to the sequence of a gene whose function is known, then it is likely that the genome sequence encodes a protein with a similar or related function.

Another major benefit of similarity searches is that they are often able to identify **homologous genes,** genes that are evolutionarily related. After the human genome was sequenced, many ORFs in it were identified as protein-coding genes based on their alignment with related genes of known function in other species. As an example, **Figure 18–5** compares portions of the human leptin gene (*LEP*) with its homolog in mice (*ob/Lep*). These two genes are over 85 percent identical in sequence. The leptin gene was first discovered in mice (recall our discussion about leptin knockout mice in Chapter 17). The match between the *LEP*-containing DNA sequence in humans and the mouse homolog sequence confirms the identity and leptin-coding function of this gene in human genomic DNA.

If homologous genes in different species are thought to have descended from a gene in a common ancestor, the genes are known as **orthologs.** In Section 18.6 we will consider the globin gene family. Mouse and human α-globin genes are orthologs evolved from a common ancestor.

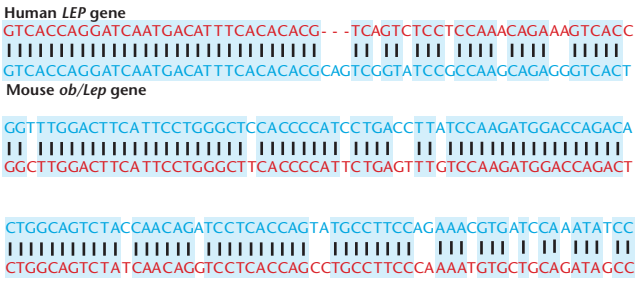


FIGURE 18–5 Comparison of the human *LEP* and mouse *ob/Lep* genes. Partial sequences for these homologs are shown with the human *LEP* gene on top and the mouse *ob/Lep* gene sequence below it. Notice from the number of identical nucleotides, indicated by vertical lines, that the nucleotide sequence for these two genes is very similar. Gaps are indicated by horizontal dashes.

Homologous genes in the same species are called **paralogs.** The α- and β-globin subunits in humans are paralogs resulting from a gene-duplication event. Paralogs often have similar or identical functions.

Predicting Function from Structural Analysis of Protein Domains and Motifs

When a gene sequence is used to predict a polypeptide sequence, the polypeptide can be analyzed for specific structural domains and motifs. Identification of **protein domains,** such as ion channels, membrane-spanning regions, DNA-binding regions, secretion and export signals, and other structural aspects of a polypeptide that are encoded by a DNA sequence, can in turn be used to predict protein function. Recall from earlier in the text (see Chapter 15), for example, that the structures of many DNA-binding proteins have characteristic patterns, or **motifs,** such as the helix-turn-helix, leucine zipper, or zinc-finger motifs. These motifs can often easily be searched for using bioinformatics software, and their identification in a sequence is a common strategy for inferring the possible functions of a protein.

ESSENTIAL POINT

Functional genomics predicts gene function based on sequence analysis. ■

18.4 The Human Genome Project Revealed Many Important Aspects of Genome Organization in Humans

Now that you have a general idea of the basic strategies used for analyzing a genome, let's look at the largest genomics project completed to date. The **Human Genome Project** (**HGP**) was a coordinated international effort to determine the sequence of the human genome and to identify all the genes it contains. It has produced a plethora of information, much of which is still being analyzed and interpreted. What is clear from all the different kinds of genomes sequenced is that humans and all other species share a common set of genes essential for cellular function and reproduction, confirming that all living organisms arose from a common ancestor.

Origins of the Project

The publicly funded Human Genome Project began in 1990 under the direction of James Watson, the co-discoverer of the double-helix structure of DNA. Eventually the public project was led by Dr. Francis Collins, who had previously led a research team involved in identifying the *CFTR* gene as the cause of cystic fibrosis. In the United States, the Collins-led HGP was coordinated by the Department of Energy and the National Center of Human Genome Research, a division of the National Institutes of Health. It established a 15-year plan with a proposed budget of $3 billion to identify all human genes, originally thought to number between 80,000 and 100,000, to sequence and map them all, and to sequence the approximately 3 billion base pairs thought to comprise the 24 chromosomes (22 autosomes, plus X and Y) in humans. Other primary goals of the HGP included the following:

- To establish functional categories for all human genes
- To analyze genetic variations between humans, including the identification of **single-nucleotide polymorphisms** (**SNPs**)
- To map and sequence the genomes of several model organisms used in experimental genetics, including *E. coli*, *S. cerevisiae*, *C. elegans*, *D. melanogaster,* and *M. musculus* (mouse)
- To develop new sequencing technologies, such as high-throughput computer-automated sequencers, in order to facilitate genome analysis
- To disseminate genome information among both scientists and the general public

Lastly, to deal with the impact that genetic information would have on society, the HGP set up the ELSI program (standing for Ethical, Legal, and Social Implications) to consider ethical, legal, and social issues arising from the HGP and to ensure that personal genetic information would be safeguarded and not used in discriminatory ways.

As the HGP grew into an international effort, scientists in 18 countries were involved in the project. Much of the work was carried out by the International Human Genome Sequence Consortium, involving nearly 3000 scientists

working at 20 centers in six countries (China, France, Germany, Great Britain, Japan, and the United States).

In 1999, a privately funded human genome project led by J. Craig Venter at Celera Genomics (aptly named from a word meaning "swiftness") was announced. Celera's goal was to use whole-genome shotgun sequencing and computer-automated high-throughput DNA sequencers to sequence the human genome more rapidly than HGP. The public project had proposed using a clone-by-clone approach to sequence the genome. Recall that Venter and colleagues had proven the potential of shotgun sequencing in 1995 when they completed the genome for *H. influenzae*. Celera's announcement set off an intense competition between the two teams, which both aspired to be first with the human genome sequence. This contest eventually led to the HGP finishing ahead of schedule and under budget after scientists from the public project began to use high-throughput sequencers and whole-genome sequencing strategies as well.

Major Features of the Human Genome

In June 2000, the leaders of the public and private genome projects met at the White House with President Clinton and jointly announced the completion of a draft sequence of the human genome. In February 2001, they each published an analysis covering about 96 percent of the euchromatic region of the genome. The public project sequenced euchromatic portions of the genome 12 times and set a quality control standard of a 0.01 percent error rate for their sequence. Although this error rate may seem very low, it still allows about 600,000 errors in the human genome sequence. Celera sequenced certain areas of the genome more than 35 times when compiling the genome.

The remaining work to complete the genome consisted of filling in gaps clustered around centromeres, telomeres, and repetitive sequences (regions rich in GC base pairs can be particularly tough to sequence and interpret), correcting misaligned segments, and re-sequencing portions of the genome to ensure accuracy. In 2003 genome sequencing and error fixing were deemed sufficient to pass the international project's definition of completion—that it contained fewer than 1 error per 10,000 nucleotides and that it covered 95 percent of the gene-containing portions of the genome. Yet even at the time of "completion" there were still some 350 gaps in the sequence that continued to be worked on.

And of course the HGP did not sequence the genome of every person on Earth. The assembled genomes largely consist of haploid genomes pooled from different individuals so that they provide a *reference genome* representative of major, common elements of a human genome widely shared among populations of humans. Examples of major features of the human genome are summarized in **Table 18.1.** As you can see in this table, many unexpected observations have provided us with major new insights. The genome is not static! Genome variations, including the abundance of repetitive sequences scattered throughout the genome, verify that the genome is indeed dynamic, revealing many evolutionary examples of sequences

TABLE 18.1 Major Features of the Human Genome

- The human genome contains 3.1 billion nucleotides, but protein-coding sequences make up only about 2 percent of the genome.
- The genome sequence is ~99.9 percent similar in individuals of all nationalities. SNPs and copy number variations (CNVs) account for genome diversity from person to person.
- The genome is dynamic. At least 50 percent of the genome is derived from transposable elements, such as SINEs, LINEs, and *Alu* sequences, retrotransposons, and other repetitive DNA sequences.
- The human genome contains approximately 20,000 protein-coding genes, far fewer than the originally predicted number of 80,000–100,000 genes.
- The average size of a human gene is ~25 kb, including gene-regulatory regions, introns, and exons. On average, mRNAs produced by human genes are ~3000 nt long.
- Many human genes produce more than one protein through alternative splicing, thus enabling human cells to produce a much larger number of proteins (perhaps as many as 200,000) from only ~20,000 genes.
- More than 50 percent of human genes show a high degree of sequence similarity to genes in other organisms; however, more than 40 percent of the genes identified have no known molecular function.
- Human genes created by duplication events are evident in gene families.
- Genes are not uniformly distributed on the 24 human chromosomes. Gene-rich clusters are separated by gene-poor "deserts" that account for 20 percent of the genome. These deserts correlate with G bands seen in stained chromosomes. Chromosome 19 has the highest gene density, and chromosome 13 and the Y chromosome have the lowest gene densities.
- Chromosome 1 contains the largest number of genes, and the Y chromosome contains the smallest number.
- Human genes are larger and contain more and larger introns than genes in the genomes of invertebrates, such as *Drosophila*. The largest known human gene encodes dystrophin, a muscle protein. This gene, associated in mutant form with muscular dystrophy, is 2.5 Mb in length (Chapter 14), larger than many bacterial chromosomes. Most of this gene is composed of introns.
- The number of introns in human genes ranges from 0 (in histone genes) to 234 (in the gene for *titin,* which encodes a muscle protein).

that have changed in structure and location. In many ways, the HGP has revealed just how little we know about our genome.

Two of the biggest surprises discovered by the HGP were that less than 2 percent of the genome codes for proteins and that there are only around 20,000 protein-coding genes. Recall that the number of genes had originally been estimated to be about 100,000, based in part on a prediction that human cells produce about 100,000 proteins. At least half of the genes show sequence similarity to genes shared by many other organisms, and as you will learn in Section 18.7, a majority of human genes are similar in sequence to genes from closely related species such as chimpanzees. There is still no consensus among scientists worldwide about the exact number of human genes. One reason is that it is unclear whether or not many of the presumed genes produce functional proteins. Genome scientists continue to annotate the genome, and as mentioned earlier, functional genomics studies have important roles in determining whether or not computational predictions about the number of protein-coding and non–protein-coding genes are accurate.

The number of genes is much lower than the number of predicted proteins in part because many genes code for multiple proteins through **alternative splicing.** Recall from earlier in the text (see Chapter 12), that alternative splicing patterns can generate multiple mRNA molecules, and thus multiple proteins, from a single gene, through different combinations of intron–exon splicing arrangements. Initial estimates suggested that over 50 percent of human genes undergo alternative splicing to produce multiple transcripts and multiple proteins. Recent studies suggest that ~94–95 percent of human pre-mRNAs contain multiple exons that are processed to produce multiple transcripts and potentially multiple different protein products. Clearly, alternative splicing produces an incredible diversity of proteins beyond simple predictions based on the number of genes in the human genome.

Functional categories have been assigned for human genes, primarily on the basis of (1) functions determined previously (for example, from recombinant DNA cloning of human genes and known mutations involved in human diseases), (2) comparison to known genes and predicted protein sequences from other species, and (3) predictions based on annotation and analysis of protein functional domains and motifs (**Figure 18–6**). Although functional categories and assignments continue to be revised, the functions of over 40 percent of human genes remain unknown. Determining human gene functions, deciphering complexities of gene-expression regulation and gene interaction, and uncovering the relationships between human genes and phenotypes are among the many challenges for genome scientists.

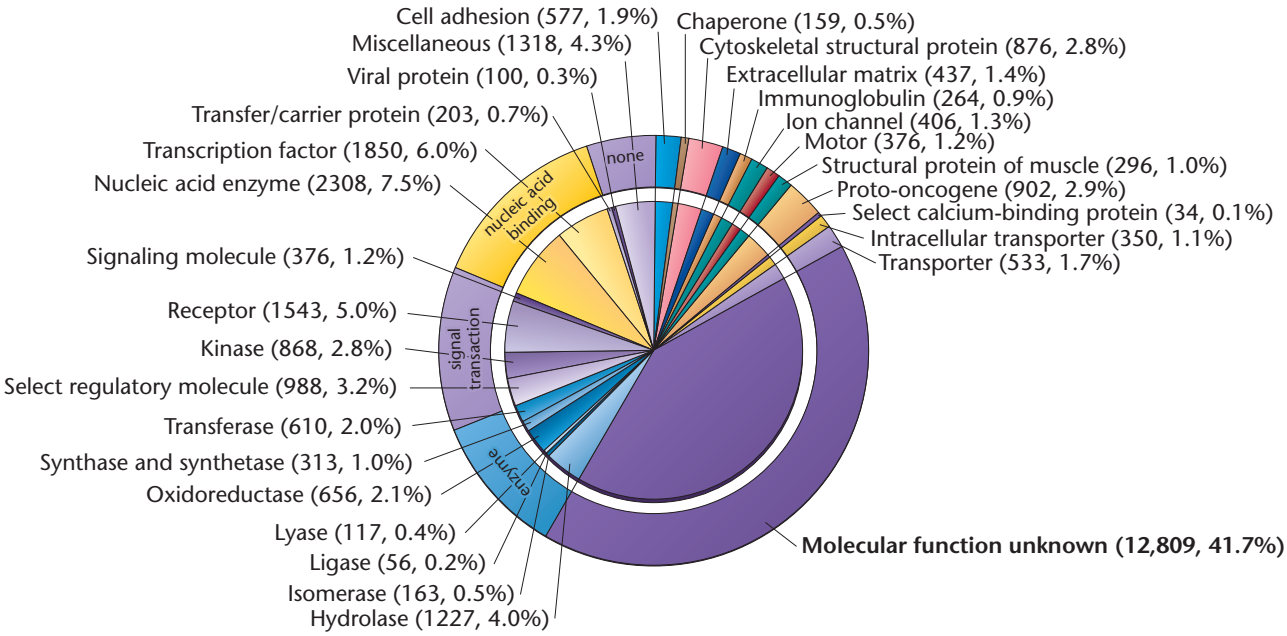


FIGURE 18–6 A representation of the functional categories to which genes in the human genome have been assigned on the basis of similarity to proteins of known function. Among the most common genes are those involved in nucleic acid metabolism (7.5 percent of all genes identified), transcription factors (6.0 percent), receptors (5 percent), hydrolases (4 percent), protein kinases (2.8 percent), and cytoskeletal structural proteins (2.8 percent). A total of 12,809 predicted proteins (41 percent) have unknown functions, indicative of the work that is still needed to fully decipher our genome.

Individual Variations in the Human Genome

The HGP has also shown us that in all humans, regardless of racial and ethnic origins, the genomic sequence is approximately 99.9 percent the same. As we discuss in other chapters, most genetic differences between humans result from **single-nucleotide polymorphisms** (**SNPs**) and **copy number variations** (**CNVs**). Recall that SNPs are single-base changes in the genome and variations of many SNPs are associated with disease conditions. For example, SNPs cause sickle-cell anemia and cystic fibrosis. Later in the text (see Chapter 19), we will examine how SNPs can be detected and used for diagnosis and treatment of disease.

After the draft sequence of the human genome was completed, it initially appeared that most genetic variations between individuals (the 0.1 percent differences) were due to SNPs. While SNPs are important contributing factors to genome variation, structural differences that we discussed earlier in the text (see Chapter 11) such as deletions, duplications, inversions, and CNVs, which can span millions of bp of DNA, play much more important roles in genome variation than previously thought. As we discussed earlier in the text (see Chapters 6 and 11), recall that CNVs are duplications or deletions of relatively large sections of DNA on the order of several hundred or several thousand base pairs. Many of the CNVs that vary the most among genomes appear to be at least 1 kilobase.

Although most human DNA is present in two copies per cell, one from each parent, CNVs are segments of DNA that are duplicated or deleted, resulting in variations in the number of copies of a DNA segment inherited by individuals. In some cases CNVs are major deletions removing entire genes; other deletions affect gene function by frameshifts in the reading code. CNV sequences that are duplicated can result in overexpression of a particular gene, yet many deleted and duplicated CNVs do not present clearly identifiable phenotypes.

Current estimates of the number of CNVs in an individual genome range from about 12 CNVs to perhaps 4–5 dozen per person. Some studies estimate that there may be as many as 1500 CNVs greater than 1 kb among the human genome. Other studies claim there are more than 1.5 million deletions of less than 100 bp that contribute to genome variation between individuals.

Accessing the Human Genome Project on the Internet

It is now possible to access databases and other sites on the Internet that display maps for all human chromosomes. You will visit a number of these databases in Exploring Genomics exercises. **Figure 18–7(a)** displays a partial gene map for chromosome 12 that was taken from an NCBI database called Map Viewer. You may already have used Map Viewer for the Exploring Genomics exercises ealier in the text (see Chapter 7). This image shows an ideogram, or cytogenetic map, of chromosome 12. To the right of the ideogram is a column showing the contigs (arranged lying vertically) that were aligned to sequence this chromosome. The Hs UniG column displays a histogram representation of gene density on chromosome 12. Notice that relatively few genes are located near the centromere. Gene symbols, loci, and gene names (by description) are provided for selected genes; in this figure only 20 genes are shown. When accessing these maps on the Internet, one can magnify, or zoom in on, each region of the chromosome, revealing all genes mapped to a particular area.

You can see that most of the genes listed here have been assigned descriptions based on the functions of their products, some of which are transmembrane proteins, some enzymes such as kinases, some receptors, including several involved in olfaction, and so on. Other genes are described in terms of hypothetical products; they are presumed to be genes based on the presence of ORFs, but their function remains unknown [Figure 21–1(a)].

The HGP's most valuable contribution will perhaps be the identification of disease genes and the development of new treatment strategies as a result. Thus, extensive maps have been developed for genes implicated in human disease conditions. The disease gene map of chromosome 21 shown in **Figure 18–7(b)** indicates genes involved in amyotrophic lateral sclerosis (ALS), Alzheimer disease, cataracts, deafness, and several different cancers. Later in the text (see Chapter 19) we discuss implications of the HGP for the identification of genes involved in human genetic diseases, and for disease diagnosis, detection, and gene therapy applications.

ESSENTIAL POINT

The Human Genome Project revealed many surprises about human genetics, including gene number, the high degree of DNA sequence similarity between individuals and between humans and other species, and showed that many genes encode multiple proteins. ■

18.5 After the Human Genome Project: What Is Next?

The Human Genome Project and the development of genomics techniques have been largely responsible for launching a new era of biological research—the era of "omics." It seems that every year, more areas of biological

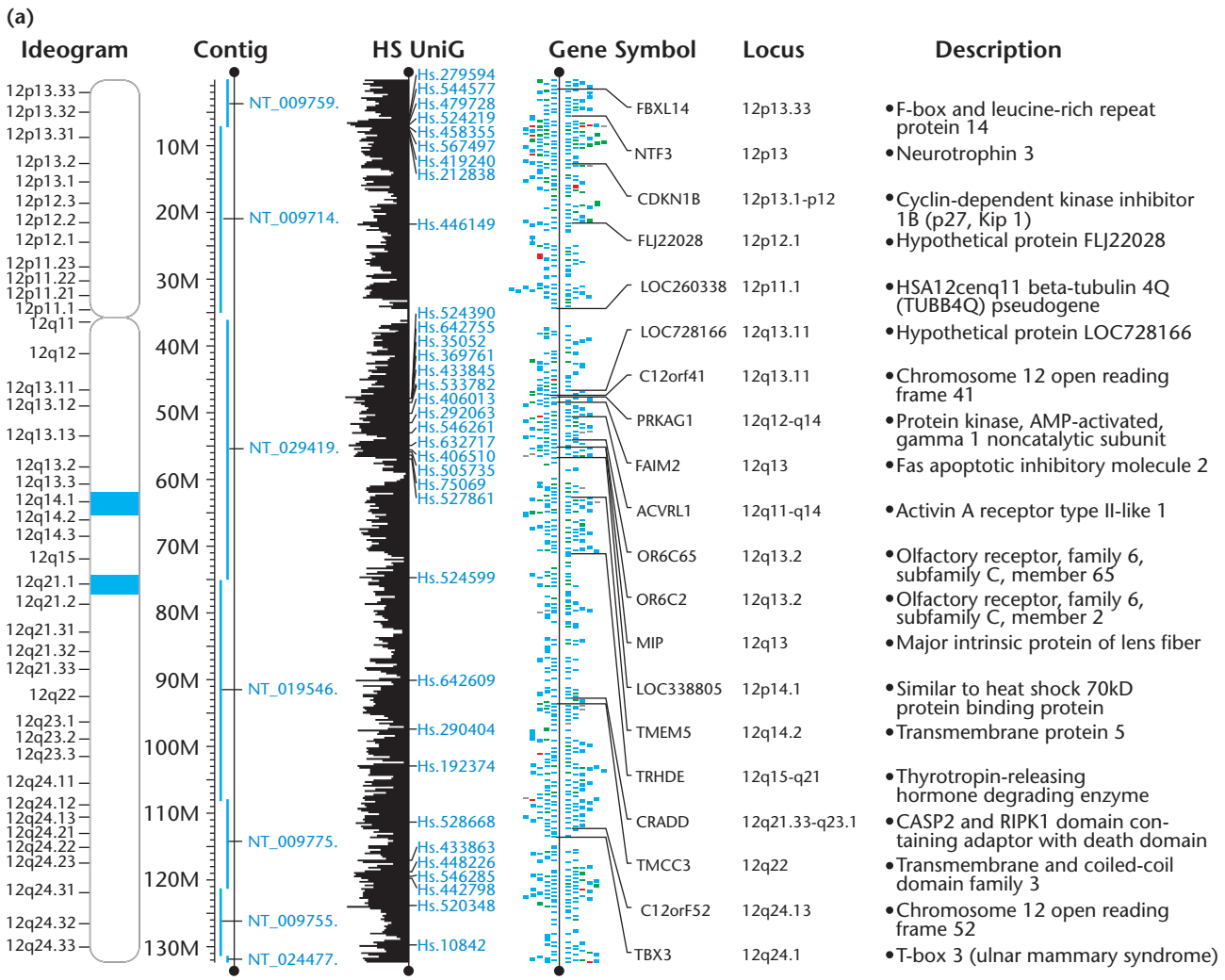


FIGURE 18–7 (a) A gene map for chromosome 12 from the NCBI database Map Viewer. (b) Partial map of disease genes on human chromosome 21. Maps such as this depict genes thought to be involved in human genetic disease conditions.

research are being described as having an omics connection. Some examples of "omics" are

- proteomics—the analysis of all the proteins in a cell or tissue
- metabolomics—the analysis of proteins and enzymatic pathways involved in cell metabolism
- glycomics—the analysis of the carbohydrates of a cell or tissue
- toxicogenomics—the analysis of the effects of toxic chemicals on genes, including mutations created by toxins and changes in gene expression caused by toxins
- metagenomics—the analysis of genomes of organisms collected from the environment
- nutrigenomics—understanding the relationships between genes and diet
- pharmacogenomics—the development of customized medicine based on a person's genetic profile for a particular condition
- transcriptomics—the analysis of all expressed genes in a cell or tissue

We will consider several of these genomics disciplines in other parts of this chapter.

Since completion of a reference sequence of the human genome, studies have continued at a very rapid pace. For example, as a result of the HGP, many other major theme areas for human genome research have emerged, including cancer genome projects, analysis of the epigenome (including a Human Epigenome Project that is creating hundreds of maps of epigenetic changes in different cell and tissue types and evaluating potential roles of epigenetics in complex diseases), characterization of SNPs (the International HapMap Project) and CNVs for their role in genome variation, disease, and pharmacogenomics applications. We have discussed aspects of a cancer genome project (Cancer Genome Atlas Project) earlier in the text (see Chapter 16). The epigenome is covered in depth later in the text (see Special Topic Chapter 1—Epigenetics). SNPs and pharmacogenomics are discussed later as well (see Special Topic Chapter 4—Genomics and Personalized Medicine). Here we consider several examples of genome research that are extensions of the HGP.

ESSENTIAL POINT

Genomics has led to a number of other related "omics" disciplines that are rapidly changing how modern biologists study DNA, RNA, and proteins and many aspects of cell function. ■

Personal Genome Projects and Personal Genomics

As we discussed earlier in this chapter and earlier in the text (see Chapter 17), new sequencing technologies, capable of generating longer sequence reads at higher speeds with greater accuracy, have greatly reduced the cost of DNA sequencing, and expectations for continued cost reductions along with continued technological advances are high (see **Figure 18–8**). These expectations led several companies to

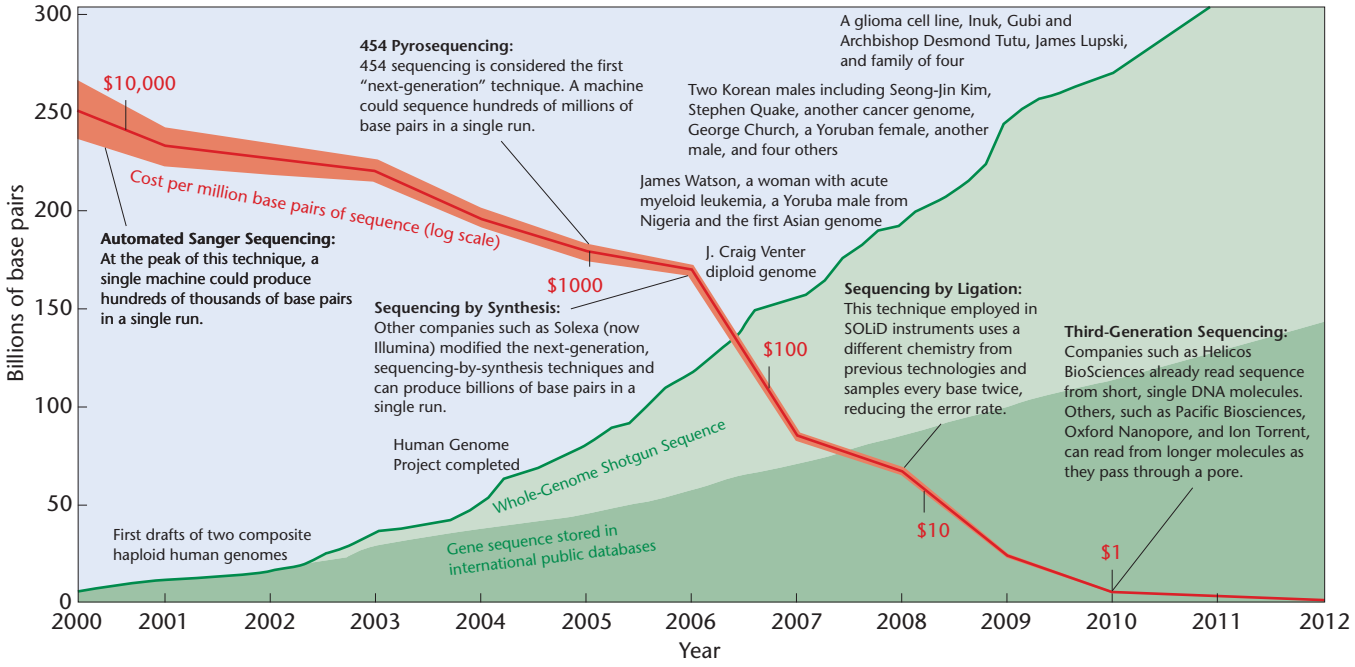


FIGURE 18–8 Human genome sequence explosion. Sequencing costs have steadily declined since 2000 due to innovations in sequencing technology. As a result, notice that the amount of whole-genome shotgun sequencing data stored in public databases—which include data on several individual genomes—has dramatically increased.

propose WGS for individual people—a personal genomics approach. competition. Two programs funded by the National Institutes of Health challenged scientists to develop sequencing technologies to complete a human genome for $1000 by 2014 (see the "Genetics, Technology, and Society" essay in Chapter 19. Since 2005, the cost of DNA sequencing has dropped from about $1000 per megabase to ten cents per megabase!).

In 2012, Life Technologies announced that their Ion Proton technology was used to sequence a genome for $1000. Whether the $1000 mark represents the costs of reagents to sequence a genome or actual costs when sequence preparation, labor, and analysis of the genome are taken into account can be debated. Whether the accuracy and completeness of the sequence coverage reported by Life Technologies is sufficient to definitively state that the $1000 genome threshold has been achieved has been challenged by other scientists.

As you will learn later in the text (see Chapter 19), having somebody such as a geneticist analyze genome data and consider how genome variations may affect a person's health takes a lot of time and money. So even if the cost of sequencing a person's genome is less than $1000, interpreting genome data to make sense for medical treatment may cost hundreds of thousands of dollars. Pursuit of the $1000 genome was an indicator that DNA sequencing may eventually be affordable enough for individuals to consider acquiring a readout of their own genetic blueprint. The genome of James D. Watson, who together with Francis Crick discovered the structure of DNA, was the focus of "Project Jim" by the Connecticut company 454 Life Sciences, which wanted to sequence the genome of a high-profile person and decided that the co-discoverer of DNA structure and the first director of the U.S. Human Genome Project should be that person.

Human genome pioneer J. Craig Venter had his genome completed by the J. Craig Venter Institute and deposited into GenBank in May 2007. George Church of Harvard and his colleagues started a **Personal Genome Project** (**PGP**) and recruited volunteers to provide DNA for individual genome sequencing on the understanding that the genome data will be made publicly available. Church's genome has been completed and is available online. The concept of a personal genome project raises the obvious question: would you have your genome sequenced for $10,000, $1000, or even for free?

Since the Watson and Venter genomes were completed, in 2008 the first complete genome sequence was provided for an individual "ancient" human, a Palaeo-Eskimo, obtained from ~4000-year-old permafrost-preserved hair. This work recovered about 78 percent of the diploid genome and revealed many interesting SNPs (of which about 7 percent have not been previously reported). As of 2014, more than 30,000 individual human genomes have been sequenced.

Exome Sequencing

The focus of personal genome projects has shifted toward **exome sequencing,** sequencing the 180,000 exons in a person's genome. This can be done at a cost of less than $1000 with $<100\times$ coverage. Exome sequencing reveals mutations involved in disease by focusing only on exons as protein-coding segments of the genome. Of course, a limitation of this approach is its failure to identify mutations in gene-regulatory regions that influence gene expression. As an example, a group of scientists called the 1000 Genomes Project Consortium reported on the genomes of 1092 individuals from 14 populations representing Europe, East Asia, sub-Saharan Africa, and the Americas. Whole-genome and exome-sequencing data revealed more than 38 million SNPs and many other structural variations (CNVs). One interpretation of this work is that it reveals clear variations in individuals and associates particular diseases with geographic or ancestral background. Thus sequencing genomes of individuals from diverse populations can help us better understand the spectrum of human genetic variation and to learn the causes of genetic diseases across diverse groups. We will come back to the topic of exome sequencing later in the text (see Chapter 19), when we discuss genetic testing.

Another particularly beneficial aspect of personal genome projects is the insight they are providing into genome variation. The HGP combined samples from different individuals to create a reference genome for a *haploid genome.* Personal genome projects sequence a diploid genome; consequently, such projects indicate that haploid genome comparisons may underestimate the extent of genome variation between individuals by five-fold or more. For example, when Venter's genome was analyzed, over 4 million variations were found between his maternal and paternal chromosomes alone. From what we are learning about personal genomes, genome variation between individuals may be closer to 0.5 percent than 0.1 percent, and in a 3-billion-bp genome this is a significant difference in sequence variation. Integrating genome data from several complete individual genomes of individuals from different ethnic groups will also be of great value in evolutionary genetics to address fundamental questions about human diversity, ancestry, and migration patterns.

In a related matter, PGPs are revealing that there can be significant *mosaicism* in human somatic cells. Thus, cells in an individual person do not all contain identical genomes. Because of the sophistication of WGS methods, mosaicisms for SNPs and CNVs have been found in skin, brain, blood, and stem cells from the same individual. We

are only beginning to understand the frequency and effects of genetic mosaicism on health and disease.

Later in the text (see Chapter 19), we will consider how various approaches to personal genomics can be used for genetic testing.

ESSENTIAL POINT

Personal genome sequencing and exome sequencing will provide insight into individual variations in genomes and has tremendous potential for diagnosis and treatment of genetic diseases. ■

Encyclopedia of DNA Elements (ENCODE) Project

In 2003, a few months after the announcement that the human genome had been sequenced, a group of about three dozen research teams around the world began the **Encyclopedia of DNA Elements (ENCODE) Project.** A main goal of ENCODE was to use both experimental approaches and bioinformatics to identify and analyze functional elements of the genome, such as transcriptional start sites, promoters, and enhancers, which regulate the expression of human genes.

Recall from our previous discussions that only a relatively small percentage (less than 2 percent) of the human genome codes for proteins. ENCODE focused not on genes but on all of the sequences, commonly referred to as "junk" DNA. So what are all of the other bases in the genome doing? We know that such sequences are important for chromosome structure, the regulation of gene expression, and other roles. Just because these sequences themselves do not code for protein does not mean that they are all unimportant. Non—protein-coding sequences are discussed in greater detail later in the text (see Special Topic Chapter 2—Emerging Roles of RNA).

ENCODE studied gene expression in 147 different cell types because genome activity differs from cell to cell. After about a decade of research and a cost of $288 million, in 2012 a group of 30 research papers were published revealing the major findings of the ENCODE project. Highlights of what ENCODE revealed include the following.

- The majority, ~80 percent, of the human genome is considered functional. This is partly because large segments of the genome are transcribed into RNA. Most of these RNAs do not encode proteins. These various RNAs include tRNA, rRNAs, and miRNAs. For example, at least 13,000 sequences specify long noncoding RNAs (lncRNAs). Other reports suggest there may be over 17,000 lncRNAs. It may turn out that the number of noncoding RNA sequences will outnumber protein-coding genes.
- The functional sequences also include gene-regulatory regions: ~70,000 promoter regions and nearly 400,000 enhancer regions.
- There are 20,687 protein-coding genes in the human genome.
- A total of 11,224 sequences are characterized as pseudogenes, previously thought to be inactive in all individuals. Some of these are inactive in most individuals but occasionally active in certain cell types of some individuals, which may eventually warrant their reclassification as active, transcribed genes and not pseudogenes.
- SNPs associated with disease are enriched within noncoding functional elements of the genome, often residing near protein-coding genes.

The ENCODE findings have broadly defined the functional roles of the genome to include encoding proteins or noncoding RNAs and displaying biochemical properties such as binding regulatory proteins that influence transcription or chromatin structure. A relatively large body of geneticists and other scientists do not agree with ENCODE's definition of functional sequences. One reason cited is that ENCODE did not adequately address many of the repetitive sequences in the genome such as transposons, LINEs, SINEs, and other sequences such as telomeres and centromeres. There has also been significant debate about the value of ENCODE, given the cost of the project. But research teams are already using information from ENCODE to identify risk factors for certain diseases, with the hopes of developing appropriate cures and treatments.

EVOLVING CONCEPT OF A GENE

Based on the work of the ENCODE project, we now know that DNA sequences that have previously been thought of as "junk DNA," which do not encode proteins, are nonetheless often transcribed into what we call noncoding RNA (ncRNA). Since the function of some these RNAs is now being determined, we must consider whether the concept of the gene should be expanded to include DNA sequences that encode ncRNAs. At this writing, there is no consensus, but it is important for you to be aware of these current findings as you develop your final interpretation of a gene. ■

The Human Microbiome Project

In 2007 the National Institutes of Health announced plans for the **Human Microbiome Project (HMP)**, a $170 million project to complete the genomes of an estimated 600—1000 microorganisms, bacteria, viruses, and yeast that live on and inside humans. Microorganisms outnumber human cells by about 10 to 1. Many microbes, such as *E. coli* in the digestive tract, have important roles in human health, and

of course other microbes make us ill. The HMP has several major goals, including:

- Determining if individuals share a core human microbiome.
- Understanding whether changes in the microbiome can be correlated with changes in human health.
- Developing new methods, including bioinformatics tools, to support analysis of the microbiome.
- Addressing ethical, legal, and social implications raised by human microbiome research. Does this sound familiar? Recall that addressing ethical, legal, and social issues was a goal of the HGP.

The HMP has involved about 200 scientists at 80 institutions. In 2012 a series of papers were published summarizing recent findings from the HMP. The HMP analyzed 15 body sites from males and 18 sites from females from 242 healthy individuals in the United States and applied WGS of genomes for the microbes and viruses present at these sites. Each person was sampled up to three times over nearly two years. Researchers used bioinformatics to compare microbial and viral genome sequences obtained to sequences in publicly available databases. In addition to WGS analysis, sequences for 16*S* rRNA gene sequences in particular were used to compare bacterial samples. More than 2000 microbial sequences isolated from the human body have been sequenced to date.

The HMP has amassed more than 1000 times the sequencing data generated by the Human Genome Project. What concepts have we formulated about the human microbiome so far?

- Sequence data from the HMP have identified an estimated 81 to 99 percent of the microbes and viruses distributed among body areas in human males and females.
- As many as 1000 bacterial strains may be present in each person.
- An estimated 10,000 bacterial species may be part of the human microbiome.
- The microbiome starts at birth. Babies pick up bacteria from their mothers' microbiome.
- A surprise to HMP scientists is that the microbiome can be substantially different from person to person. Also, sequences for disease-causing bacteria are present in everyone's microbiome.
- In the human gut, for example, although the microbiome differs from person to person, it remains relatively stable over time in individuals.

There is no single "reference" human microbiome to which people can be compared. Microbial diversity varies greatly from individual to individual, and a personalization of the microbiome occurs in individuals. For instance, comparing sequences of the microbiomes from two healthy people of equivalent age reveals microbiomes that can be quite different. There are, however, similarities in certain parts of the body, with signature bacteria and characteristic genes associated with a particular location in the body.

Knowledge about the personalized nature of the microbiome will be valuable for improving human health and medicine, which in the future may include microbiome-specific therapeutic drugs. Scientists are trying to establish criteria for a healthy microbiome, which is expected to help determine, for example, how bacteria help maintain normal health, how antibiotics can disturb a person's microbiome, and why certain individuals are susceptible to certain diseases, especially chronic conditions such as psoriasis, irritable bowel syndrome, and potentially even obesity.

Related to this project, a team of researchers at the University of California, Los Angeles, analyzed DNA sequences from 101 college students, 49 of whom had acne and 52 of whom did not. Over 1000 strains of *Propionibacterium acnes* (*P. acnes*) were isolated. Using WGS and bioinformatics, researchers clustered these strains into ten strain types (related strains). Six of these types were more common among acne-prone students, and one type appeared repeatedly in skin samples from students without acne. Sequence analysis of types associated with acne indicated gene clusters that may contribute to the skin disease. Further analysis of these strain types may help dermatologists develop new drugs targeted at killing acne-causing strains of *P. acnes*.

No Genome Left Behind and the Genome 10K Plan

Without question, new sequencing technologies that have been developed as a result of the HGP are an important part of the transformational effect the HGP has had on modern biology. In the late 1990s, a room full of sequencers and several million dollars were required to sequence the 97-Mb genome of *C. elegans*. As a sign of modern times in the world of genomics, recently two sequencers and $500,000 produced a reasonably complete draft of the 750-Mb cod genome—in a month!

Recent headline-grabbing genomes that have been completed include:

- the tomato, which has 31,760 genes, more genes than humans!
- the potato, a vegetable that shares 92 percent of its DNA with tomatoes, a fruit

- chickpea, the second most widely grown legume after the soybean
- the red-spotted newt, which has a genome of almost 10 billion base pairs!

Modern sequencing technologies are asking some to consider the question, "What would you do if you could sequence everything?" Partners around the world, including genome scientists and museum curators, have proposed sequencing 10,000 vertebrate genomes, the **Genome 10K** plan. Shortly after the HGP finished, the National Human Genome Research Institute (NHGRI) assembled a list of mammals and other vertebrates as priorities for genome sequencing in part because of their potential benefit for learning about the human genome through comparative genomics. Genome 10K will also provide insight into genome evolution and speciation.

Stone-Age Genomics

In yet another example of how genomics has taken over areas of DNA analysis, a number of labs around the world are involved in analyzing "ancient" DNA. These so-called **stone-age genomics** studies are generating fascinating data from miniscule amounts of ancient DNA obtained from bone and other tissues such as hair that are tens of thousands to about 700,000 years old, and often involve samples from extinct species. Analysis of DNA from a 2400-year-old Egyptian mummy, bison, mosses, platypus, mammoths, Pleistocene-age cave bears and polar bears, coelacanths and Neanderthals are some of the most prominent examples of stone-age genomics. In 2013, scientists reported the oldest complete genome sequence generated to date. It came from a 700,000-year-old bone fragment from an ancient horse uncovered from the frozen ground in the Yukon Territory of Canada. This result is interesting in part because evolutionary biologists have used genomic data to estimate that ancient ancestors of modern horses branched off from other animal lineages around 4 million years ago— about twice as long ago as prior estimates.

In 2005, researchers from McMaster University in Canada and Pennsylvania State University published about 13 million bp from a 27,000-year-old woolly mammoth. This study revealed a ~98.5 percent sequence identity between mammoths and African elephants. Subsequent studies by other scientists have used whole-genome shotgun sequencing of mitochondrial and nuclear DNA from Siberian mammoths to provide data on the mammoth genome. These studies suggest that the mammoth genome differs from the African elephant by as little as 0.6 percent. These studies are also great demonstrations of how stable DNA can be under the right conditions, particularly when frozen. In the future, it may be possible to produce complete genome sequences from samples that are several million years old.

Perhaps even more intriguing are similarities that have been revealed between the mammoth and human genomes. For example, 18–8, when the gene sequences from human chromosomes were aligned with sequences from the mammoth genome, approximately 50 percent of mammoth genes showed sequence alignment with human genes on autosomes.

In Section 18.6 we will discuss recent work on the Neanderthal genome. Obtaining the genome of a human ancestor this old was previously unimaginable. This work is providing new insights into our understanding of human evolution.

ESSENTIAL POINT

Since completion of the Human Genome Project, human genome research has focused on individual human genomes (personalized genomics) and other efforts such as the Human Microbiome Project. ■

18.6 Comparative Genomics Analyzes and Compares Genomes from Different Organisms

As of 2014, over 4400 whole genomes have been sequenced—including many model organisms and a number of viruses. About 200 of the completed genomes are from eukaryotes. This is quite extraordinary progress in a relatively short time span. Among these organisms are yeast (*Saccharomyces cerevisiae)*—the first eukaryotic genome to be sequenced to bacteria such as *E. coli,* the nematode roundworm (*Caenorhabditis elegans)*, the thale cress plant (*Arabidopsis thaliana)*, mice (*Mus musculus*), zebrafish (*Danio rerio*), and of course *Drosophila.* In the past few years, genomes for chimpanzees, dogs, chickens, gorillas, sea urchins, honey bees, pigs, pufferfish, rice, and wheat have all been sequenced.

These studies have demonstrated not only significant differences in genome organization between prokaryotes and eukaryotes but also many similarities between genomes of nearly all species. Similar gene sets are used by organisms for basic cellular functions, such as DNA replication, transcription, and translation. These genetic relationships are the rationale for using model organisms to study inherited human disorders, the effects of the environment on genes, and interactions of genes in complex diseases, such as cardiovascular disease, diabetes, neurodegenerative conditions, and behavioral disorders.

In this section we discuss interesting aspects of genomes in selected organisms.

Comparative genomics compares the genomes of different organisms to answer questions about genetics and other aspects of biology. It is a field with many research and practical applications, including gene discovery and the development of model organisms to study human diseases. It also incorporates the study of gene and genome evolution and the relationship between organisms and their environment. Comparative genomics can reveal genetic differences and similarities between organisms to provide insight into how those differences contribute to differences in phenotype, life cycle, or other attributes, and to ascertain the evolutionary history of those genetic differences.

Prokaryotic and Eukaryotic Genomes Display Common Structural and Functional Features and Important Differences

Since most prokaryotes have small genomes amenable to shotgun cloning and sequencing, many early genome projects have focused on prokaryotes, and more than 900 additional projects to sequence prokaryotic genomes are now under way. Many of the prokaryotic genomes already sequenced are from organisms that cause human diseases, such as cholera, tuberculosis, and leprosy. Traditionally, the bacterial genome has been thought of as relatively small (less than 5 Mb) and contained within a single circular DNA molecule. *E. coli,* used as the prototypical bacterial model organism in genetics, has a genome with these characteristics. However, the flood of genomic information now available has challenged the validity of this viewpoint for bacteria in general. Although most prokaryotic genomes are small, their sizes vary across a surprisingly wide range. In fact, there is some overlap in size between larger bacterial genomes (30 Mb in *Bacillus megaterium*) and smaller eukaryotic genomes (12.1 Mb in yeast). Gene number in bacterial genomes also demonstrates a wide range, from less than 500 to more than 5000 genes, a ten-fold difference.

In addition, although many bacteria have a single, circular chromosome, there is substantial variation in chromosome organization and number among bacterial species. An increasing number of genomes composed of linear DNA molecules are being identified, including the genome of *Borrelia burgdorferi,* the organism that causes Lyme disease. Sequencing of the *Vibrio cholerae* genome (the organism responsible for cholera) revealed the presence of two circular chromosomes.

Other bacteria that have genomes with two or more chromosomes include *Rhizobium radiobacter* (formerly *Agrobacterium tumefaciens*), *Deinococcus radiodurans,* and *Rhodobacter sphaeroides.* The finding that some bacterial species have multiple chromosomes raises questions both about how replication and segregation of their chromosomes are coordinated during cell division and about what undiscovered mechanisms of gene regulation may exist in bacteria. The answers may provide clues about the evolution of multichromosome eukaryotic genomes.

We can make two generalizations about the organization of protein-coding genes in bacteria. First, gene density is very high, averaging about one gene per kilobase of DNA. For example, the genome of *E. coli* strain K12, which was sequenced in 1997 as the second prokaryotic genome to be sequenced, is 4.6 Mb in size, and it contains 4289 protein-coding genes in its single, circular chromosome. This close packing of genes in prokaryotic genomes means that a very high proportion of the DNA (approximately 85 to 90 percent) serves as coding DNA. Typically, only a small amount of a bacterial genome is noncoding DNA, often in the form of regulatory sequences or of transposable elements that can move from one place to another in the genome.

The second generalization we can make is that bacterial genomes contain operons. Recall from an earlier chapter (see Chapter 15) that operons contain multiple genes functioning as a transcriptional unit whose protein products are part of a common biochemical pathway). In *E. coli,* 27 percent of all genes are contained in operons (almost 600 operons).

The basic features of eukaryotic genomes are similar in different species, although genome size in eukaryotes is highly variable (**Table 18.2**). Genome sizes range from about 10 Mb in fungi to over 100,000 Mb in some flowering plants (a ten thousand-fold range); the number of chromosomes per genome ranges from two to the hundreds (about a hundred-fold range), but the number of genes varies much less dramatically than either genome size or chromosome number.

Eukaryotic genomes have several features not found in prokaryotes:

- **Gene density.** In prokaryotes, gene density is close to 1 gene per kilobase. In eukaryotic genomes, there is a wide range of gene density. In yeast, there is about 1 gene/2 kb, in *Drosophila*, about 1 gene/13 kb, and in humans, gene density varies greatly from chromosome to chromosome. Human chromosome 22 has about 1 gene/64 kb, while chromosome 13 has 1 gene/155 kb of DNA.
- **Introns.** Most eukaryotic genes contain introns. There is wide variation among genomes in the number of introns they contain and also wide variation from gene to gene. The entire yeast genome has only 239 introns, whereas just a single gene in the human genome can contain more than 100 introns. Regarding intron

TABLE 18.2 Comparison of Selected Genomes

Organism (Scientific Name)	Approximate Size of Genome (in million [megabase, Mb] or billion [gigabase, Gb] bases) (Date Completed)	Number of Genes	Approximate Percentage of Genes Shared with Humans
Bacterium (*Escherichia coli*)	4.6 Mb (1997)	4403	not determined
Chicken (*Gallus gallus*)	1 Gb (2004)	~20,000–23,000	60%
Dog (*Canis familiaris*)	2.5 Gb (2003)	~18,400	75%
Chimpanzee (*Pan troglodytes*)	~3 Gb (2005)	~20,000–24,000	98%
Fruit fly (*Drosophila melanogaster*)	165 Mb (2000)	~13,600	50%
Human (*Homo sapiens*)	3.1 Gb (2004)	~20,000	100%
Mouse (*Mus musculus*)	~2.5 Gb (2002)	~30,000	80%
Pig (*Sus scrofa*)	~3 Gb (2012)	21,640	84%
Rat (*Rattus norvegicus*)	~2.75 Gb (2004)	~22,000	80%
Rhesus macaque (*Macaca mulatta*)	2.87 Gb (2007)	~20,000	93%
Rice (*Oryza sativa*)	389 Mb (2005)	~41,000	not determined
Roundworm (*Caenorhabditis elegans*)	97 Mb (1998)	19,099	40%
Sea urchin (*Strongylocentrotus purpuratus*)	814 Mb (2006)	~23,500	60%
Thale cress (plant) (*Arabidopsis thaliana*)	140 Mb (2000)	~27,500	not determined
Yeast (*Saccharomyces cerevisiae*)	12 Mb (1996)	~5700	30%

Adapted from Palladino, M. A. *Understanding the Human Genome Project,* 2nd ed. Benjamin Cummings, 2006.
Note: Billion bp (gigabase, Gb).

size, generally the size in eukaryotes is correlated with genome size. Smaller genomes have smaller average introns, and larger genomes have larger average intron sizes. But there are exceptions. For example, the genome of the pufferfish (*Fugu rubripes*) has relatively few introns.

- **Repetitive sequences.** The presence of introns and the existence of repetitive sequences are two major reasons for the wide range of genome sizes in eukaryotes. In some plants, such as maize, repetitive sequences are the dominant feature of the genome. The maize genome has about 2500 Mb of DNA, and more than two-thirds of that genome is composed of repetitive DNA. In the human, as discussed previously, about half of the genome is repetitive DNA.

ESSENTIAL POINT

Genomic analysis of prokaryotes and eukaryotes has revealed similarities and important fundamental differences in genome size, gene number, and genome organization. ■

Comparative Genomics Provides Novel Information about the Genomes of Model Organisms and the Human Genome

As mentioned earlier, the Human Genome Project sequenced genomes from a number of model nonhuman organisms too, including *E. coli, Arabidopsis thaliana, Saccharomyces cerevisiae, Drosophila melanogaster,* the nematode roundworm *Caenorhabditis elegans,* and the mouse *Mus musculus.* Complete genome sequences of such organisms have been invaluable for comparative genomics studies of gene function in these organisms and in humans. As shown in Table 18.2, the number of genes humans share with other species is very high, ranging from about 30 percent of the genes in yeast to ~80 percent in mice and ~98 percent in chimpanzees. The human genome even contains around 100 genes that are also present in many bacteria.

Comparative genomics has shown us that many mutated genes involved in human disease are also present in model organisms. For instance, approximately 60 percent of genes mutated in nearly 300 human diseases are also found in *Drosophila.* These include genes involved in prostate, colon, and pancreatic cancers; cardiovascular disease; cystic fibrosis; and several other conditions. Here we consider how comparative genomics studies of several model organisms (sea urchins, dogs, chimpanzees, and Rhesus monkeys) and the Neanderthal genome have revealed interesting elements of the human genome.

The Sea Urchin Genome

In 2006, researchers from the Sea Urchin Genome Sequencing Consortium completed the 814 million bp genome of the sea urchin *Strongylocentrotus purpuratus.* Sea urchins are shallow-water marine invertebrates that have served

as important model organisms, particularly for developmental biologists. One reason is that the sea urchin is a nonchordate deuterostome, and humans, with their spinal cord, are chordate deuterostomes. Fossil records indicate that sea urchins appeared during the Early Cambrian period, around 520 mya.

A combination of whole-genome shotgun sequencing and map-based cloning in BACs was used to complete the genome. Sea urchins have an estimated 23,500 genes, including representative genes for just about all major vertebrate gene families. Sequence alignment and homology searches demonstrate that the sea urchin contains many genes with important functions in humans, yet interestingly, important genes in flies and worms, such as certain cytochrome P-450 genes that play a role in the breakdown of toxic compounds, are missing from sea urchins. The sea urchin genome also has an abundance (~25 to 30 percent) of **pseudogenes**—nonfunctional relatives of protein-coding genes (we meet pseudogenes again in the next subsection). Sea urchins have a smaller average intron size than humans, supporting the general trend revealed by comparative genomics that intron size is correlated with overall genome size.

Urchins have nearly 1000 genes for sensing light and odor, indicative of great sensory abilities. In this respect, their genome is more typical of vertebrates than invertebrates. A number of orthologs of human genes involved in hearing and balance are present in the sea urchin, as are many human-disease-associated orthologs, including protein kinases, GTPases, transcription factors, innate immunity, transporters, and low-density lipoprotein receptors. Sea urchins and humans share approximately 7000 orthologs.

The Dog Genome

In 2005 the genome for "man's best friend" was completed, and it revealed that we share about 75 percent of our genes with dogs (*Canis familiaris*), providing a useful model with which to study our own genome. Dogs have a genome that is similar in size to the human genome: about 2.5 billion base pairs with an estimated 18,400 genes. The dog offers several advantages for studying heritable human diseases. Dogs share many genetic disorders with humans, including over 400 single-gene disorders, sex-chromosome aneuploidies, multifactorial diseases (such as epilepsy), behavioral conditions (such as obsessive-compulsive disorder), and genetic predispositions to cancer, blindness, heart disease, and deafness.

The molecular causes of at least 60 percent of inherited diseases in dogs, such as point mutations and deletions, are similar or identical to those found in humans. In addition, at least 50 percent of the genetic diseases in dogs are breed-specific, so that the mutant allele segregates in relatively homogeneous genetic backgrounds. Dog breeds resemble isolated human populations in having a small number of founders and a long period of relative genetic isolation. These properties make individual dog breeds useful as models of human genetic disorders.

Dog breeders are now using genetic tests to screen dogs for inherited disease conditions, for coat color in Labrador retrievers and poodles, and for fur length in Mastiffs. Undoubtedly, we can expect many more genetic tests for dogs in the near future, including DNA analysis for size, type of tail, speed, sense of smell, and other traits deemed important by breeders and owners.

The Chimpanzee Genome

Although the chimpanzee (*Pan troglodytes*) genome was not part of the HGP, its nucleotide sequence was completed in 2004. Overall, the chimp and human genome sequences differ by less than 2 percent, and 98 percent of the genes are the same. Comparisons between these genomes offer some interesting insights into what makes some primates humans and others chimpanzees.

The speciation events that separated humans and chimpanzees occurred less than 6.3 million years ago (mya). Genomic analysis indicates that these species initially diverged but then exchanged genes again before separating completely. Their separate evolution after this point is exhibited in such differences as that seen between the sequence of chimpanzee chromosome 22 and its human ortholog, chromosome 21 (chimps have 48 chromosomes and humans have 46, so the numbering is different). These chromosomes have accumulated nucleotide substitutions that total 1.44 percent of the sequence.

The most surprising difference is the discovery of 68,000 nucleotide insertions or deletions, collectively called **indels,** in the chimp and human chromosomes, a frequency of 1 indel every 470 bases. Many of these are *Alu* insertions in human chromosome 21. Although the overall difference in the nucleotide sequence is small, there are significant differences in the encoded genes. Only 17 percent of the genes analyzed encode identical proteins in both chromosomes; the other 83 percent encode genes with one or more amino acid differences.

Differences in the time and place of gene expression also play a major role in differentiating the two primates. Using DNA microarrays (discussed in Section 18.9), researchers compared expression patterns of 202 genes in human and chimp cells from brain and liver. They found more species-specific differences in expression of brain genes than liver genes. To further examine these differences, Svante Pääbo and colleagues compared expression of 10,000 genes in human and chimpanzee brains and found that 10 percent of genes examined differ in expression in

one or more regions of the brain. More importantly, these differences are associated with genes in regions of the human genome that have been duplicated subsequent to the divergence of chimps and humans. This finding indicates that genome evolution, speciation, and gene expression are interconnected. Further work on these segmental duplications and the genes they contain may identify genes that help make us human.

The Rhesus Monkey Genome

The Rhesus macaque monkey (*Macaca mulatta*), another primate, has served as one of the most important model organisms in biomedical research. Macaques have played central roles in our understanding of cardiovascular disease, aging, diabetes, cancer, depression, osteoporosis, and many other aspects of human health. They have been essential for research on AIDS vaccines and for the development of polio vaccines. The macaque's genome is the first monkey genome to have been sequenced. A main reason geneticists are so excited about the completion of this sequencing project is that macaques provide a more distant evolutionary window that is ideally suited for comparing and analyzing human and chimpanzee genomes. As we discussed in the preceding section, humans and chimpanzees shared a common ancestor approximately 6 mya. But macaques split from the ape lineage that led to chimpanzees and humans about 25 mya. The macaque and human genome have thus diverged farther from one another, as evidenced by the ~93 percent sequence identity between humans and macaques compared to the ~98 percent sequence identity shared by humans and chimpanzees.

The macaque genome was published in 2007, and it was no surprise to learn that it consists of 2.87 billion bp (similar to the size of the human genome) contained in 22 chromosomes (20 autosomes, an X, and a Y) with ~20,000 protein-coding genes. Although comparative analyses of this genome are ongoing, a number of interesting features have been revealed so far. As in humans, about 50 percent of the genome consists of repeat elements (transposons, LINEs, SINEs). Gene duplications and gene families are abundant, including cancer gene families found in humans.

A number of interesting surprises have also been observed. For instance, recall from earlier in the text (see Chapter 4) and elsewhere our discussion about the genetic disorder phenylketonuria (PKU), an autosomal recessive inherited condition in which individuals cannot metabolize the amino acid phenylalanine due to mutation of the phenylalanine hydroxylase (*PAH*) gene. The histidine substitution encoded by a mutation in the *PAH* gene of humans with PKU appears as the wild-type amino acid in the protein from healthy macaques. Further analysis of the macaque genome and comparison to the human and chimpanzee genome will be invaluable for geneticists studying genetic variations that played a role in primate evolution.

The Neanderthal Genome and Modern Humans

In early 2009, a team of scientists led by Svante Pääbo at the Max Planck Institute for Evolutionary Anthropology in Germany and 454 Life Sciences reported completion of a rough draft of the Neanderthal (*Homo neanderthalensis*) genome encompassing more than 3 billion bp of Neanderthal DNA and about two-thirds of the genome. Previously, in 1997, Pääbo's lab sequenced portions of Neanderthal mitochondrial DNA from a fossil. In late 2006, Pääbo's group along with a number of scientists in the United States reported the first sequence of ~65,000 bp of nuclear DNA isolated from Neanderthal bone samples. Bones from three females who lived in Vindija Cave in Croatia about 38,000 to 44,000 years ago were used to produce the draft sequence of the Neanderthal nuclear genome.

Because Neanderthals are members of the human family, and closer relatives to humans than chimpanzees, the Neanderthal genome is expected to provide an unprecedented opportunity to use comparative genomics to advance our understanding of evolutionary relationships between modern humans and our predecessors. In particular, scientists are interested in identifying areas in the genome where humans have undergone rapid evolution since splitting (diverging) from Neanderthals. Much of this analysis involves a comparative genomics approach to compare the Neanderthal genome to the human and chimpanzee genomes.

The human and Neanderthal genomes are 99 percent identical. Comparative genomics has identified 78 protein-coding sequences in humans that seem to have arisen since the divergence from Neanderthals and that may have helped modern humans adapt. Some of these sequences are involved in cognitive development and sperm motility. Of the many genes shared by these species, *FOXP2* is a gene that has been linked to speech and language ability. There are many genes that influence speech, so this finding does not mean that Neanderthals spoke as we do. But because Neanderthals had the same modern human *FOXP2* gene scientists have speculated that Neanderthals possessed linguistic abilities.

The realization that modern humans and Neanderthals lived in overlapping ranges as recently as 30,000 years ago has led to speculation about the interactions between modern humans and Neanderthals. Genome studies suggest that interbreeding took place between Neanderthals and modern humans an estimated 37,000 to 80,000 years ago in the eastern Mediterranean. In fact, the genome

of non-African *H. sapiens* contains approximately 1–4 percent of sequence inherited from Neanderthals. Recent work by Pääbo's lab on a 45,000-year-old leg bone from Siberia has produced the oldest genome sequence for *Homo sapiens* to date. About 2 percent of this genome was derived from Neanderthals. These exciting studies, previously thought to be impossible, are having ramifications in many areas of human evolution, and it will be interesting indeed to follow the progress of this work.

ESSENTIAL POINT

Studies in comparative genomics are revealing fascinating similarities and variations in genomes from different organisms. ■

18.7 Comparative Genomics Is Useful for Studying the Evolution and Function of Multigene Families

Comparative genomics has also proven to be valuable for identifying members of **multigene families,** groups of genes that share similar but not identical DNA sequences through duplication and descent from a single ancestral gene. Their gene products frequently have similar functions, and the genes are often, but not always, found at a single chromosomal locus. A group of related multigene families is called a **superfamily.** Sequence data from genome projects are providing evidence that multigene families are present in many, if not all, genomes. One of the best-studied examples of gene family evolution is the **globin gene superfamily,** whose members encode very similar but not identical polypeptide chains with closely related functions (**Figure 18–9**). Other well-characterized gene superfamilies include the histone, tubulin, actin, and immunoglobulin (antibody) gene superfamilies.

Recall that paralogs, which we defined in Section 18.3, are homologous genes present in the same single organism, believed to have evolved by gene duplication. The globin genes that encode the polypeptides in hemoglobin molecules are a paralogous multigene superfamily that arose by duplication and dispersal to occupy different chromosomal sites.

In this family, an ancestral gene encoding an oxygen transport protein was duplicated about 800 mya, producing two sister genes, one of which evolved into the modern-day myoglobin gene. **Myoglobin** is an oxygen-carrying protein found in muscle. The other gene underwent further duplication and divergence about 500 mya and formed prototypes of the α-globin and β-globin genes. These genes encode proteins found in **hemoglobin,** the oxygen-carrying molecule in red blood cells. Additional duplications within these genes occurred within the last 200 million years. Events subsequent to each duplication dispersed these gene subfamilies to different chromosomes, and in the human genome, each now resides on a separate chromosome.

18.8 Metagenomics Applies Genomics Techniques to Environmental Samples

Metagenomics, also called **environmental genomics,** is the use of whole-genome shotgun approaches to sequence genomes from entire communities of microbes in environmental samples of water, air, and soil. Oceans, glaciers, deserts, and virtually every other environment on Earth are being sampled for metagenomics projects. Human

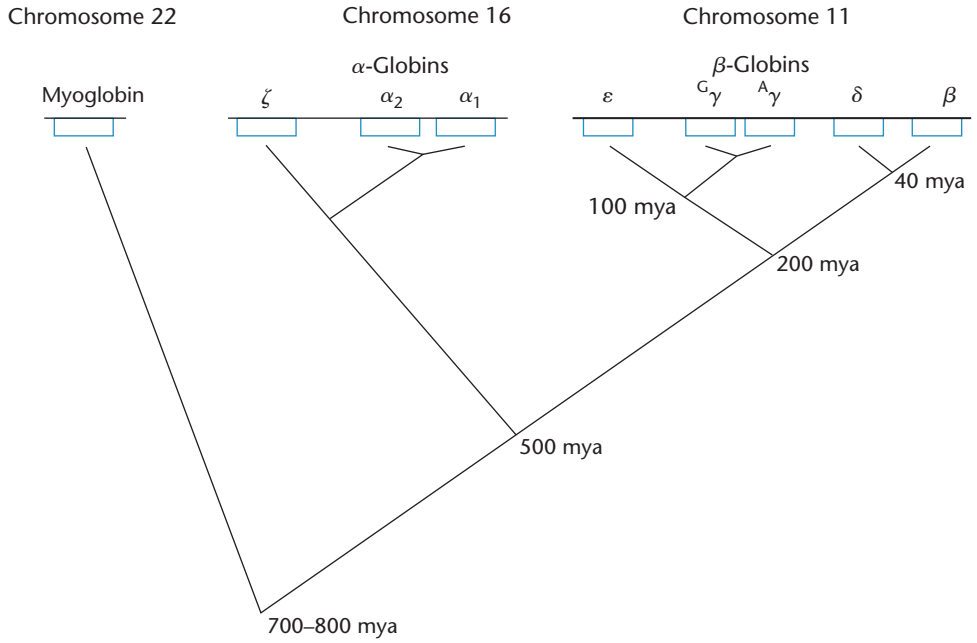


FIGURE 18–9 The evolutionary history of the globin gene superfamily. A duplication event in an ancestral gene gave rise to two lineages about 700 to 800 million years ago (mya). One line led to the myoglobin gene, which is located on chromosome 22 in humans; the other underwent a second duplication event about 500 mya, giving rise to the ancestors of the α-globin and β-globin gene subfamilies. Duplications beginning about 200 mya formed the β-globin gene subfamilies. In humans, the α-globin genes are located on chromosome 16, and the β-globin genes are on chromosome 11.

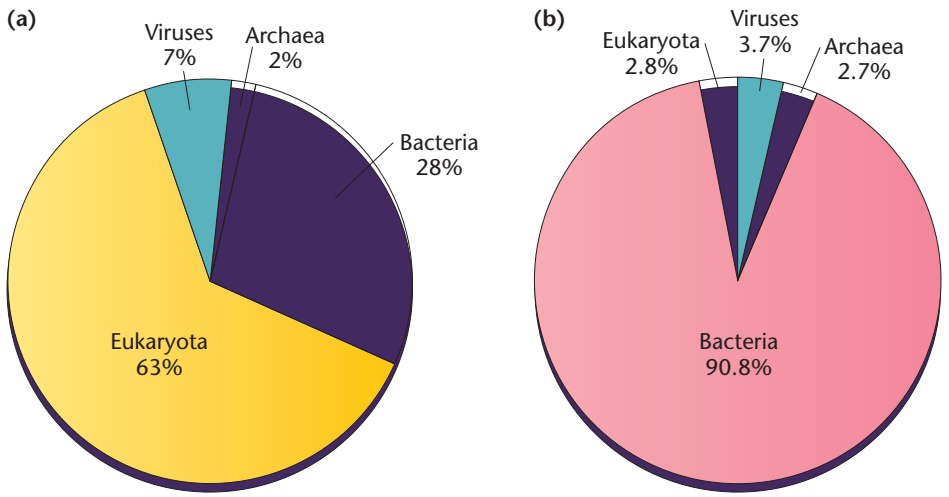


FIGURE 18–10 (a) Kingdom identifications for predicted proteins in NCBInr, NCBI Prokaryotic Genomes, the Institute for Genomics Research Gene Indices, and Ensembl databases. Notice that the publicly available databases of sequenced genomes and the predicted proteins they encode are dominated by eukaryotic sequences. (b) Kingdom identifications for novel predicted proteins in the Global Ocean Sampling (GOS) database. Bacterial sequences dominate this database, demonstrating the value of metagenomics for revealing new information about microbial genomes and microbial communities.

genome pioneer J. Craig Venter left Celera to form the J. Craig Venter Institute, and his group has played a central role in developing metagenomics as an emerging area of genomics research.

One of the institute's major initiatives has been a global expedition to sample marine and terrestrial microorganisms from around the world and to sequence their genomes. Through this project, called the *Sorcerer II* Global Ocean Sampling (GOS) Expedition, Venter and his researchers traveled the globe by yacht, in a sailing voyage described as a modern-day version of Charles Darwin's famous voyage on the *H.M.S. Beagle.*

A key benefit of metagenomics is its potential for teaching us more about millions of species of bacteria, of which only a few thousand have been well characterized. Many new viruses, particularly bacteriophages, are also identified through metagenomics studies of water samples. Metagenomics is providing important new information about genetic diversity in microbes that is key to understanding complex interactions between microbial communities and their environment, as well as allowing phylogenetic classification of newly identified microbes. Metagenomics also has great potential for identifying genes with novel functions, some of which may have valuable applications in medicine and biotechnology.

The general method used in metagenomics to sequence genomes for all microbes in a given environment involves isolating DNA directly from an environmental sample without requiring cultures of the microbes or viruses. Such an approach is necessary because often it is difficult to replicate the complex array of growth conditions the microbes need to survive in culture.

For the *Sorcerer II* GOS project, samples of water from different layers in the water column were passed through high-density filters of various sizes to capture the microbes. DNA was then isolated from the microbes and subjected to shotgun sequencing and genome assembly. High-throughput sequencers on board the yacht operated nearly around the clock.

By early 2007, the GOS database contained approximately 6 billion bp of DNA from more than 400 uncharacterized microbial species! These sequences included 7.7 million previously uncharacterized sequences, encoding more than 6 million different potential proteins. This is almost twice the total number of previously characterized proteins in all other known databases worldwide. **Figure 18–10(a)** shows the kingdom assignments for predicted protein sequences in publicly available databases worldwide, such as the NCBI-nonredundant protein database (NCBInr), which accesses GenBank, Ensembl, and other well-known databases. Eukaryotic sequences comprise the majority (63 percent) of predicted proteins in these databases. Reviewing the kingdom assignments of approximately 6 million predicted proteins in the Global Ocean Sampling (GOS) dataset shows that, in contrast, the largest majority (90.8 percent) of sequences in this database are from the bacterial kingdom [**Figure 18–10(b)**].

The GOS Expedition also examined protein families corresponding to the predicted proteins encoded by the genome sequences in the GOS database: 17,067 families were medium (between 20 and 200 proteins) and large-sized (>200 proteins) clusters. These data demonstrate the value of the GOS Expedition and of metagenomics for identifying novel microbial genes and potential proteins.

In Section 18.5 you learned about the Human Microbiome Project. This project represents an example of a metagenomics project in that it is intended to sequence the genomes of microbes and viruses present in and on humans as the "environment" being sampled. Many other high-profile metagenomics applications have emerged recently, including the use of metagenomics to identify viruses and fungi thought to be involved in colony collapse disorder. One such malady has resulted in the loss of 50–90 percent

of the honey bee population in beekeeping operations throughout the United States.

ESSENTIAL POINT

Metagenomics, or environmental genomics, sequences genomes of organisms in environmental samples, often identifying new sequences that encode proteins with novel functions. ■

18.9 Transcriptome Analysis Reveals Profiles of Expressed Genes in Cells and Tissues

Once any genome has been sequenced and annotated, a formidable challenge still remains: that of understanding genome function by analyzing the genes it contains and the ways the genes expressed by the genome are regulated. **Transcriptome analysis,** also called **transcriptomics** or **global analysis of gene expression,** studies the expression of genes by a genome both qualitatively—by identifying which genes are expressed and which genes are not expressed—and quantitatively—by measuring varying levels of expression for different genes.

Even though in theory all cells of an organism possess the same gene in any cell or tissue type, certain genes will be highly expressed, others expressed at low levels, and some not expressed at all. Transcriptome analysis reveals gene-expression profiles that for the same genome may vary from cell to cell or from tissue type to tissue type. Identifying genes expressed by a genome is essential for understanding how the genome functions. Transcriptome analysis provides insights into (1) normal patterns of gene expression that are important for understanding how a cell or tissue type differentiates during development, (2) how gene expression dictates and controls the physiology of differentiated cells, and (3) mechanisms of disease development that result from or cause gene-expression changes in cells. Later in the text (see Chapter 19), we will consider why gene-expression analysis is gradually becoming an important diagnostic tool in certain areas of medicine. For example, examining gene-expression profiles in a cancerous tumor can help diagnose tumor type, determine the likelihood of tumor metastasis (spreading), and develop the most effective treatment strategy.

Microarray Analysis

A number of different techniques can be used for transcriptome analysis. PCR-based methods are useful because of their ability to detect genes that are expressed at low levels. For many years **DNA microarray analysis** was widely used because it enables researchers to analyze all of a sample's expressed genes simultaneously. Although as DNA sequencing technologies have developed and, most recently, techniques for **RNA sequencing** (**RNA-seq**) have developed including *in situ* RNA sequencing, it is expected that microarrays will become antiquated relatively soon.

Most microarrays, also known as **gene chips,** consist of a glass microscope slide onto which single-stranded DNA molecules are attached, or "spotted," using a computer-controlled high-speed robotic arm called an arrayer. Arrayers are fitted with a number of tiny pins. Each pin is immersed in a small amount of solution containing millions of copies of a different single-stranded DNA molecule. For example, many microarrays are made with single-stranded sequences of complementary DNA (cDNA) or expressed sequenced tags (ESTs)—short fragments of cloned DNA from expressed genes. The arrayer fixes the DNA onto the slide at specific locations (points, or spots) that are recorded by a computer. A single microarray can have over 20,000 different spots of DNA (and over 1 million for exon-specific microarrays), each containing a unique sequence that serves as a probe for a different gene.

Probes for entire genomes are available on microarrays, including the human genome. As you will learn later in the text (see Chapter 19), researchers are also using microarrays to compare patterns of gene expression in tissues in response to different conditions, to compare gene-expression patterns in normal and diseased tissues, and to identify pathogens. One approach to using a microarray for transcriptome analysis is shown in **Figure 18–11**.

Microarrays have dramatically changed the way gene-expression patterns are analyzed. As discussed earlier in the text (see Chapter 17), Northern blot analysis was one of the earliest methods used for analyzing gene expression. Then PCR techniques proved to be rapid and more sensitive approaches. The biggest advantage of microarrays is that they enable thousands of genes to be studied simultaneously. As a result, however, they can generate an overwhelming amount of gene-expression data. Over 1 million gene-expression datasets are now available in publicly accessible databases. Most of these datasets have been generated in the past decade largely through microarray analysis. In addition, even when properly controlled, microarrays often yield variable results. For example, one experiment under certain conditions may not always yield similar patterns of gene expression as another identical experiment. Some of these differences can be due to real differences in gene expression, but others can be the result of variability in chip preparation, cDNA synthesis, probe hybridization, or washing conditions, all of which must be carefully controlled to limit such variability. Commercially available microarrays can reduce

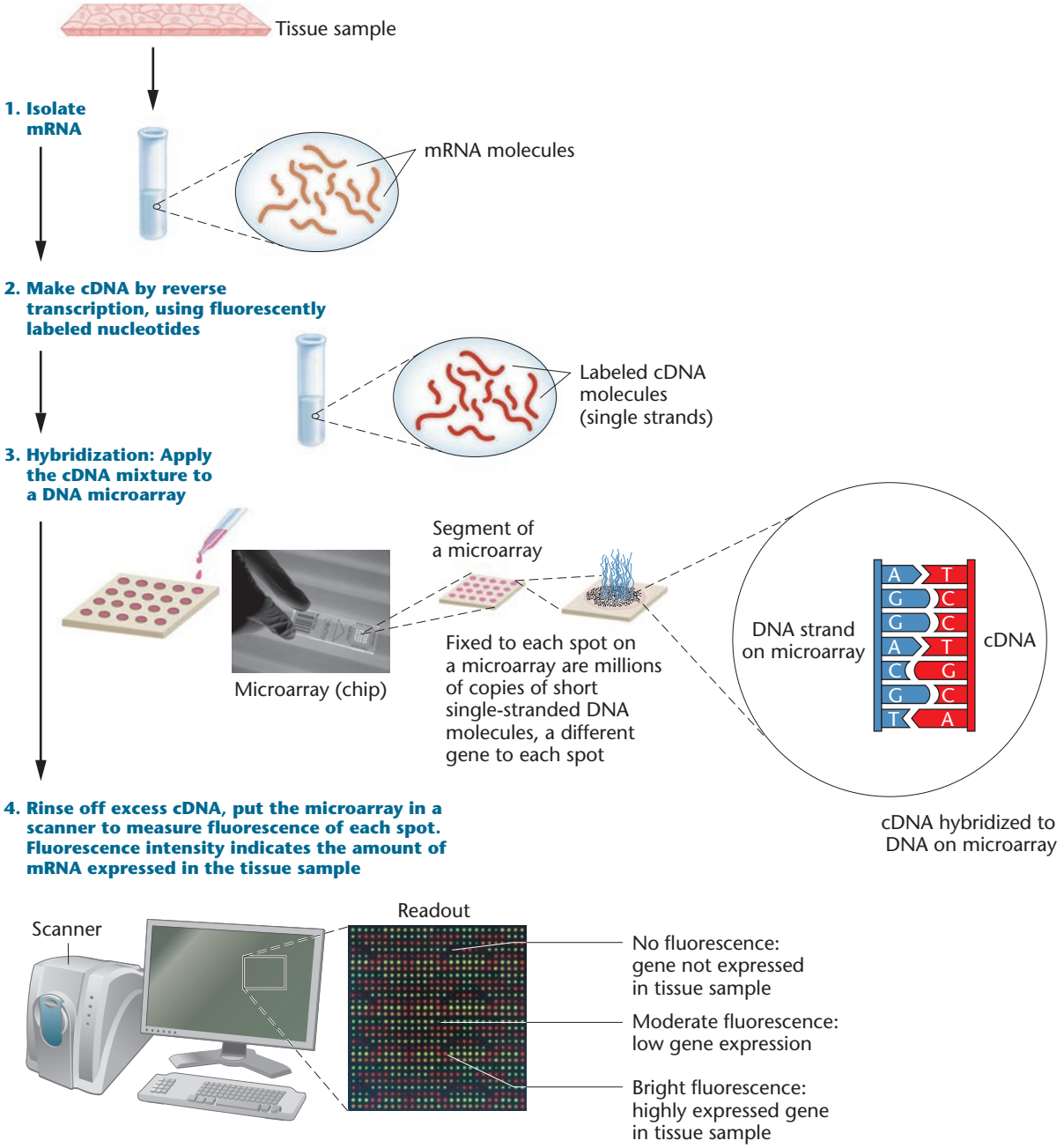


FIGURE 18–11 Microarray analysis for analyzing gene-expression patterns in a tissue.

the variability that can result when individual researchers make their own arrays. As mentioned previously, you should also be aware that new methods for directly sequencing RNA (RNA-Seq, also called whole-transcriptome shotgun sequencing) will soon render microarrays obsolete.

Now that we have considered genomes and transcriptomes, we turn our attention to the ultimate end products of most genes, the proteins encoded by a genome.

ESSENTIAL POINT

DNA microarrays or gene chips have been valuable for transcriptome analysis by studying expression patterns for thousands of genes simultaneously. ■

18.10 Proteomics Identifies and Analyzes the Protein Composition of Cells

As more genomes have been sequenced and studied, biologists have focused increasingly on understanding the complex structures and functions of the proteins the genomes encode. This interest is not surprising given that in most of the genomes sequenced to date, many newly discovered genes and their putative proteins have no known function. Keep in mind, in the ensuing discussion, that although every

cell in the body contains an equivalent set of genes, not all cells express the same genes and proteins. **Proteome** is a term that represents the complete set of proteins encoded by a genome, but it is also often used to mean the entire complement of proteins in a cell. This definition would then include proteins that a cell acquired from another cell type.

Proteomics—the complete identification, characterization, and quantitative analysis of the proteome of a cell, tissue, or organism—can be used to reconcile differences between the number of genes in a genome and the number of different proteins produced. But equally important, proteomics also provides information about a protein's structure and function; posttranslational modifications; protein–protein, protein–nucleic acid, and protein–metabolite interactions; cellular localization of proteins; protein stability and aspects of translational and posttranslational levels of gene-expression regulation; and relationships (shared domains, evolutionary history) to other proteins.

Proteomics is also of clinical interest because it allows comparison of proteins in normal and diseased tissues, which can lead to the identification of proteins as biomarkers for disease conditions. Proteomic analysis of mitochondrial proteins during aging, proteomic maps of atherosclerotic plaques from human coronary arteries, and protein profiles in saliva as a way to detect and diagnose diseases are examples of such work.

Reconciling the Number of Genes and the Number of Proteins Expressed by a Cell or Tissue

Recall the one-gene:one-polypeptide hypothesis of George Beadle and Edward Tatum (see Chapter 13). Genomics has revealed that the link between gene and gene product is often much more complex. Genes can have multiple transcription start sites that produce several different types of transcripts. Alternative splicing and editing of pre-mRNA molecules can generate dozens of different proteins from a single gene. Remember the current estimate that over 50 percent of human genes produce more than one protein by alternative splicing. As a result, proteomes are substantially larger than genomes. For instance, the ~20,000 genes in the human genome encode ~100,000 proteins, although some estimates suggest that the human proteome may be as large as 150,000–200,000 proteins.

Proteomes undergo dynamic changes that are coordinated in part by regulation of gene-expression patterns—the transcriptome. However, a number of other factors affect the proteome profile of a cell, further complicating the analysis of protein function. For instance, many proteins are modified by co-translational or posttranslational events, such as cleavage of signal sequences that target a protein for an organelle pathway, propeptides, or initiator methionine residues; by linkage to carbohydrates and lipids; or by the addition of chemical groups through methylation, acetylation, and phosphorylation and other modifications. Over a hundred different mechanisms of posttranslational modification are known. In addition, many proteins work via elaborate protein–protein interactions or as part of a large molecular complex.

Well before a draft sequence of the human genome was available, scientists were already discussing the possibility of a "Human Proteome Project." One reason such a project never came to pass is that there is no single human proteome: different tissues produce different sets of proteins. But the idea of such a project led to the **Protein Structure Initiative (PSI)** by the National Institute of General Medical Sciences (NIGMS), a division of the National Institutes of Health, involving over a dozen research centers. PSI is a multiphase project designed to analyze the three-dimensional structures of more than 4000 protein families. Proteins with interesting potential therapeutic properties are a top priority for the PSI, and to date the structures of over 6000 proteins have been determined. Developing computation protein structural prediction methods, solving unique protein structures, disseminating PSI information, and focusing on the biological relevance of the work are major goals. There also are a number of other ongoing projects dedicated to identifying proteome profiles that correlate with diseases such as cancer and diabetes.

Proteomics Technologies: Two-Dimensional Gel Electrophoresis for Separating Proteins

With proteomics technologies, scientists have the ability to study thousands of proteins simultaneously, generating enormous amounts of data quickly and dramatically changing ways of analyzing the protein content of a cell.

The early history of proteomics dates back to 1975 and the development of **two-dimensional gel electrophoresis (2DGE)** as a technique for separating hundreds to thousands of proteins with high resolution (**Figure 18–12**). In this technique, proteins isolated from cells or tissues of interest are loaded onto a polyacrylamide tube gel and first separated by *isoelectric focusing*, which causes proteins to migrate according to their electrical charge in a pH gradient. Then in a second migration, perpendicular to the first, the proteins are separated by their molecular mass using **sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE).**

Proteins in the 2D gel are visualized by staining with Coomassie blue, silver stain, or other dyes that reveal the separated proteins as a series of spots in the gel (Figure 18–12). It is not uncommon for a 2D gel loaded with a complex mixture of proteins to show several thousand spots in the gel, as in Figure 18–12, which displays the

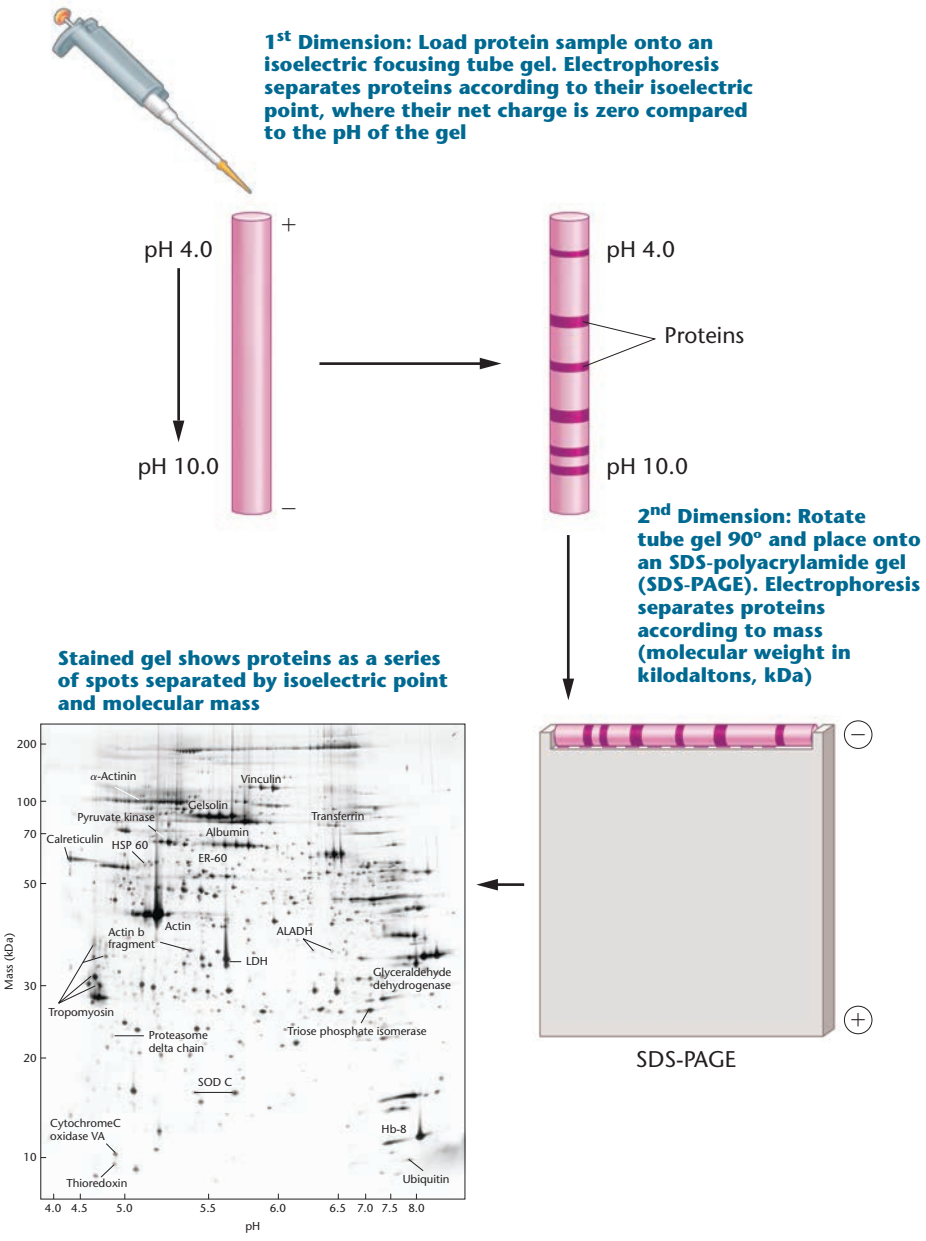


FIGURE 18–12 Two-dimensional gel electrophoresis (2DGE) is a useful method for separating proteins in a protein extract from cells or tissues that contains a complex mixture of proteins with different biochemical properties. The two-dimensional gel photo shows separations of human platelet proteins. Each spot represents a different polypeptide separated by molecular weight (*y*-axis) and isoelectric point, pH (*x*-axis). Known protein spots are labeled by name based on identification by comparison to a reference gel or by determination of protein sequence using mass spectrometry techniques. Notice that many spots on the gel are unlabeled, indicating proteins of unknown identity.

complex mixture of proteins in human platelets (thrombocytes). Particularly abundant protein spots in this gel have been labeled with the names of identified proteins. With thousands of different spots on the gel, how are the identities of the proteins ascertained?

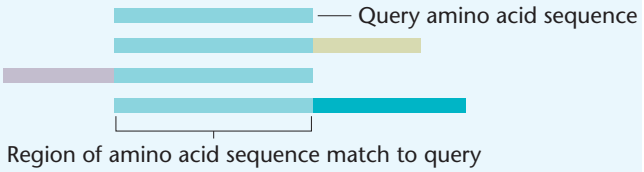
In some cases, 2D gel patterns from experimental samples can be compared to gels run with reference standards containing known proteins with well-characterized migration patterns. Many reference gels for different biological samples such as human plasma are available, and computer software programs can be used to align and compare the spots from different gels. In the early days of 2DGE, proteins were often identified by cutting spots out of a gel and sequencing the amino acids the spots contained. Only relatively small sequences of amino acids can typically be generated this way; rarely can an entire polypeptide be sequenced using this technique. BLAST and similar programs can be used to search protein databases containing amino acid sequences of known proteins. However, because of alternative splicing or posttranslational modifications, peptide sequences may not always match easily with the final product, and the identity of the protein may have to be confirmed by another approach. As you will learn in the next section, proteomics has incorporated other techniques to aid in protein identification, and one of these techniques is mass spectrometry.

Proteomics Technologies: Mass Spectrometry for Protein Identification

As important as 2DGE has been for protein analysis, **mass spectrometry** (**MS**) has been instrumental to the development of proteomics. Mass spectrometry techniques analyze ionized samples in gaseous form and measure the **mass-to-charge** (***m/z***) **ratio** of the different ions in a sample. Proteins analyzed by mass spectra generate *m/z* spectra that can be correlated with an *m/z* database containing known protein sequences to discover the protein's identity. Certain MS applications can provide peptide sequences directly from spectra. Some of the most valuable proteomics applications of this technology are to identify an unknown protein or proteins in a complex mix of proteins, to sequence peptides, to identify posttranslational modifications of proteins, and to characterize multiprotein complexes.

NOW SOLVE THIS

18–2 Annotation of a proteome attempts to relate each protein to a function in time and space. Traditionally, protein annotation depended on an amino acid sequence comparison between a query protein and a protein with known function. If the two proteins shared a considerable portion of their sequence, the query would be assumed to share the function of the annotated protein. Following is a representation of this method of protein annotation involving a query sequence and three different human proteins. Note that the query sequence aligns to common domains within the three other proteins. What argument might you present to suggest that the function of the query is not related to the function of the other three proteins?



HINT: *This problem asks you to think about sequence similarities between four proteins and predict functional relationships. The key to its solution is to remember that although protein domains may have related functions, proteins can contain several different interacting domains that determine protein function.*

For MS analysis, proteins are first extracted from cells or tissues of interest and separated by 2DGE, after which MS is used to identify the proteins in the different spots. **Figure 18–13** shows an example in which two different sets of cells grown in culture are analyzed for protein differences. Just about any source providing a sufficient number of cells can be used: blood, whole tissues, and organs; tumor samples; microbes; and many other substances. Many proteins involved in cancer have been identified by the use of MS to compare protein profiles in normal tissue and tumor samples.

Protein spots are cut out of the 2D gel, and proteins are purified out of each gel spot. Computer-automated high-throughput instruments are available that can pick all of the spots out of a 2D gel. Isolated proteins are then enzymatically digested with a protease (a protein-digesting enzyme) such as trypsin to create a series of peptides. This proteolysis produces a complex mixture of peptides determined by the cleavage sites for the protease in the original protein. Each type of protein produces a characteristic set of peptide fragments, and these are identified by MS (**Figure 18–14** on p. 389).

Databases of *m/z* spectra for different peptides can be analyzed to look for matches between *m/z* spectra of unknown samples and those of known proteins. One limitation of this approach is database quality. An unknown protein from a 2D gel can only be identified by MS if proteomics databases have a MS spectrum for that protein. But as is occurring with genomics databases, proteomics databases with thousands of well-characterized proteins from different organisms are rapidly developing.

As we mentioned when discussing genomics, high-throughput 2DGE instruments and mass spectrometers can process thousands of samples in a single day. Instruments with faster sample-processing times and increased sensitivity are under development. These instruments may soon make "shotgun proteomics" a viable approach for characterizing entire proteomes.

In late 2014, two different research teams reported results from mass spectrometry analysis of the human proteome that accounted for approximately 84 percent and 92 percent of the proteins encoded by the human genome. These studies have created proteome catalogs that will be available for researchers around the world.

Protein microarrays are also becoming valuable tools for proteomics research. These are designed around the same basic concept as microarrays (gene chips) and are often constructed with antibodies that specifically recognize and bind to different proteins. These microarrays are used, among other applications, for examining protein–protein interactions, for detecting protein markers for disease diagnosis, and for biosensors designed to detect pathogenic microbes and potentially infectious bioweapons.

ESSENTIAL POINT

Proteomics methods such as mass spectometry are valuable for analyzing proteomes—the protein content of a cell. ■

NOW SOLVE THIS

18–3 Because of its accessibility and biological significance, the proteome of human plasma has been intensively studied and used to provide biomarkers for such conditions as myocardial infarction (troponin) and congestive heart failure (B-type natriuretic peptide). Polanski and Anderson (Polanski, M., and Anderson, N. L., *Biomarker Insights,* 2: 1–48, 2006) have compiled a list of 1261 proteins, some occurring in plasma, that appear to be differentially expressed in human cancers. Of these 1261 proteins, only 9 have been recognized by the FDA as tumor-associated proteins. First, what advantage should there be in using plasma as a diagnostic screen for cancer? Second, what criteria should be used to validate that a cancerous state can be assessed through the plasma proteome?

HINT: *This problem asks you to consider criteria that are valuable for using plasma proteomics as a diagnostic screen for cancer. The key to its solution is to consider proteomics data that you would want to evaluate to determine whether a particular protein is involved in cancer.*

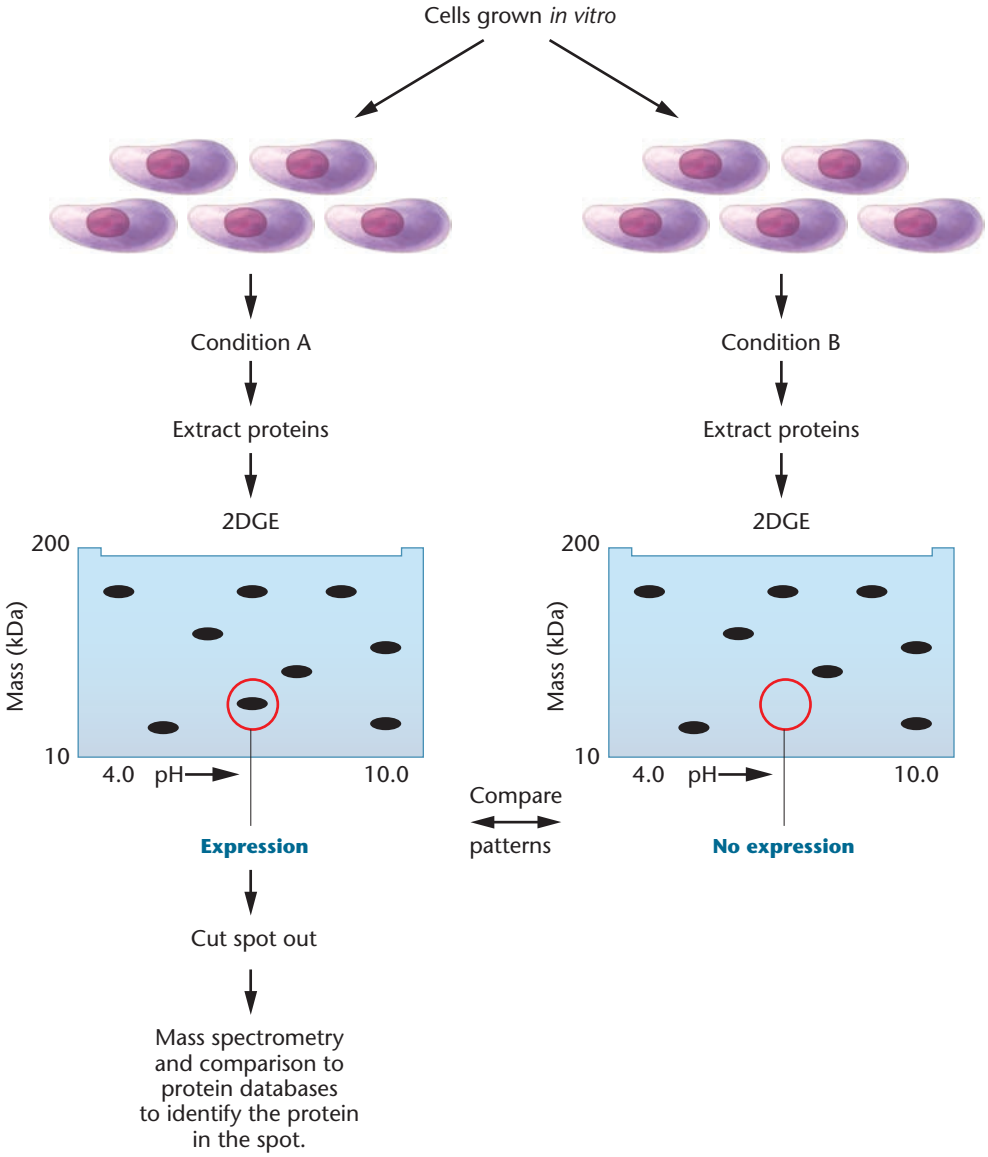
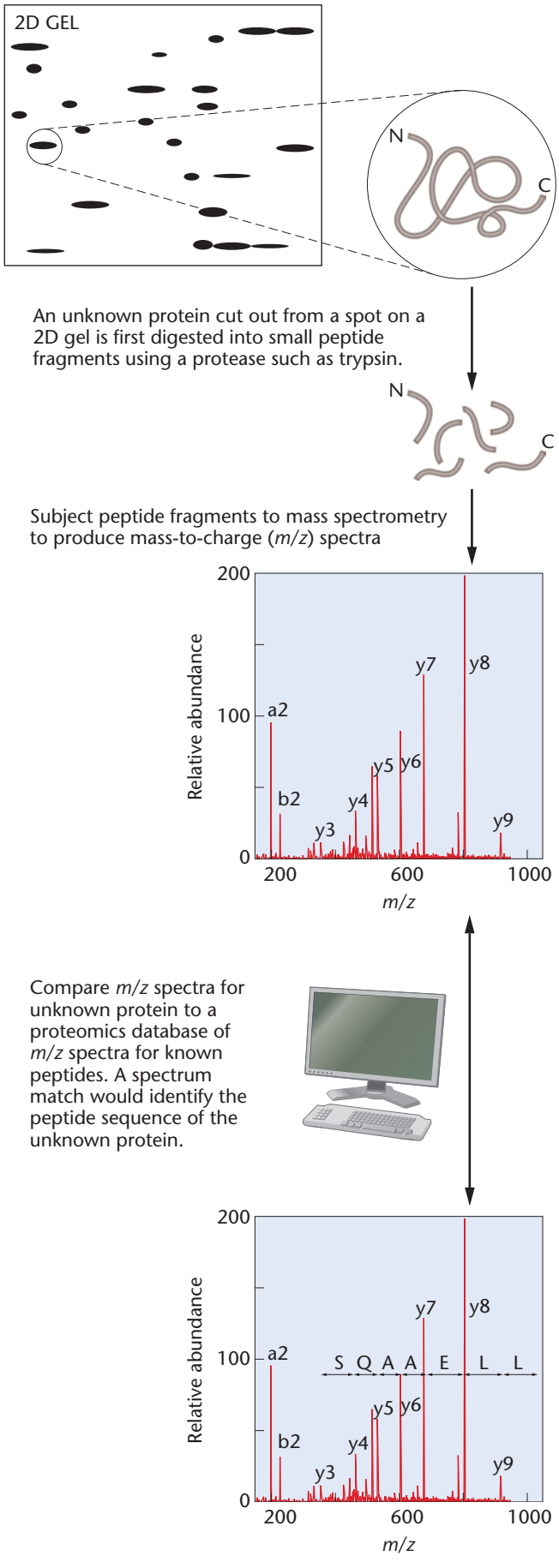


FIGURE 18–13 In a typical proteomic analysis, cells are exposed to different conditions (such as different growth conditions, drugs, or hormones). Then proteins are extracted from these cells and separated by 2DGE, and the resulting patterns of spots are compared for evidence of differential protein expression. Spots of interest are cut out from the gel, digested into peptide fragments, and analyzed by mass spectrometry to identify the protein they contain.

18.11 Systems Biology Is an Integrated Approach to Studying Interactions of All Components of an Organism's Cells

We conclude this chapter by discussing **systems biology,** an emerging discipline that incorporates data from genomics, transcriptomics, proteomics, and other areas of biology, as well as engineering applications and problem-solving approaches.

In many ways, systems biology is interpreting genomic information in the context of the structure, function, and regulation of biological pathways. As is well known, biological systems are very complex. By studying relationships between all components in an organism, biologists are trying to build a "systems"-level understanding of how organisms function. Systems biologists typically combine recently acquired genomics and proteomics data with years of more traditional studies of gene and protein structure and function. Much of this data is retrieved from databases such as PubMed, GenBank, and other newly emerging genomics, transcriptomics, and proteomics resources. Systems models are used to diagram interactions within a cell or an entire organism, such as protein–protein interactions, protein–nucleic acid interactions, and protein–metabolite interactions (e.g.,



enzyme-substrate binding). These models help systems biologists understand the components of interacting pathways and the interrelationships of molecules in an interacting pathway. In recent years, the term **interactome** has arisen to describe the interacting components of a cell. Systems biologists use several different types of models to diagram protein interaction pathways. One of the most common model types is a **network map**—a sketch showing interacting proteins, genes, and other molecules. These diagrams are essentially the equivalent of an electrical wiring diagram. One disadvantage of network maps is that they are static diagrams that typically lack information about when and where each interaction occurs. Even so, they are a useful foundation for generating computational models that allow the running of simulations to determine how signaling events occur. For example, major groups of kinases, enzymes that phosphorylate other proteins to affect their activity, have been network mapped to show their interactions with each other. Because kinases play such important roles in the regulation of most critical cellular processes, such information about the "kinome" has been very valuable for companies developing drug treatments targeted to certain metabolic pathways.

Network maps are helping scientists model intricate potential interactions of molecules involved in normal and disease processes. **Figure 18–15** shows an example of a network map. This map depicts a human disease network model illustrating the complexity of interactions between genes involved in 22 different human diseases. Look at the cluster of turquoise-colored nodes corresponding to genes involved in several different cancers. One aspect of the map that should be immediately obvious is that a number of cancers share interacting genes even though the cancers affect different organs. Knowing the genes involved and the protein interaction networks for different cancers is a major breakthrough for informing scientists about target genes and proteins to consider for therapeutic purposes. Systems biology is becoming increasingly important in the drug discovery and development process, where its approaches can help scientists and physicians develop a conceptual framework of gene and protein interactions in human disease that can then serve as the rationale for effective drug design.

ESSENTIAL POINT

Systems biology approaches are designed to provide an integrated understanding of interactions between genes, proteins and other molecules that govern complex biological processes. ■

FIGURE 18–14 Mass spectrometry for identifying an unknown protein isolated from a 2D gel. The mass-to-charge spectrum (*m/z*) for trypsin-digested peptides from the unknown protein can be compared to a proteomics database for a spectrum match to identify the unknown protein. The peptide in this example was revealed to have the amino acid sequence serine (S)-glutamine (Q)-alanine (A)-alanine (A)-glutamic acid (E)-leucine (L)-leucine (L), shown in single-letter amino acid code.

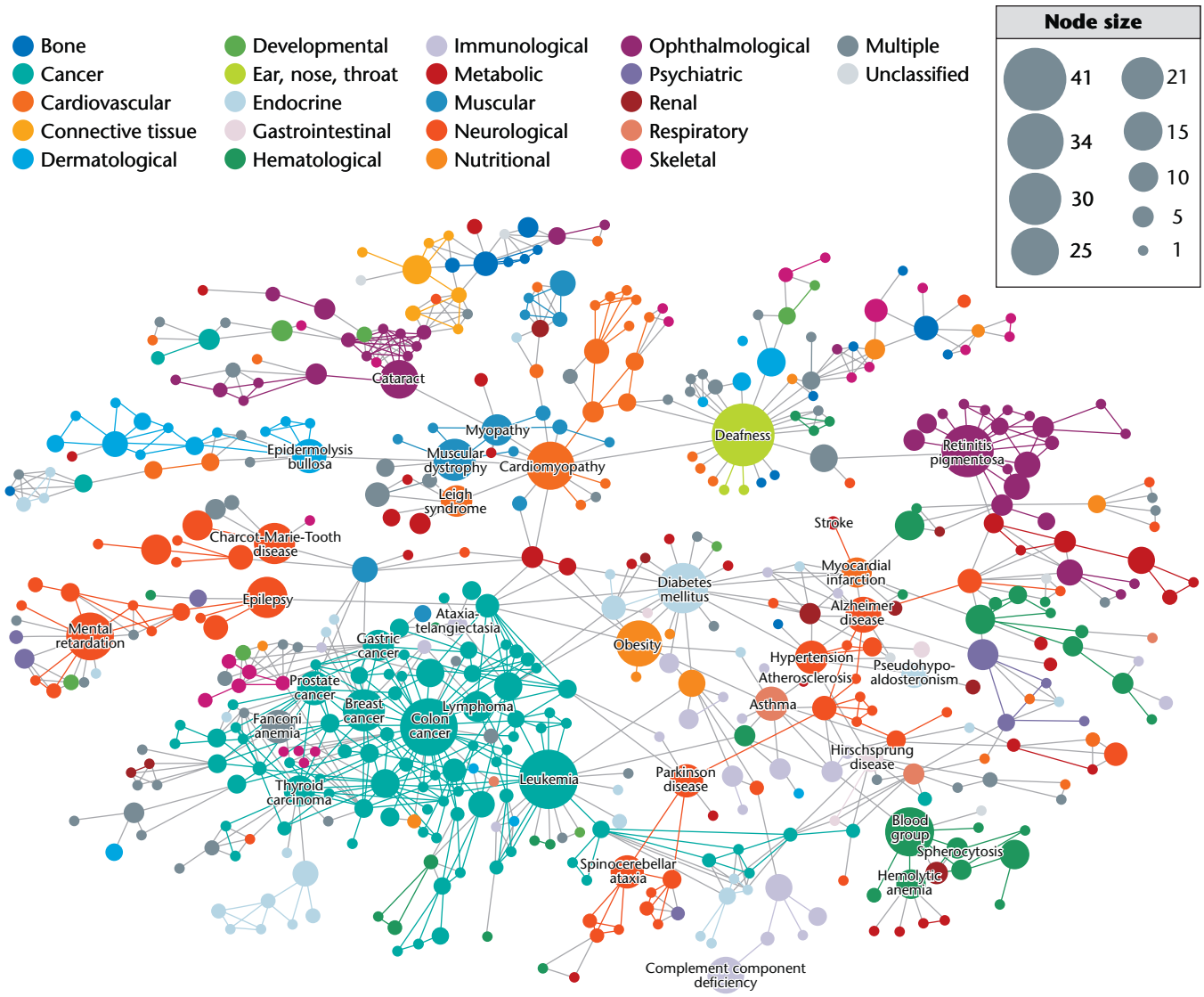
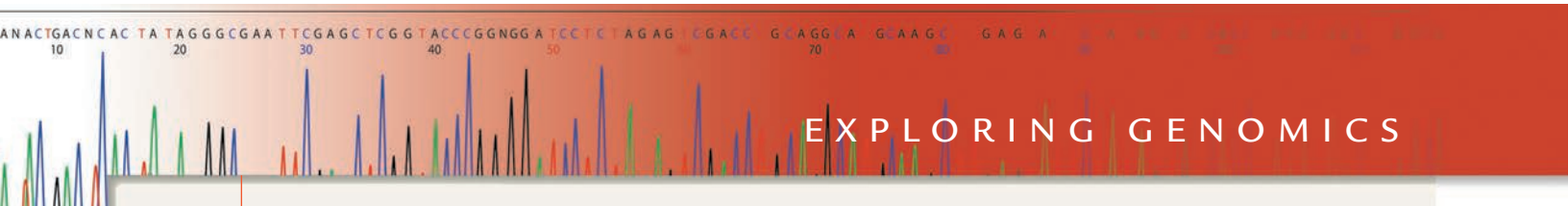


FIGURE 18–15 A systems biology model of human disease gene interactions. The model shows nodes corresponding to 22 specific disorders colored by class. Node size is proportional to the number of genes contributing to the disorder.



Contigs, Shotgun Sequencing, and Comparative Genomics

MasteringGenetics™ Visit the Study Area: Exploring Genomics

In this chapter, we discussed how whole-genome shotgun sequencing methods can be used to assemble chromosome maps. Recall that in the technique of shotgun cloning, chromosomal DNA is digested with different restriction enzymes to create a series of overlapping DNA fragments called contiguous sequences, or "contigs." The contigs are then subjected to DNA sequencing, after which bioinformatics-based programs are used to arrange the contigs in their correct order on the basis of short overlapping sequences of nucleotides.

In this Exploring Genomics exercise you will carry out a simulation of contig alignment to help you to understand the underlying logic of this approach to creating sequence maps of a chromosome. For this purpose, you will use the **National Center for Biotechnology**

Information BLAST site and apply a DNA alignment program called bl2seq.

Exercise I – Arranging Contigs to Create a Chromosome Map

1. Access BLAST from the NCBI Web site at http://blast.ncbi.nlm.nih.gov/Blast.cgi. Locate and select the "Align two sequences using BLAST (bl2seq)" category at the bottom of the BLAST homepage. The bl2seq feature allows you to compare two DNA sequences at a time to check for sequence similarity alignments.
2. Go to the Companion Web site for *Essentials of Genetics* and open the Exploring Genomics exercise for this chapter. Listed are eight contig sequences, called Sequences A through H, taken from an actual human chromosome sequence deposited in GenBank. For this exercise we have used short fragments; however, in reality contigs are usually several thousand base pairs long. To complete this exercise, copy and paste two sequences into the Align feature of BLAST and then run an alignment (by clicking on "Align"). Repeat these steps with other combinations of two sequences to determine which sequences overlap, and then use your findings to create a sequence map that places overlapping contigs in their proper order. Here are a few tips to consider:
   - Develop a strategy to be sure that you analyze alignments for all pairs of contigs.
   - Only consider alignment overlaps that show 100 percent sequence similarity.
3. On the basis of your alignment results, answer the following questions, referring to the sequences by their letter codes (A through H):
   a. What is the correct order of overlapping contigs?
   b. What is the length, measured in number of nucleotides, of each sequence overlap between contigs?
   c. What is the total size of the chromosome segment that you assembled?
   d. Did you find any contigs that do not overlap with any of the others? Explain.
4. Run a nucleotide-nucleotide BLAST search (BLASTn) on any of the overlapping contigs to determine which chromosome these contigs were taken from, and report your answer.

CASE STUDY

Your microbiome may be a risk factor for disease

A number of genes involved in susceptibility to inflammatory bowel disorders (IBDs), including Crohn disease and ulcerative colitis, have been identified. However, it is clear that other risk factors, both genetic and nongenetic, are important in triggering the onset of these diseases. Recent research has centered on understanding the role of the gut microbiome and its interactions with the host genome in IBD. It is known that the microbiome of those with IBD is different from that of those whose IBD is in remission, and it is also different from that of people who do not have IBD. These observations suggest that transfer of microbiota from unaffected individuals via fecal microbial transplantation (FMT) might be a successful treatment for IBD. This idea is supported by the use of FMT as an effective treatment in IBD individuals for a potentially life-threatening infection caused by the bacterium *Clostridium difficile*. Currently, four clinical trials are underway to evaluate the use of FMT as a treatment for IBD.

1. If you had IBD, how would you react if your physician recommended that you enroll in one of these clinical studies to evaluate fecal transplants as a treatment?
2. Current treatment of IBD involves the use of anti-inflammatory drugs, but these drugs achieve remission only in some cases. If genetic analysis reveals that you carry susceptibility alleles for IBD that respond to periodic FMT as a therapy, would you agree to try this method?
3. Before agreeing to FMT, what would you want to know about the microbiomes of individuals who do not have IBD?

INSIGHTS AND SOLUTIONS

1 One of the main problems in annotation is deciding how long a putative ORF must be before it is accepted as a gene. Shown at the right are three different ORF scans of the same *E. coli* genome region—the region containing the *lacY* gene. Regions shaded in brown indicate ORFs. The top scan was set to accept ORFs of 50 nucleotides as genes. The middle and bottom scans accepted ORFs of 100 and 300 nucleotides as genes, respectively. How many putative genes are detected in each scan? The longest ORF covers 1254 bp; the next longest, 234 bp; and the shortest, 54 bp. How can we decide the actual number of genes in this region? In this type of ORF scan, is it more likely that the number of genes in the genome will be overestimated or underestimated? Why?

Solution: Generally, one can examine conserved sequences in other organisms to indicate that an ORF is likely a coding region. One can also match a sequence to previously described sequences that are known to code for proteins. The

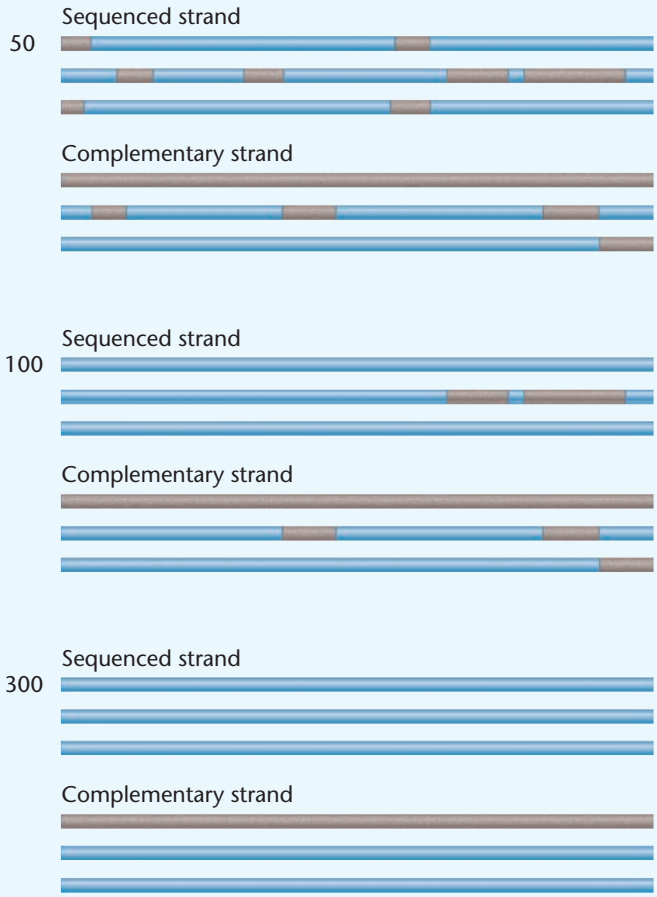
(continued)

Insights and Solutions—continued

problem is not easily solved—that is, deciding which ORF is actually a gene. The shorter the ORFs scan, the more likely the overestimate of genes because ORFs longer than 200 are less likely to occur by chance. For these scans, notice that the 50-bp scans produce the highest number of possible genes, whereas the 300-bp scan produces the lowest number (1) of possible genes.

2 Sequencing of the heterochromatic regions (repeat-rich sequences concentrated in centromeric and telomeric areas) of the *Drosophila* genome indicates that within 20.7 Mb, there are 297 protein-coding genes (Bergman et al. 2002. http://genomebiology.com/2002/3/12/research/0086). Given that the euchromatic regions of the genome contain 13,379 protein-coding genes in 116.8 Mb, what general conclusion is apparent?

Solution: Gene density in euchromatic regions of the *Drosophila* genome is about one gene per 8730 base pairs, while gene density in heterochromatic regions is one gene per 70,000 bases (20.7 Mb/297). Clearly, a given region of heterochromatin is much less likely to contain a gene than the same-sized region in euchromatin.



Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

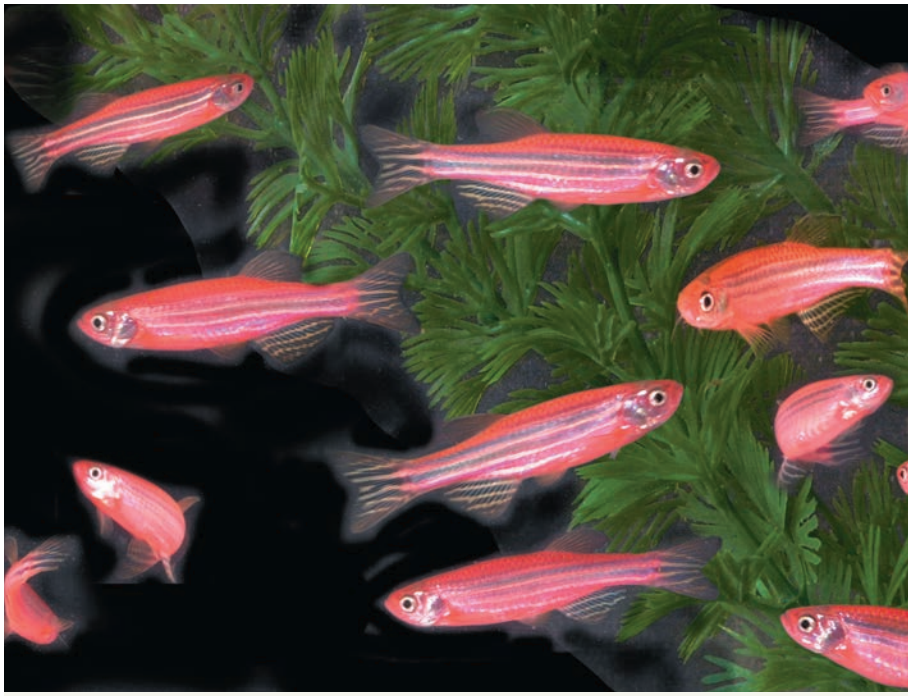
1. In this chapter, we focused on the analysis of genomes, transcriptomes, and proteomes and considered important applications and findings from these endeavors. At the same time, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions:
   (a) How do we know which contigs are part of the same chromosome?
   (b) How do we know if a genomic DNA sequence contains a protein-coding gene?
   (c) What evidence supports the concept that humans share substantial sequence similarities and gene functional similarities with model organisms?
   (d) How can proteomics identify differences between the number of protein-coding genes predicted for a genome and the number of proteins expressed by a genome?
   (e) How have microarrays demonstrated that, although all cells of an organism have the same genome, some genes are expressed in almost all cells, whereas other genes show cell- and tissue-specific expression?

CONCEPT QUESTION

2. Review the Chapter Concepts list on page 361. All of these pertain to how genomics, bioinformatics, and proteomics approaches have changed how scientists study genes and proteins. Write a short essay that explains how recombinant DNA techniques were used to identify and study genes compared to how modern genomic techniques have revolutionized the cloning and analysis of genes. ■
3. What is functional genomics? How does it differ from comparative genomics?
4. Compare and contrast whole-genome shotgun sequencing to a map-based cloning approach.
5. What is bioinformatics, and why is this discipline essential for studying genomes? Provide two examples of bioinformatics applications.
6. List and describe three major goals of the Human Genome Project.
7. Intron frequency varies considerably among eukaryotes. Provide a general comparison of intron frequencies in yeast and humans. What about intron size?
8. BLAST searches and related applications are essential for analyzing gene and protein sequences. Define BLAST, describe basic features of this bioinformatics tool, and provide an example of information provided by a BLAST search.

9. Describe the human genome in terms of genome size, the percentage of the genome that codes for proteins, how much is composed of repetitive sequences, and how many genes it contains. Describe two other features of the human genome.
10. The Human Genome Project has demonstrated that in humans of all races and nationalities approximately 99.9 percent of the sequence is the same, yet different individuals can be identified by DNA fingerprinting techniques. What is one primary variation in the human genome that can be used to distinguish different individuals? Briefly explain your answer.
11. Archaea (formerly known as archaebacteria) is one of the three major divisions of living organisms; the other two are eubacteria and eukaryotes. *Nanoarchaeum equitans* is in the Archaea domain and has one of the smallest genomes known, about 0.5 Mb. How can an organism complete its life cycle with so little genetic material?
12. Through the Human Genome Project (HGP), a relatively accurate human genome sequence was published in 2003 from combined samples from different individuals. It serves as a reference for a haploid genome. Recently, genomes of a number of individuals have been sequenced under the auspices of the Personal Genome Project (PGP). How do results from the PGP differ from those of the HGP?
13. The term *paralog* is often used in conjunction with discussions of hemoglobin genes. What does this term mean, and how does it apply to hemoglobin genes?
14. It can be said that modern biology is experiencing an "omics" revolution. What does this mean? Explain your answer.
15. In what way will the discipline called *metagenomics* contribute to human health and welfare?
16. What are gene microarrays? How are microarrays used?
17. Annotations of the human genome have shown that genes are not randomly distributed, but form clusters with gene "deserts" in between. These "deserts" correspond to the dark bands on G-banded chromosomes. Comparisons between the human transcriptome map and the genome sequence show that highly expressed genes are also clustered together. In terms of genome organization, how is this an advantage?
18. Genomic sequencing has opened the door to numerous studies that help us understand the evolutionary forces shaping the genetic makeup of organisms. Using databases containing the sequences of 25 genomes, scientists (Kreil, D.P. and Ouzounis, C.A., *Nucl. Acids Res.* 29: 1608–1615, 2001) examined the relationship between GC content and global amino acid composition. They found that it is possible to identify thermophilic species on the basis of their amino acid composition alone, which suggests that evolution in a hot environment selects for a certain whole organism amino acid composition. In what way might evolution in extreme environments influence genome and amino acid composition? How might evolution in extreme environments influence the interpretation of genome sequence data?
19. Systems biology models the complex networks of interacting genes, proteins, and other molecules that contribute to human genetic diseases, such as cancer, diabetes, and hypertension. These interactomes show the contribution of each piece towards the whole and where diseases overlap, and provide models for drug discovery and development. Describe some of the differences that might be seen in the interactomes of normal and cancerous cells taken from the same tissue, and explain how these differences could lead to drugs specifically targeted against cancer cells.
20. Exome sequencing is a procedure to help physicians identify the cause of a genetic condition that has defied diagnosis by traditional means. The implication here is that exons in the nuclear genome are sequenced in the hopes that, by comparison with the genomes of nonaffected individuals, a diagnosis might be revealed.
    (a) What are the strengths and weaknesses of this approach?
    (b) If you were ordering exome sequencing for a patient, would you also include an analysis of the patient's mitochondrial genome?

19 Applications and Ethics of Genetic Engineering and Biotechnology



GloFish, marketed as the world's first GM pet, are a controversial product of genetic engineering.

CHAPTER CONCEPTS

- Recombinant DNA technology, genetic engineering, and biotechnology have revolutionized medicine and agriculture.
- Genetically modified plants and animals can serve as bioreactors to produce therapeutic proteins and other valuable protein products.
- Genetic modifications of plants have resulted in herbicide- and pest-resistant crops, and crops with improved nutritional value; similarly, transgenic animals are being created to produce therapeutic proteins and to protect animals from disease.
- A synthetic genome has been assembled and transplanted into a donor bacterial strain, elevating interest in potential applications of synthetic biology.
- Applications of recombinant DNA technology and genomics have become essential for diagnosing genetic disorders, determining genotypes, and scanning the human genome to detect diseases.
- Genome-wide association studies (GWAS) scan for hundreds or thousands of genetic differences in an attempt to link genome variations to particular traits and diseases.
- Medical clinics are adopting whole-genome sequencing of an individual's DNA for disease diagnosis and treatment.
- Computational services for predicting offspring based on a couple's genetics are being advertised to consumers.
- Almost all applications of genetic engineering and biotechnology present unresolved ethical dilemmas that involve important moral, social, and legal issues.

Since the dawn of recombinant DNA technology in the 1970s, scientists have harnessed **genetic engineering** not only for biological research, but also for applications in medicine, agriculture, and biotechnology. Genetic engineering refers to the alteration of an organism's genome and typically involves the use of recombinant DNA technologies to add a gene or genes to a genome, but it can also involve gene removal. The ability to manipulate DNA *in vitro* and to introduce genes into living cells has allowed scientists to generate new varieties of plants, animals, and other organisms with specific traits. These organisms are called **genetically modified organisms (GMOs).**

Biotechnology is the use of living organisms to create a product or a process that helps improve the quality of life for humans or other organisms. Biotechnology as a modern industry began in earnest shortly after recombinant DNA technology developed. But biotechnology is actually a science dating back to ancient civilization and the use of microbes to make many important products, including beverages such as wine and beer, vinegar, breads, and cheeses. Modern biotechnology relies heavily on recombinant DNA technology, genetic engineering, and genomics applications, and these areas will be highlighted in this chapter. Existing products and new developments that occur seemingly every day make the biotechnology industry one of the most rapidly developing branches of the workforce worldwide, encompassing nearly 5000 companies in 54 countries.

The development of the biotechnology industry and the rapid growth in the number of applications for DNA technologies have raised serious concerns about using our power to manipulate genes and to apply gene technologies. Genetic engineering and biotechnology have the potential to provide solutions to major problems globally and to significantly alter how humans deal with the natural world; hence, they raise ethical, social, and economic questions that are unprecedented in human experience. These complex issues cannot be fully explored in the context of an introductory genetics textbook.

This chapter will therefore present only a selection of applications that illustrate the power of genetic engineering and biotechnology and the complexity of the dilemmas they engender. We will begin by explaining how genetic engineering occurs in animals. We briefly describe how genetic engineering has affected the production of pharmaceutical products, and we examine the impact of genetic technologies on the diagnosis and treatment of human diseases, including gene therapy approaches. Finally, we explore some of the social, ethical, and legal implications of genetic engineering and biotechnology.

Please note that many of the topics discussed in this chapter are covered in more detail later in the text (see Special Topic Chapter 3—DNA Forensics, Special Topic Chapter 4—Genomics and Personalized Medicine, Special Topic Chapter 5—Genetically Modified Foods, and Special Topic Chapter 6—Gene Therapy).

19.1 Genetically Engineered Organisms Synthesize a Wide Range of Biological and Pharmaceutical Products

The most successful and widespread application of recombinant DNA technology has been production by the biotechnology industry of recombinant proteins as **biopharmaceutical** products—particularly, therapeutic proteins to treat diseases. Prior to the recombinant DNA era, therapeutic proteins such as insulin, clotting factors, or growth hormones were purified from tissues such as the pancreas, blood, or pituitary glands. These tissues were in limited supply, and the purification processes were expensive. In addition, products derived from these natural sources could be contaminated by disease agents such as viruses. Since human genes encoding important therapeutic proteins can be cloned and expressed in a number of host-cell types, we have more abundant, safer, and less expensive sources of biopharmaceuticals. **Biopharming** is a commonly used term to describe the production of valuable proteins in genetically modified (GM) animals and plants.

In this section, we outline several examples of therapeutic products that are produced by expression of cloned genes in transgenic host cells and organisms. It should not surprise you that cancers, arthritis, diabetes, heart disease, and infectious diseases such as AIDS are among the major diseases that biotechnology companies are targeting for treatment by recombinant therapeutic products. **Table 19.1** provides a short list of important recombinant products currently synthesized in transgenic bacteria, plants, yeast, and animals.

Insulin Production in Bacteria

Many therapeutic proteins have been produced by introducing human genes into bacteria. In most cases, the human gene is cloned into a plasmid, and the recombinant vector is introduced into the bacterial host. Large quantities of the transformed bacteria are grown, and the recombinant human protein is recovered and purified from bacterial extracts.

The first human gene product manufactured by recombinant DNA technology was human insulin, called Humulin, which was licensed for therapeutic use in 1982 by the **U.S. Food and Drug Administration** (**FDA**), the government agency responsible for regulating the safety of food and drug products and medical devices. In 1977, scientists at Genentech, the San Francisco biotechnology company cofounded in 1976 by Herbert Boyer (one of the pioneers of using plasmids for recombinant DNA technology) and Robert Swanson, isolated and cloned the gene for insulin and expressed it in bacterial cells. Genentech, short for "genetic engineering technology," is also generally regarded as the world's first biotechnology company.

Previously, insulin was chemically extracted from the pancreas of cows and pigs obtained from slaughterhouses. **Insulin** is a protein hormone that regulates glucose metabolism. Individuals who cannot produce insulin have diabetes, a disease that, in its more severe form (type I), affects more than 2 million individuals in the United States. Although synthetic human insulin can now be produced by another process, a look at the original genetic engineering method is instructive, as it shows both the promise and the difficulty of applying recombinant DNA technology.

Clusters of cells embedded in the pancreas synthesize a precursor polypeptide known as preproinsulin. As this polypeptide is secreted from the cell, amino acids are cleaved from the end and the middle of the chain. These cleavages produce the mature insulin molecule, which contains two polypeptide chains (the *A* and *B* chains) joined by disulfide bonds. The *A* subunit contains 21 amino acids, and the *B* subunit contains 30.

TABLE 19.1 Examples of Genetically Engineered Biopharmaceutical Products Available or Under Development

Gene Product	Condition Treated	Host Type
Erythropoitin	Anemia	*E. coli*; cultured mammalian cells
Interferons	Multiple sclerosis, cancer	*E. coli*; cultured mammalian cells
Tissue plasminogen activator tPA	Heart attack, stroke	Cultured mammalian cells
Human growth hormone	Dwarfism	Cultured mammalian cells
Monoclonal antibodies against vascular endothelial growth factor (VEGF)	Cancers	Cultured mammalian cells
Human clotting factor VIII	Hemophilia A	Transgenic sheep, pigs
C1 inhibitor	Hereditary angioedema	Transgenic rabbits
Recombinant human antithrombin	Hereditary antithrombin deficiency	Transgenic goats
Hepatitis B surface protein vaccine	Hepatitis B infections	Cultured yeast cells, bananas
Immunoglobulin IgG1 to HSV-2	Herpesvirus infections	Transgenic soybeans glycoprotein B
Recombinant monoclonal antibodies	Passive immunization against rabies (also used in diagnosing rabies), cancer, rheumatoid arthritis	Transgenic tobacco, soybeans, cultured mammalian cells
Norwalk virus capsid protein	Norwalk virus infections	Potato (edible vaccine)
E. coli heat-labile enterotoxin	*E. coli* infections	Potato (edible vaccine)

In the original bioengineering process, genes that encode the *A* and *B* subunits were constructed by oligonucleotide synthesis (63 nucleotides for the *A* polypeptide and 90 nucleotides for the *B* polypeptide). Each synthetic oligonucleotide was inserted into a separate vector, adjacent to the *lacZ* gene encoding the bacterial form of the enzyme -galactosidase. When transferred to a bacterial host, the *lacZ* gene and the adjacent synthetic oligonucleotide were transcribed and translated as a unit. The product is a **fusion protein**—that is, a hybrid protein consisting of the amino acid sequence for -galactosidase attached to the amino acid sequence for one of the insulin subunits (**Figure 19–1**). The fusion proteins were purified from bacterial extracts and treated with cyanogen bromide, a chemical that cleaves the fusion protein from the -galactosidase. When the fusion products were mixed, the two insulin subunits spontaneously united, forming an intact, active insulin molecule. The purified injectable insulin was then packaged for use by diabetics.

Shortly after insulin became available, growth hormone—used to treat children who suffer from a form of dwarfism—was cloned. Soon, recombinant DNA technology made that product readily available too, as well as a wide variety of other medically important proteins that were once difficult to obtain in adequate amounts. Since recombinant insulin ushered in the biotechnology era, well over 200 recombinant products have entered the market worldwide. In recent years, the development of many other, nonbiopharmaceutical products has been a very active area of research. One example includes the production in *E. coli* of the antioxidant lycopene found in tomatoes. Lycopene produced by *E. coli* is not yet available for human consumption.

Transgenic Animal Hosts and Pharmaceutical Products

Although bacteria have been widely used to produce therapeutic proteins, there are some disadvantages in using prokaryotic hosts to synthesize eukaryotic proteins. One problem is that bacterial cells often cannot process and modify eukaryotic protein for full biological activity, rendering them inactive. In addition, eukaryotic proteins produced in prokaryotic cells often do not fold into the proper three-dimensional conformation and are therefore inactive. To overcome these difficulties and increase yields, many biopharmaceuticals are produced in eukaryotic hosts. As seen in Table 19.1, eukaryotic hosts may include cultured eukaryotic cells (plant or animal) or transgenic farm animals. For example, a herd of goats or cows can serve as very effective **bioreactors** or **biofactories**—living factories—that could continuously make milk containing the desired therapeutic protein that can then be isolated in a noninvasive way.

Yeast are also valuable hosts for expressing recombinant proteins. Even insect cells are valuable for this purpose, through the use of a gene delivery system (virus) called **baculovirus.** Recombinant baculovirus containing a gene of interest is used to infect insect cell lines, which then express the protein at high levels. Baculovirus-insect cell expression is particularly useful for producing human recombinant proteins that are heavily glycosylated.

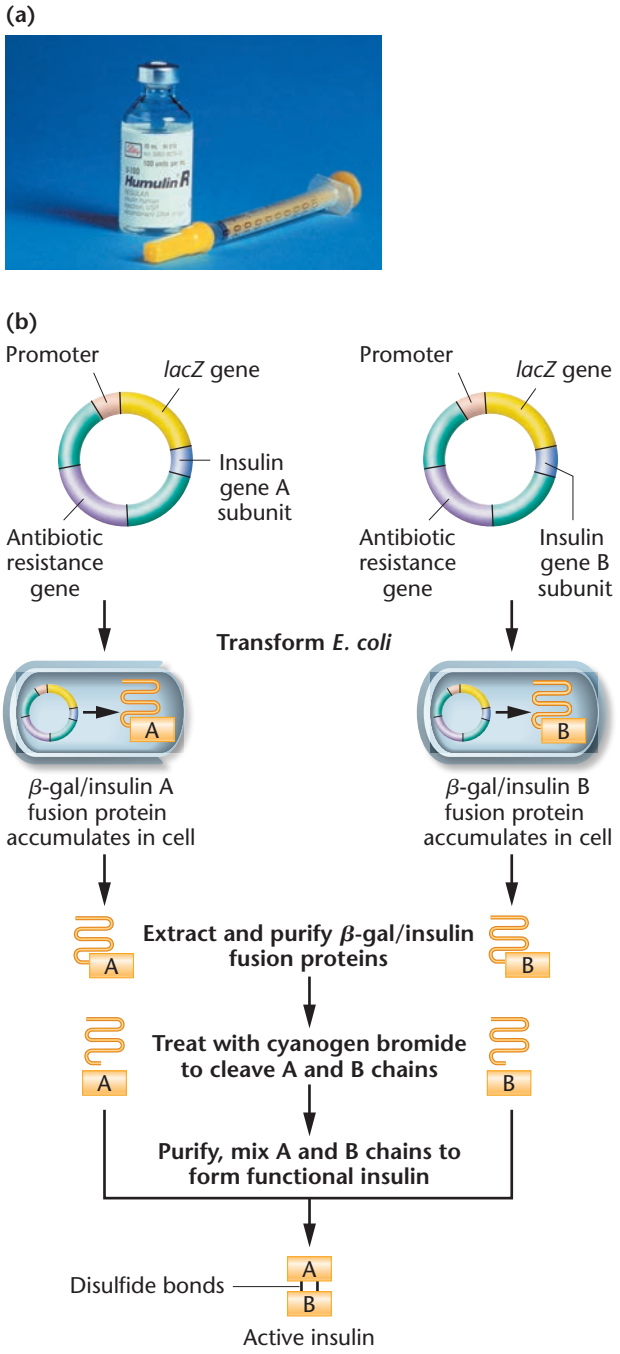


FIGURE 19–1 (a) Humulin, a recombinant form of human insulin, was the first therapeutic protein produced by recombinant DNA technology to be approved for use in humans. (b) To synthesize recombinant human insulin, synthetic oligonucleotides encoding the insulin A and B chains were inserted (in separate vectors) at the tail end of a cloned *E. coli lacZ* gene. The recombinant plasmids were transformed into *E. coli* host cells, where the -gal/insulin fusion protein was synthesized and accumulated in the cells. Fusion proteins were then extracted from the host cells and purified. Insulin chains were released from -galactosidase by treatment with cyanogen bromide. The insulin subunits were purified and mixed to produce a functional insulin molecule.

Regardless of the host, therapeutic proteins may then be purified from the host cells—or when **transgenic animals** are used, isolated from animal products such as milk.

Refer to our discussion earlier in the text (see Chapter 17) on how transgenic animals can be created to allow for expression of a transgene of interest. Biotechnology companies are working on expressing many different genes in transgenic animals for the purpose of expressing, isolating, and purifying commercially valuable proteins.

In 2006, recombinant human **antithrombin,** an anticlotting protein, became the world's first drug extracted from the milk of farm animals to be approved for use in humans. Scientists at GTC Biotherapeutics introduced the human antithrombin gene into goats. By placing the gene adjacent to a promoter for beta casein, a common protein in milk, GTC scientists were able to target antithrombin expression in the mammary gland. As a result, antithrombin protein is highly expressed in the milk. In one year, a single goat will produce the equivalent amount of antithrombin that in the past would have been isolated from 90,000 blood collections.

Recombinant DNA Approaches for Vaccine and Antibody Production

Another successful application of recombinant DNA technology for therapeutic purposes is the production of vaccines. Vaccines stimulate the immune system to produce antibodies against disease-causing organisms and thereby confer immunity against specific diseases. Traditionally, two types of vaccines have been used: **inactivated vaccines,** which are prepared from killed samples of the infectious virus or bacteria; and **attenuated vaccines,** which are live viruses or bacteria that can no longer reproduce but can cause a mild form of the disease. Inactivated vaccines include the vaccines for rabies and influenza; vaccines for tuberculosis, cholera, and chickenpox are examples of attenuated vaccines.

Genetic engineering is being used to produce **subunit vaccines,** which consist of one or more surface proteins from the virus or bacterium but not the entire virus or bacterium. Often the surface protein is produced through recombinant DNA technology by cloning and expressing the genes encoding the protein to be used for the vaccine. This surface protein acts as an antigen that stimulates the immune system to make antibodies that act against the organism from which it was derived. One of the first subunit vaccines was made against the **hepatitis B virus,** which causes liver damage and cancer. The gene that encodes the hepatitis B surface protein was cloned into a yeast expression vector, and the cloned gene was expressed in yeast host cells. The protein was then extracted and purified from the host cells and packaged for use as a vaccine.

In 2006, the FDA approved **Gardasil,** a subunit vaccine produced by the pharmaceutical company Merck and the first cancer vaccine to receive FDA approval. Gardasil targets four strains of **human papillomavirus** (**HPV**) that cause 70 percent of cervical cancers. Approximately 70 percent of sexually active women will be infected by an HPV strain during their lifetime. Gardasil is designed to provide immune protection against HPV prior to infection but is not effective against existing infections. You may have heard of Gardasil through media coverage of the legislation proposed in several states that would require all adolescent school girls to receive a Gardasil vaccination regardless of whether or not they are sexually active.

Vaccine Proteins and Antibodies Can Be Produced by Plants

Plants offer several other advantages for expressing recombinant proteins. For instance, once a transgenic plant is made, it can easily be grown and replicated in a greenhouse or field, and it will provide a constant source of recombinant protein. In addition, the cost of expressing a recombinant protein in a transgenic plant is typically much lower than making the same protein in bacteria, yeast, or mammalian cells.

No recombinant proteins expressed in transgenic plants have yet been approved for use by the FDA as therapeutic proteins for humans, although about a dozen products are close to making it through final clinical trials. Some edible vaccines are now in clinical trials. For example, a vaccine against a bacterium that causes cholera has been produced in genetically engineered potatoes and used to successfully vaccinate human volunteers.

In 2014, an outbreak of Ebola virus in West Africa killed over 1500 people, with many more cases unreported. Ebola causes hemorrhagic fever and produces fatality rates of approximately 90 percent. There is no effective treatment for curing or preventing Ebola virus infection. But antibodies against Ebola expressed in tobacco leaves are showing promise in ongoing clinical trials. Mice were used to create monoclonal antibodies against the virus. The antibody genes were then introduced into tobacco plants. The transgenic tobacco plants express high quantities of the antibody proteins, which can then be isolated and purified for use in humans. Transgenic tobacco plants are commonly used for expressing recombinant proteins because of the large size of their leaves and relatively high yield of recombinant proteins compared to other plants.

DNA-Based Vaccines

DNA-based vaccines have been attempted for many years, and recently there has been renewed interest in using these vaccines to protect against viral pathogens. In this approach, DNA encoding proteins from a particular pathogen are inserted into plasmid vectors, which are then injected directly into an individual or delivered via a viral vector similar to the way certain viruses are used for gene therapy. The idea here is that pathogen proteins encoded by the delivered DNA would be produced and trigger an immune response that could provide protection should an immunized person be exposed to the pathogen in the future.

For example, trials are underway using plasmid DNA encoding protein antigens from HIV as an attempt to vaccinate individuals against HIV. Thus far, a major limitation of DNA-based vectors has been that they typically result in very low production of protein encoded by delivered genes, and thus the immune response in vaccinated persons is insufficient to provide the desired protection. Work on DNA-based vaccines continues to be an active area of exploration, but whether they will ever have significant roles in the vaccine market remains to be seen.

ESSENTIAL POINT

Recombinant DNA technology can be used to produce valuable biopharmaceutical protein products such as therapeutic proteins for treating disease. ■

NOW SOLVE THIS

19–1 In order to vaccinate people against diseases by having them eat antigens (such as the cholera toxin) or antibodies expressed in an edible vaccine, the antigen must reach the cells of the small intestine. What are some potential problems of this method?

■ **HINT:** *This problem asks you to consider why edible vaccines may not be effective. The key to its solution is to consider the molecular structure of the antigen or antibody and its recognition by the immune system.*

19.2 Genetic Engineering of Plants Has Revolutionized Agriculture

For millennia, farmers have manipulated the genetic makeup of plants and animals to enhance food production. Until the advent of genetic engineering 30 years ago, these genetic manipulations were primarily restricted to **selective breeding**—the selection and breeding of naturally occurring or mutagen-induced variants. In the last 50 to 100 years, genetic improvement of crop plants through the traditional methods of artificial selection and genetic crosses has resulted in dramatic increases in productivity and nutritional enhancement. For example, maize yields have

FIGURE 19–2 Selective breeding is one of the oldest methods of genetic alteration of plants. Shown here is teosinte (*Zea canina*, left), a selectively bred hybrid (center), and modern corn (*Zea mays*).

increased fourfold over the last 60 years, and more than half of this increase is due to genetic improvement by artificial selection and selective breeding (**Figure 19–2**). Modern maize has substantially larger ears and kernels than the predecessor crops, including hybrids from which it was bred.

Recombinant DNA technology provides powerful new tools for altering the genetic constitution of agriculturally important organisms. Scientists can now identify, isolate, and clone genes that confer desired traits, then specifically and efficiently introduce these into organisms. As a result, it is possible to quickly introduce insect resistance, herbicide resistance, or nutritional characteristics into farm plants and animals, a primary purpose of **agricultural biotechnology.**

There are many examples of applications of recombinant DNA and genomics involving the use of plants; we will defer those discussions in detail. Genetic modifications of plant crops are discussed in greater detail later in the text (see Special Topic Chapter 5—Genetically Modified Foods). In this section we primarily consider genetic manipulations to produce transgenic crop plants of agricultural value. In Section 19.3, we will discuss examples of genetic manipulations of agriculturally important animals.

Worldwide, over 4 billion acres of genetically engineered crops have been planted, particularly herbicide- and pest-resistant soybeans, corn, cotton, and canola; over 50 different transgenic crop varieties are available, including alfalfa, corn, rice, potatoes, tomatoes, tobacco, wheat, and cranberries.

The main reasons for generating transgenic crops include:

- Improving the growth characteristics and yield of agriculturally valuable crops
- Increasing the nutritional value of crops
- Providing crop resistance against insect and viral pests, drought, and herbicides

In addition, many new GM crops and microalgae are being designed for ethanol production and for making biodiesel fuel—that is, for providing sustainable sources of energy.

Insights from plant genome sequencing projects will undoubtedly be the catalyst for analysis of genetic diversity in crop plants, identification of genes involved in crop domestication and breeding traits, and subsequent enhancement of a variety of desirable traits through genetic engineering. In the past several years genome projects have been completed for many major food and industrial crops, including the three crops that account for most of the world's caloric intake: maize, rice, wheat. The genome for a popular crop species of coffee plants was recently sequenced. Plant scientists expect to use genome data to improve coffee crop growth and eventually to improve crop phenotypes to produce the most desirable attributes for coffee seeds.

ESSENTIAL POINT

Genetically modified (GM) plants, designed to improve crop yield and nutritional value and to increase resistance to herbicides, pests, and severe weather, are becoming more prevalent worldwide. ■

19.3 Transgenic Animals Serve Important Roles in Biotechnology

Although genetically engineered plants are major players in modern agriculture, commercial applications of transgenic animals are less widespread. Most transgenic animals are created for research purposes to study gene function. Nonetheless, some high-profile examples of genetically engineered farm animals have aroused public interest and controversy. It is expected that in the future transgenic animals may increasingly be available for commercial purposes.

Examples of Transgenic Animals

Oversize mice containing a human growth hormone transgene were some of the first transgenic animals created. Attempts to create farm animals containing transgenic growth hormone genes have not been particularly successful, probably because growth is a complex, multigene trait. One notable exception is the transgenic Atlantic salmon,

bearing copies of a Chinook salmon growth hormone gene adjacent to a constitutive promoter (see Special Topic Chapter 5—Genetically Modified Foods).

As discussed in Section 19.1, currently, the major uses for transgenic farm animals are as bioreactors to produce useful pharmaceutical products, but a number of other interesting transgenic applications are under development. Several of these applications are designed to increase milk production or increase nutritional value of milk. Significant research efforts are also being made to protect farm animals against common pathogens that cause disease and animal loss (including potential bioweapons that could be used in a terrorist attack on food animals) and put the food supply at risk. For instance, controlling mastitis in cattle by creating transgenic cows has shown promise (**Figure 19–3**). **Mastitis** is an infection of the mammary glands. It is the most costly disease affecting the dairy industry, leading to over $2 billion in losses in the United States. Mastitis can block milk ducts, reducing milk output, and can also contaminate the milk with pathogenic microbes. Infection by the bacterium *Staphylococcus aureus* is the most common cause of mastitis, and most cattle with mastitis typically do not respond well to conventional treatments with antibiotics. As a result, mastitis is a significant cause of herd reduction.

In an attempt to create cattle resistant to mastitis, transgenic cows were generated that possessed the lysostaphin gene from *Staphylococcus simulans.* Lysostaphin is an enzyme that specifically cleaves components of the *S. aureus* cell wall. Transgenic cows expressing this protein in milk produce a natural antibiotic that wards off *S. aureus* infections. These transgenic cows do not completely solve the mastitis problem because lysostaphin is not effective against other microbes such as *E. coli* and *S. uberis* that occasionally cause mastitis; moreover, there is also the potential that *S. aureus* may develop resistance to lysostaphin. Nonetheless, scientists are cautiously optimistic that transgenic approaches have a strong future for providing farm animals with a level of protection against major pathogens.

Researchers in New Zealand have engineered a cow to produce hypoallergenic milk. This research effort has been spurred by the fact that an estimated 2–3 percent of babies are allergic to milk from dairy cows and develop a reaction to a protein called -lactoglobulin (BLG). The approach of these researchers involved designing miRNAs to inhibit BLG and then using a transgenic approach to introduce these genes into cow embryos. Of 100 GM cow embryos, only one produced a calf, Daisy. As Daisy began lactating, researchers found that her milk did not have any detectable levels of BLG. Currently, studies are underway to determine if Daisy's milk is less allergenic to mice, with future plans to test whether humans are allergic to Daisy's milk.

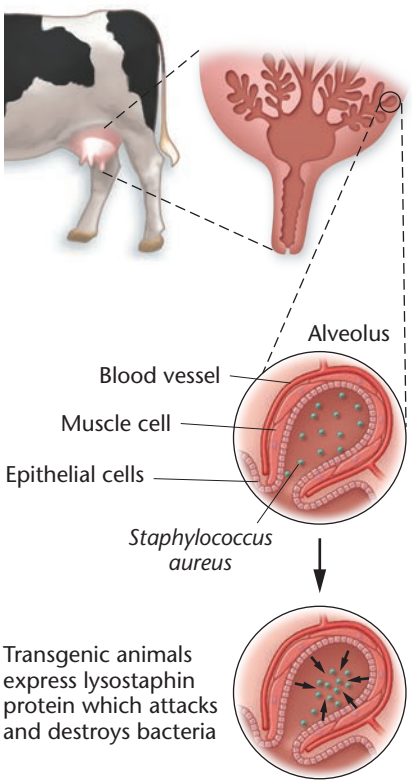


FIGURE 19–3 Transgenic cows for battling mastitis. The mammary glands of nontransgenic cows are highly susceptible to infection by the skin microbe *Staphylococcus aureus.* Transgenic cows express the lysostaphin transgene in milk, where it can kill *S. aureus* before they can multiply in sufficient numbers to cause inflammation and damage mammary tissue.

Scientists at Yorktown Industries of Austin, Texas, created the **GloFish,** a transgenic strain of zebrafish (*Danio rerio*) containing a red fluorescent protein gene from sea anemones. Marketed as the first GM pet in the United States, GloFish fluoresce bright pink when illuminated by ultraviolet light (see the opening photograph at the beginning of this chapter). GM critics describe these fish as an abuse of genetic technology. However, GloFish may not be as frivolous a use of genetic engineering as some believe. A variation of this transgenic model, incorporating a heavy-metal-inducible promoter adjacent to the red fluorescent protein gene, has shown promise in a bioassay for heavy metal contamination of water. When these transgenic zebrafish are in water contaminated by mercury and other heavy metals, the promoter becomes activated, inducing transcription of the red fluorescent protein gene. In this way, zebrafish fluorescence can be used as a bioassay to measure heavy metal contamination and uptake by living organisms.

ESSENTIAL POINT

Transgenic animals with improved growth characteristics or desirable phenotypes are being genetically engineered for a number of different applications. ■

19.4 Synthetic Genomes and the Emergence of Synthetic Biology

Studying genomes has led to a fundamental question: "what is the minimum number of genes necessary to support life?" Determining the answer to this question is the first step in the ultimate creation of synthetic genomes that can, in turn, lead to the production of artificial cells or organisms.

To help advance synthetic genome work, we can use the small genomes of obligate parasites. For example, the bacterium *Mycoplasma genitalium,* a human parasitic pathogen, is among the simplest self-replicating prokaryotes known and has served as a model for understanding the minimal elements of a genome necessary for a self-replicating cell. *M. genitalium* has a genome of 580 kb.

In 2010, scientists from the J. Craig Venter Institute (JCVI) published the first report of a functional synthetic genome. In this approach they designed and had chemically synthesized more than one thousand 1080-bp segments called cassettes covering the entire 1.08-Mb *M. mycoides* genome (**Figure 19.4**). To assemble these segments correctly, the sequences had 80-bp sequences at each end which overlapped with their neighbor sequences. These sequences were cloned in *E. coli*. Then, using the yeast *Saccharomyces cerevisiae,* a homologous recombination approach was used to organize the sequences into 11 separate 100-kb assemblies that were eventually combined to completely span the entire 1.08-Mb *M. mycoides* genome.

The entire assembled genome, called JCVI-syn1.0, was then subjected to the ultimate test of functionality: transplantation into another cell, in this case, another bacterium. They transplanted it into a close relative, *M. capricolum*. This resulted in cells with the JCVI-syn1.0 genotype and phenotype. Transformation of *M. capricolum* into JCBI-syn 1.0 *M. mycoides* was verified, in part, because these cells were shown to express the *lacZ* gene which was only present in the synthetic genome. The recipient cells also made proteins characteristic of *M. mycoides* and not *M. capricolum*, verifying the strain conversion.

The synthetic genome effectively rebooted the *M. capricolum* recipient cells to change them from one form to another. When this work was announced, J. Craig Venter claimed: "This is equivalent to changing a Macintosh computer into a PC by inserting a new piece of PC software." This is tedious work. Ninety-nine percent of the experiments involved failed! A single base error among a million bp would derail the project for several months.

Venter's recent work with *M. mycoides* JCVI-syn1.0, a decade-long project that cost about $40 million, is being hailed as a defining moment in the emerging field of **synthetic biology.** Synthetic biology applies engineering principles and designs to biological systems.

There are many fundamental questions about synthetic genomes and genome transplantation that need to be answered. But clearly these studies provided key "proof of concept" that synthetic genomes could be produced, assembled, and successfully transplanted to create a microbial strain encoded by a synthetic genome and bring scientists closer to producing novel synthetic genomes incorporating genes for specific traits of interest. An international effort is now underway to produce synthetic chromosomes comprising an entire yeast genome, about 12.5 million bases. In 2014, a synthetic version of yeast (*S. cerevisiae*) chromosome III was created, the first such eukaryotic chromosome.

What are other potential applications of synthetic genomes and synthetic biology? One of JCVI's goals is to create microorganisms that can be used to synthesize biofuels. Other possibilities exist such as the creation of synthetic microbes engineered to degrade pollutants (bioremediation), the synthesis of new biopharmaceutical products, synthesizing chemicals and fuels from sunlight and carbon dioxide, genetically programmed bacteria to help us heal, and "semisynthetic" crops that contain synthetic chromosomes encoding genes for beneficial traits such as drought resistance or improved photosynthetic efficiency.

ESSENTIAL POINT

Synthetic genomes and synthetic biology offer the potential for geneticists to create genetically engineered cells with novel characteristics that may have commercial value. ■

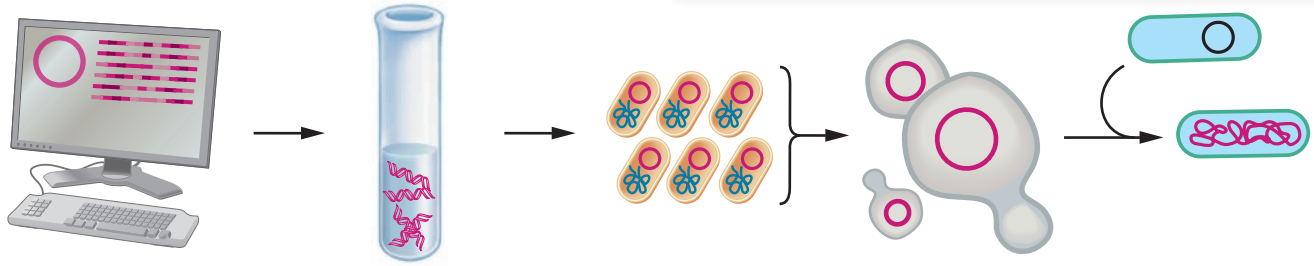


FIGURE 19–4 Building a synthetic version of the 1.08-Mb *Mycoplasma mycoides* genome JCVI-syn1.0.

19.5 Genetic Engineering and Genomics Are Transforming Medical Diagnosis

Gene-based technologies have had a major impact on the diagnosis of disease and are revolutionizing medical treatments and the development of specific and effective pharmaceuticals. In large part as a result of the Human Genome Project, researchers are identifying genes involved in both single-gene diseases and complex genetic traits. In this section, we provide an overview of representative examples that demonstrate how gene-based technologies are being used to diagnose genetic diseases.

Using DNA-based tests, scientists can directly examine a patient's DNA for mutations associated with disease. Gene testing was one of the first successful applications of recombinant DNA technology, and currently more than 900 gene tests are in use. These tests usually detect DNA mutations associated with single-gene disorders that are inherited in a Mendelian fashion. Examples of such genetic tests are those that detect sickle-cell anemia, cystic fibrosis, Huntington disease, hemophilias, and muscular dystrophies. Other genetic tests have been developed for complex disorders such as breast and colon cancers.

Gene tests are used to perform prenatal diagnosis of genetic diseases, to identify carriers, to predict the future development of disease in adults, to confirm the diagnosis of a disease detected by other methods, and to identify genetic diseases in embryos created by *in vitro* fertilization. For genetic testing of adults, DNA from white blood cells is commonly used. Alternatively, many genetic tests can be carried out on cheek cells collected by swabbing the inside of the mouth, or hair cells. Some genetic testing can be carried out on gametes.

Prenatal Genetic Testing

The genetic testing of adults is increasing, as is the screening of newborns for genetic disorders (an ethically controversial issue that we will discuss in Section 19.8). In newborns, a simple prick of the heel of a baby produces a few drops of blood that are used to check the newborn for genetic disorders. Over the past two decades more genetic tests have been used to detect genetic conditions in babies than in adults.

All states now require newborn screening for certain medical conditions. And there are about 60 conditions that can be screened for. In the United States, newborn screening identifies about 12,500 children with medical disorders out of approximately 4 million babies born each year.

For prenatal diagnosis, fetal cells are obtained by **amniocentesis** or **chorionic villus sampling.** Figure 19–5 shows the procedure for amniocentesis, in which a small volume of the amniotic fluid surrounding the fetus is removed. Amniotic fluid contains fetal cells that can be used for karyotyping, genetic testing, and other procedures. For chorionic villus sampling, cells from the fetal portion of the

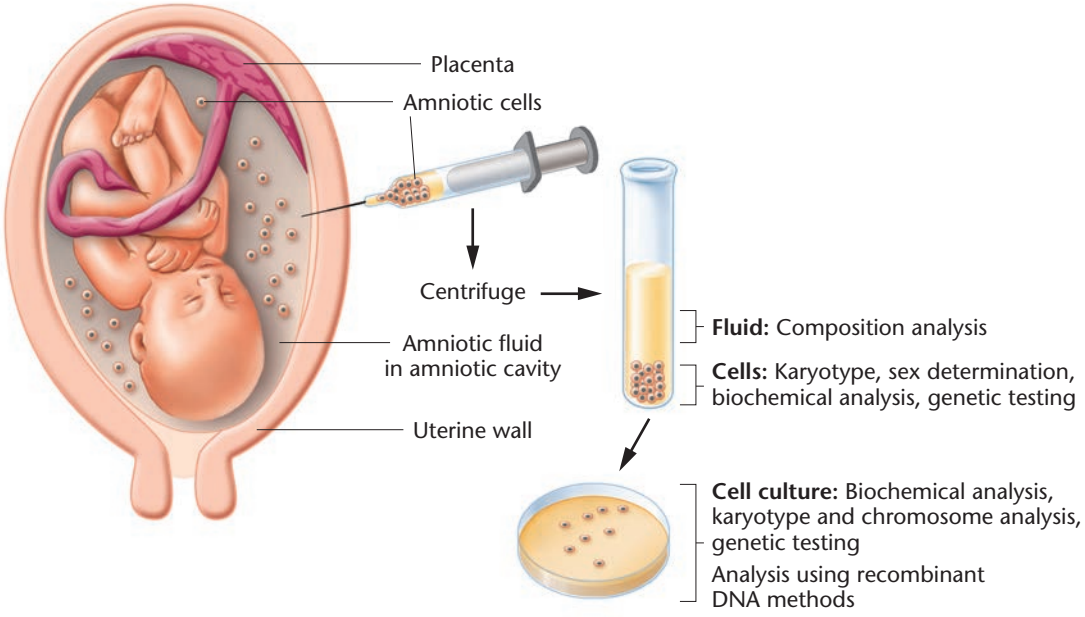


FIGURE 19–5 For amniocentesis, the position of the fetus is first determined by ultrasound, and then a needle is inserted through the abdominal and uterine walls to recover amniotic fluid and fetal cells for genetic or biochemical analysis.

placental wall (the chorionic villi) are sampled through a vacuum tube, and analyses can be carried out on this tissue. Captured fetal cells can then be subjected to genetic analysis, usually involving techniques that involve PCR (such as allele-specific oligonucleotide testing, described later in this section).

Noninvasive procedures are being developed for prenatal genetic testing of fetal DNA. These procedures are making prenatal testing easier with little to no risk to the fetus. Circulating in each person's bloodstream is cell-free DNA that is released from dead and dying cells. This DNA is cut up into small fragments by enzymes in the blood. The blood of a pregnant woman contains snippets of DNA from cells of the fetus. It is estimated that 3 to 6 percent of the DNA in a pregnant mother's blood belong to her baby. It is now possible to analyze these traces of fetal DNA to determine if the baby has certain types of genetic conditions such as Down syndrome. Such tests require about a tablespoon of blood.

DNA in the blood is sequenced to analyze **haplotypes,** contiguous segments of DNA that do not undergo recombination during gamete formation, that distinguish which DNA segments are maternal and which are from the fetus (see Figure 19–6). If a fetal haplotype contained a specific mutation, this would also be revealed by sequence analysis. In addition, nearly complete fetal genome sequences have been assembled from maternal blood. These are developed by sequencing DNA fragments from maternal blood and comparing those fragments to sequenced genomes from the mother and father. Bioinformatics software is then used to organize the genetic sequences from the fetus in an effort to assemble the fetal genome. Currently, this technology results in an assembled genome sequence with segments missing, so it does not capture the entire fetal genome. It has been shown, however, that whole-genome shotgun sequencing of maternal plasma DNA can be used to accurately sequence the entire exome of a fetus.

Tests for fetal genetic analysis based on maternal blood samples started to arrive on the market in 2011. Sequenom of San Diego, California, was one of the first companies to launch such a test—**MaterniT21®Plus,** a Down syndrome test that can also be used to test for trisomy 13 (Patau syndrome) and trisomy 18 (Edward syndrome). The MaterniT21®Plus test analyzes 36-bp fragments of DNA to identify chromosome 21 from the fetus. Sequenom claims that this test is highly accurate with a false positive rate of just 0.2 percent. The MaterniT21®Plus test can be done as early as week 10 (about the same time at which CVS sampling can be done, which is about 4 to 6 weeks earlier than amniocentesis can be performed). While not intended to replace amniocentesis, it can help prevent some women from having amniocentesis after a false positive report from ultrasound or protein marker tests.

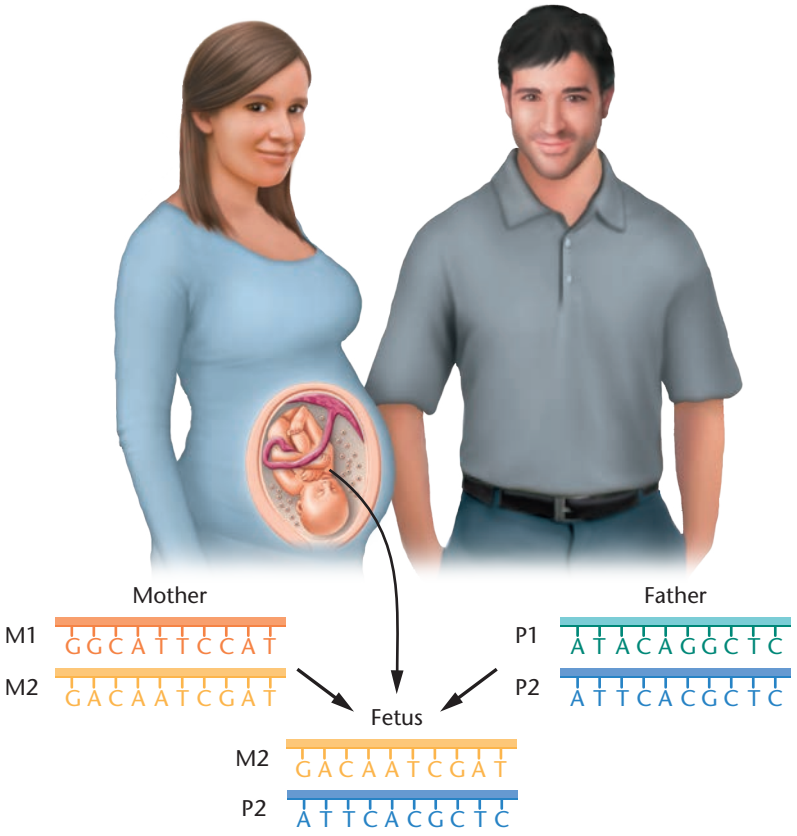


FIGURE 19–6 Deducing fetal genome sequences from maternal blood. For any given chromosome, a fetus inherits one copy of a haplotype from the mother (maternal copies, M1 or M2) and another from the father (paternal copies, P1 or P2). For simplicity, a single-stranded sequence of DNA from each haplotype is shown. Here the fetus inherited haplotypes are M2 and P2 from the mother and father, respectively. DNA from the blood of a pregnant woman would contain paternal haplotypes inherited by the fetus (P2, blue), maternal haplotypes that are not passed to the fetus (M1, orange), and maternal haplotypes that are inherited by the fetus (M2, yellow). The maternal haplotype inherited by the fetus (M2) would be present in excess amounts relative to the haplotype that is not inherited (M1). These haplotype sequences can be detected by whole-genome shotgun sequencing.

In Section 19.9 we will discuss preconception testing and recent patents for computing technologies designed to predict the genetic potential of offspring (destiny tests).

Genetic Tests Based on Restriction Enzyme Analysis

A classic method of genetic testing is **restriction fragment length polymorphism (RFLP) analysis.** As we will discuss in the next section, PCR-based methods have largely replaced RFLP analysis; however, applications of this approach are still used occasionally, and for historical purposes it is also

helpful to compare RFLP analysis to new approaches, which were largely not widespread prior to completion of the HGP. To illustrate this method, we examine the prenatal diagnosis of **sickle-cell anemia.** As we have discussed before, this disease is an autosomal recessive condition common in people with family origins in areas of West Africa, the Mediterranean basin, and parts of the Middle East and India. It is caused by a single amino acid substitution in the -globin protein, as a consequence of a single-nucleotide substitution in the -globin gene. The single-nucleotide substitution also eliminates a cutting site in the -globin gene for the restriction enzymes *Mst*II and *Cvn*I. As a result, the mutation alters the pattern of restriction fragments seen on Southern blots. These differences in restriction cutting sites are used to prenatally diagnose sickle-cell anemia and to establish the parental genotypes and the genotypes of other family members who may be heterozygous carriers of this condition.

DNA is extracted from tissue samples and digested with *Mst*II. This enzyme cuts three times within a region of the normal -globin gene, producing two small DNA fragments. In the mutant sickle-cell allele, the middle *Mst*II site is destroyed by the mutation, and one large restriction fragment is produced by *Mst*II digestion (**Figure 19–7**). The restriction-enzyme-digested DNA fragments are separated by gel electrophoresis, transferred to a nylon membrane, and visualized by Southern blot hybridization, using a probe from this region. Figure 19–7 shows the results of RFLP analysis for sickle-cell anemia in one family.

Only about 5 to 10 percent of all point mutations can be detected by restriction enzyme analysis because most mutations occur in regions of the genome that do not contain restriction enzyme cutting sites. However, now that many disease-associated mutations are known, geneticists can employ synthetic oligonucleotides to detect these mutations, as described next.

Genetic Testing Using Allele-Specific Oligonucleotides

A common method of genetic testing involves the use of synthetic DNA probes known as **allele-specific oligonucleotides** (**ASOs**). Scientists use these short, single-stranded fragments of DNA to identify alleles that differ by as little as a single nucleotide. In contrast to restriction enzyme analysis, which is limited to cases for which a mutation changes a restriction site, ASOs detect single-nucleotide changes (**single-nucleotide polymorphisms** or **SNPs**), including those that do not affect restriction enzyme cutting sites. As a result, this method offers increased resolution and wider application. The ASO is tagged with a molecule that is either radioactive or fluorescent, to allow for visualization of the ASO hybridized to DNA on the membrane. Under proper conditions, an ASO will hybridize only with its complementary DNA sequence and not with other sequences, even those that vary by as little as a single nucleotide.

Genetic testing using ASOs and PCR analysis is now available to screen for many disorders, such as sickle-cell anemia. **Figure 19–8** shows an example of ASO testing for sickle-cell anemia. This rapid, inexpensive, and accurate technique is used to diagnose a wide range of genetic disorders caused by point mutations. Although highly effective, SNPs can affect probe binding leading to false positive or false negative results that may not reflect a genetic disorder, particularly if precise hybridization conditions are not used. Sometimes DNA sequencing is carried out on amplified gene segments to confirm identification of a mutation.

Because ASO testing often involves PCR, small amounts of DNA can be analyzed. As a result, ASO testing is ideal for **preimplantation genetic diagnosis** (**PGD**). PGD is the genetic analysis of single cells from embryos created by

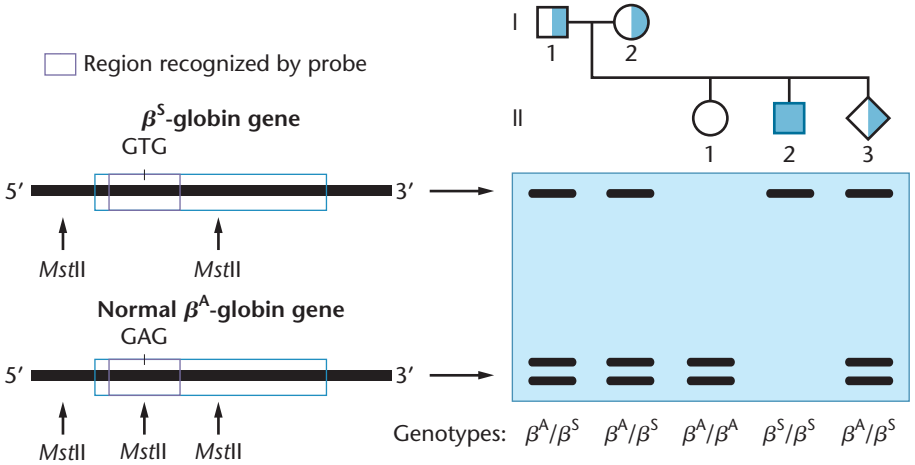


FIGURE 19–7 RFLP diagnosis of sickle-cell anemia. In the mutant -globin allele (^{s}), a point mutation (GAG GTG) has destroyed a cutting site for the restriction enzyme *Mst*II, resulting in a single large fragment on a Southern blot. In the pedigree, the family has one unaffected homozygous normal daughter (II-1), an affected son (II-2), and an unaffected carrier fetus (II-3). The genotype of each family member can be read directly from the blot and is shown below each lane.

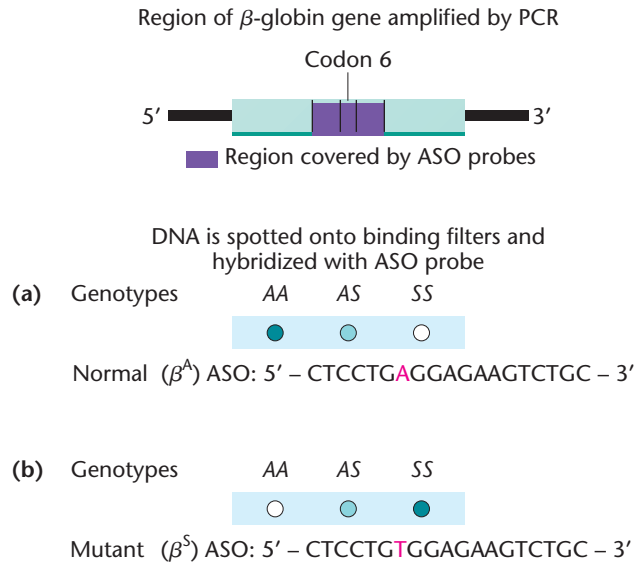


FIGURE 19–8 Allele-specific oligonucleotide (ASO) testing for the -globin gene and sickle-cell anemia. The -globin gene is amplified by PCR, using DNA extracted from white blood cells or cells obtained by amniocentesis. The amplified DNA is then denatured and spotted onto strips of DNA-binding membranes. Each strip is hybridized to a specific ASO. (a) Results observed when the three possible genotypes are hybridized to an ASO from the normal -globin allele: *AA*-homozygous individuals have normal hemoglobin that has two copies of the normal -globin gene and will show heavy hybridization; *AS*-heterozygous individuals carry one normal -globin allele and one mutant allele and will show weaker hybridization; *SS*-homozygous sickle-cell individuals carry no normal copy of the -globin gene and will show no hybridization to the ASO probe for the normal -globin allele. (b) Results observed when DNA for the three genotypes are hybridized to the probe for the sickle-cell -globin allele: no hybridization by the *AA* genotype, weak hybridization by the heterozygote (*AS*), and strong hybridization by the homozygous sickle-cell genotype (*SS*).

in vitro fertilization (IVF). When sperm and eggs are mixed to create zygotes, the early-stage embryos are grown in culture. A single cell can be removed from an early-stage embryo using a vacuum pipette to gently aspirate one cell away from the embryo. This could possibly kill the embryo, but if it is done correctly the embryo will often continue to divide normally. DNA from the removed cell is then typically analyzed by FISH (for chromosome analysis) or by ASO testing. The genotypes for each cell can then be used to decide which embryos will be implanted into the uterus.

Any alleles that can be detected by ASO testing can be used for PGD. Sickle-cell anemia, cystic fibrosis, and dwarfism are often tested for by PGD, but alleles for many other conditions are often analyzed. As you will learn in Section 19.6, it is now becoming possible to carry out whole-genome sequencing on individual cells. This method is now being applied for PGD of single cells from an embryo created by IVF.

NOW SOLVE THIS

19–2 The DNA sequence surrounding the site of the sickle-cell mutation in the -globin gene, for normal and mutant genes, is as follows.

Each type of DNA is denatured into single strands and applied to a DNA-binding membrane. The membrane containing the two spots is hybridized to an ASO of the sequence

5′-GACTCCTGAGGAGAAGT-3′

Which spot, if either, will hybridize to this probe?

5′-GACTCCTGAGGAGAAGT-3′
3′-CTGAGGACTCCTCTTCA-5′
Normal DNA

5′-GACTCCTGTGGAGAAGT-3′
3′-CTGAGGACACCTCTTCA-5′
Sickle-cell DNA

HINT: *This problem asks you to analyze results of an ASO test. The key to its solution is to understand that ASO analysis is done under conditions that allow only identical nucleotide sequences to hybridize to the ASO on the membrane.*

For more practice see Problems 8 and 21.

Genetic Analysis Using Gene-Expression Microarrays

Both RFLP and ASO analyses are efficient methods of screening for gene mutations; however, they can only detect the presence of one or a few specific mutations whose identity and locations in the gene are known. There is also a need for genetic tests that detect complex previously unknown mutations in genes associated with genetic diseases and cancers, mutations that may be associated with, or predispose, a patient to a particular disease, and mRNA expression patterns for genes associated with specific diseases. To analyze multiple genes or mRNA transcripts by genetic tests oftens requires comprehensive, high-throughput methods.

From an earlier chapter (see Chapter 18) recall that one high-throughput screening technique is based on the use of **DNA microarrays** (**Figure 19–9**). (also called DNA chips or gene chips; **Figure 19–10**). The numbers and types of single-stranded DNA sequences on a microarray are dictated by the type of analysis that is required. For example, each field on a microarray might contain a DNA sequence derived from each member of a gene family, or sequence variants from one or several genes of interest, or a sequence derived from each gene in an organism's genome.

FIGURE 19–9 A commercially available DNA microarray, called a GeneChip, marketed by Affymetrix, Inc. This microarray can be used to analyze expression for approximately 50,000 RNA transcripts. It contains 22 different probes for each transcript and allows scientists to simultaneously assess the expression levels of most of the genes in the human genome.

What makes DNA microarrays so amazing is the immense amount of information that can be simultaneously generated from a single array. DNA microarrays the size of postage stamps (just over 1 cm square) can contain up to 500,000 different fields, each representing a different DNA sequence. Earlier in the text (see Chapter 18), you learned about the use of microarrays for transcriptome analysis. In the recent past DNA microarrays have had a wide range of applications, including the detection of mutations in genomic DNA and the detection of gene-expression patterns in diseased tissues. However, in the near future whole-genome sequencing, exome sequencing, and RNA sequencing are expected to replace most applications involving microarrays and render this technology obsolete.

Human genome microarrays containing probes for most human genes are available. DNA microarrays have been designed to scan for mutations in many disease-related genes, including the *p53* gene, which is mutated in a majority of human cancers, and the *BRCA1* gene, which, when mutated, predisposes women to breast cancer.

In addition to testing for mutations in single genes, DNA microarrays can contain probes that detect SNPs. SNPs occur randomly about every 100 to 300 nucleotides throughout the human genome, both inside and outside of genes. SNPs crop up in an estimated 15 million positions in the genome where these single-based changes reveal differences from one person to the next. Certain SNP sequences at a specific locus are shared by certain segments of the population. In addition, certain SNPs cosegregate with genes associated with some disease conditions. By correlating the presence or absence of a particular SNP with a genetic disease, scientists are able to use the SNP as a genetic testing marker.

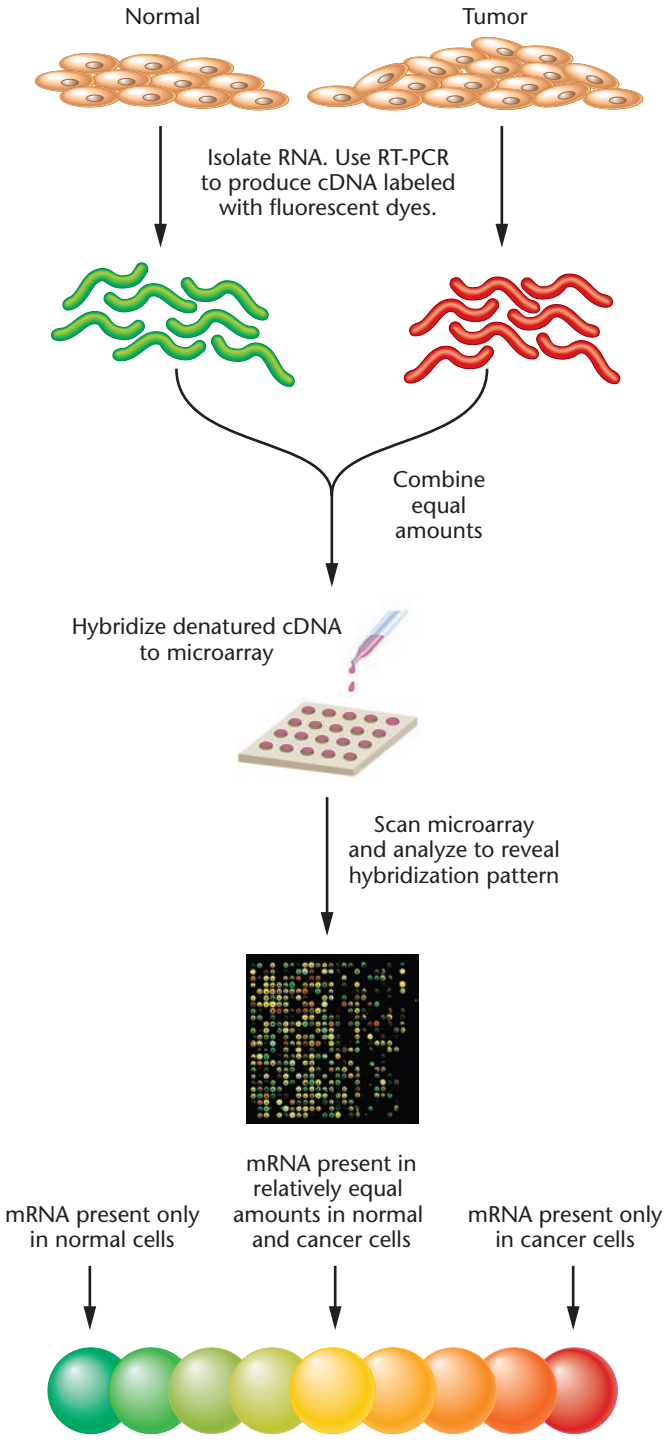


FIGURE 19–10 Microarray procedure for analyzing gene expression in normal and cancer cells. The method shown here is based on a two-channel microarray in which cDNA samples from the two different tissues are competing for binding to the same probe sets. Colors of dots on an expression microarray represent levels of gene expression. In this example, green dots represent genes expressed only in one cell type (e.g., normal cells), and red dots represent genes expressed only in another cell type (e.g., cancer cells). Intermediate colors represent different levels of expression of the same gene in the two cell types. Only a small portion of the DNA microarray is shown.

The presence of SNPs as probes on a DNA microarray allows scientists to simultaneously screen thousands of genes that might be involved in single-gene diseases as well as those involved in disorders exhibiting multifactorial inheritance. This technique, known as **genome scanning,** makes it possible to analyze a person's DNA for dozens or hundreds of disease alleles, including those that might predispose the person to heart attacks, asthma, diabetes, Alzheimer disease, and other genetically defined disease subtypes. Genome scans are occasionally used when physicians encounter patients with chronic illnesses where the underlying cause cannot be diagnosed.

Gene-expression microarrays have been widely used in both basic research and genetic testing for detecting gene-expression patterns for specific genes. Gene-expression microarrays are effective for analyzing gene-expression patterns in genetic diseases because the progression of a tissue from a healthy to a diseased state is almost always accompanied by changes in expression of hundreds to thousands of genes. Because mRNA expression is being detected through gene-expression microarrays, these arrays provide a powerful tool for diagnosing genetic disorders and gene-expression changes. Expression microarrays may contain probes for only a few specific genes thought to be expressed differently in cell types or may contain probes representing each gene in the genome. Although microarray techniques provide novel information about gene expression, keep in mind that DNA microarrays do not directly provide us with information about protein levels in a cell or tissue. We often infer what predicted protein levels may be based on mRNA expression patterns, but this may not always be accurate.

In one type of expression microarray analysis, mRNA is isolated from two different cell or tissue types—for example, normal cells and cancer cells arising from the same cell type (Figure 19–10). The mRNA samples contain transcripts from each gene that is expressed in that cell type. Some genes are expressed more efficiently than others; therefore, each type of mRNA is present at a different level. The level of each mRNA can be used to develop a gene-expression profile that is characteristic of the cell type. Isolated mRNA molecules are converted into cDNA molecules, using reverse transcriptase. The cDNAs from the normal cells are tagged with fluorescent dye-labeled nucleotides (for example, green), and the cDNAs from the cancer cells are tagged with a different fluorescent dye-labeled nucleotide (for example, red).

The labeled cDNAs are mixed together and applied to a DNA microarray. The cDNA molecules bind to complementary single-stranded probes on the microarray but not to other probes. Keep in mind that each field or feature does not consist of just one probe, but rather they contain thousands of copies of the probe. After washing off the nonbinding cDNAs, scientists scan the microarray with a laser, and a computer captures the fluorescent image pattern for analysis. The pattern of hybridization appears as a series of colored dots (often called a "heat" map), with each dot corresponding to one field of the microarray (Figure 19–10). The color patterns revealed on the microarray provide a sensitive measure of the relative levels of each cDNA in the mixture.

Expression microarray profiling has revealed that certain cancers have distinct patterns of gene expression and that these patterns correlate with factors such as the cancer's stage, clinical course, or response to treatment. As this type of analysis has been introduced into clinical use, it has been possible to adjust therapies for each group of cancer patients and to identify new specific treatments based on gene-expression profiles. Similar gene-expression profiles have been generated for many other cancers, including breast, prostate, ovarian, and colon cancer. Gene-expression microarrays are providing tremendous insight into both substantial and subtle variations in genetic diseases.

Several companies are now promoting **nutrigenomics** services in which they claim to use genotyping and gene-expression microarrays to identify allele polymorphisms and gene-expression patterns for genes involved in metabolism. For example, polymorphisms in genes such apolipoprotein A (*APOA1*), involved in lipid metabolism, and *MTHFR* (methylenetetrahydrofolate reductase), involved in metabolism of folic acid, have been implicated in cardiovascular disease. Nutrigenomics companies claim that microarray analysis of a patient's DNA sample for genes such as these and others enables them to judge whether a patient's allele variations or gene-expression profiles warrant dietary changes to potentially improve health and reduce the risk of diet-related diseases.

Application of Microarrays for Gene Expression and Genotype Analysis of Pathogens

Among their many applications, microarrays have provided infectious disease researchers with powerful new tools for studying pathogens. Genotyping microarrays are being used to identify strains of emergent viruses, such as the Ebola virus, the virus that causes the highly contagious condition called Severe Acute Respiratory Syndrome (SARS), and the H5N1 avian influenza virus, the cause of bird flu, which has killed people in Asia, leading to the slaughter of over 80 million chickens and causing concern about possible pandemic outbreaks.

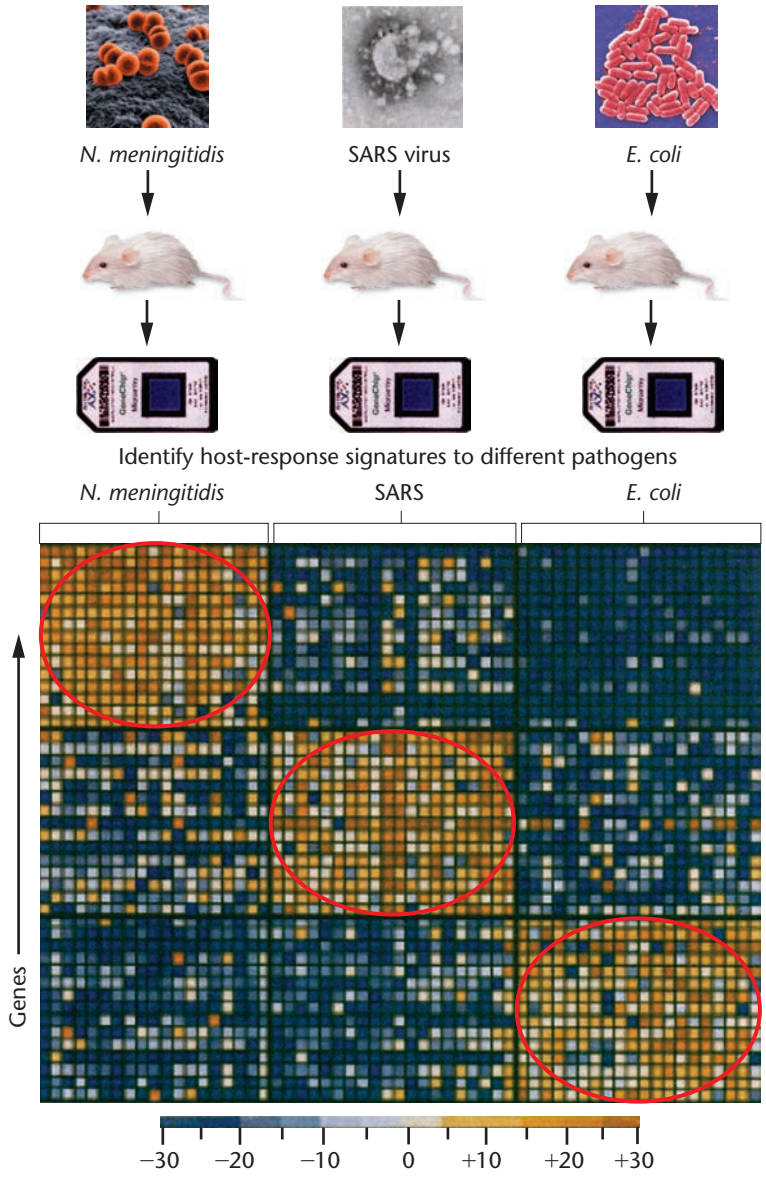


FIGURE 19–11 Gene-expression microarrays can reveal host-response signatures for pathogen identification. In this example, mice were infected with different pathogens: *Neisseria meningitidis,* the virus that causes Severe Acute Respiratory Syndrome (SARS), and *E. coli*. Mouse tissues were then used as the source of mRNA for gene-expression microarray analysis. Increased expression compared to uninfected control mice is shown in shades of yellow. Decreased expression compared to uninfected controls is indicated in shades of blue. Notice that each pathogen elicits a somewhat different response in terms of which major clusters of host genes are activated by pathogen infection (circles).

Whole-genome transcriptome analysis of pathogens is being used to inform researchers about genes that are important for pathogen infection and replication. In this approach, bacteria, yeast, protists, or viral pathogens are used to infect host cells *in vitro,* and then expression microarrays are used to analyze pathogen gene-expression profiles. Patterns of gene activity during pathogen infection of host cells and replication are useful for identifying pathogens and understanding mechanisms of infection. But of course a primary goal of infectious disease research is to prevent infection. Gene-expression profiling is also a valuable approach for identifying important pathogen genes and the proteins they encode that may prove to be useful targets for subunit vaccine development or for drug treatment strategies to prevent or control infectious disease.

Similarly, researchers are evaluating host responses to pathogens. This type of detection has been accelerated in part by the need to develop pathogen-detection strategies for military and civilian use both for detecting outbreaks of naturally emerging pathogens such as SARS and avian influenza and for potential detection of outbreaks such as anthrax (caused by the bacterium *Bacillus anthracis*) that could be the result of a bioterrorism event. Host-response gene-expression profiles are developed by exposing a host to a pathogen and then using expression microarrays to analyze host gene-expression patterns.

Figure 19–11 shows the different gene-expression profiles for mice following exposure to *Neisseria meningitidis,* the SARS virus, or *E. coli.* Comparing such host gene-expression profiles following exposure to different pathogens provides researchers with a way to quickly diagnose and classify infectious diseases. Scientists are developing databases of both pathogen and host-response expression profile data that can be used to identify pathogens efficiently. This type of approach was used to identify host genome-wide responses to Ebola virus infection of nonhuman primates. Results from these analyses helped researchers develop therapeutic drugs for combating the Ebola virus outbreak in western Africa during 2014.

ESSENTIAL POINT

A variety of different molecular techniques, including restriction fragment length polymorphism analysis, allele-specific oligonucleotide tests, and DNA microarrays, can be used to identify genotypes associated with normal and diseased phenotypes. ■

19.6 Genetic Analysis by Individual Genome Sequencing

The ability to sequence and analyze individual genomes is rapidly changing the ways that scientists and physicians evaluate a person's genetic information. Genome

sequencing is being utilized in medical clinics at an accelerated rate. Many major hospitals around the world are setting up clinical sequencing facilities for use in identifying the causes of rare diseases.

Recently, whole-genome sequencing has provided new insights into the genetics of anorexia, Alzheimer disease, autism, Proteus syndrome, and other diseases. Proteus syndrome, a rare congenital disorder that causes atypical bone development, tumors, and other conditions, was the subject of the acclaimed movie *Elephant Man*. Already there have been some very exciting success stories whereby whole-genome sequencing of an individual's DNA has led to improved treatment of diseases in children and adults.

For example, through individual genome sequencing, researchers in Newfoundland identified a mutation in a gene called *AVRD5*. Newfoundlanders have one of the highest incidences in the world of a rare condition called *arrhythmogenic right ventricular cardiomyopathy* (*ARVC*), a condition in which affected individuals often have no symptoms but then die suddenly from irregular electrical impulses within the heart. Mutations of the *AVRD5* gene lead to such cases of premature death through ARVC. Approximately 50 percent of males and 5 percent of females die by age 40, and 80 percent of males and 20 percent of females die by age 50. Individuals carrying this mutation are now being implanted with internal cardiac defibrillators that can be used to restart their hearts if electrical impulses stop or become irregular.

Recall our introduction of the concept of **exome sequencing** (see Chapter 18). Exome sequencing in clinical settings is now producing some promising results. For example, from the time he was born, Nicholas Volker had to live with unimaginable discomfort from an undiagnosed condition that was causing intestinal fistulas (holes from his gut to outside of his body) that were leaking body fluids and feces and requiring constant surgery. By 3 years of age, Nicholas had been to the operating room more than 100 times. A team at the Medical College of Wisconsin decided to have Nicholas's exome sequenced. Applying bioinformatics to compare his sequence to that of the general population, they identified a mutation in a gene on the X chromosome called *X-linked inhibitor of apoptosis* (*XIAP*). *XIAP* is known to be linked to another condition that can often be corrected by a bone marrow transplant. In 2010 a bone marrow transplant saved Nicholas's life and largely restored his health. Shortly thereafter the popular press was describing Nicholas as the first child saved by DNA sequencing.

We now have the ability to sequence the genome from a *single cell*! Single-cell sequencing typically involves isolating DNA from a single cell and then executing *whole-genome amplification (WGA)* to produce sufficient DNA to be sequenced. Reliable amplification of the genome to produce enough DNA for sequencing without introducing errors remains a major challenge.

Single-cell sequencing allows scientists to explore how the genome varies from cell to cell. These studies are revealing that different mutant genes can vary greatly between individual cells. Cancer cells from a tumor in particular often show genetic diversity that is only recently being appreciated by researchers and clinicians. Understanding variations in gene mutations and expression by individual cells within a tumor could lead to better treatment choices.

Genome sequencing is quickly becoming standard practice in medical clinics, and there is every reason to expect genome-based diagnosis to become an essential part of mainstream medicine in the future.

19.7 Genome-Wide Association Studies Identify Genome Variations That Contribute to Disease

Microarray-based genomic analysis has led geneticists to employ powerful strategies called **genome-wide association studies (GWAS)** in their quest to identify genes that may influence disease risk. During the past 8 years there has been a dramatic expansion in the number of GWAS being reported. For example, GWAS for height differences, autism, obesity, diabetes, macular degeneration, myocardial infarction, arthritis, hypertension, several cancers, bipolar disease, autoimmune diseases, Crohn disease, schizophrenia, amyotrophic lateral sclerosis, multiple sclerosis, and behavioral traits (such as intelligence) are among the many GWAS that have been widely publicized in the scientific literature and popular press.

In a GWAS, the genomes of thousands of unrelated individuals with a particular disease are analyzed, typically by microarray analysis, and results are compared with genomes of individuals without the disease as an attempt to identify genetic variations that may confer risk of developing the disease. Many GWAS involve large-scale use of SNP microarrays that can probe on the order of 500,000 SNPs to evaluate results from different individuals. Other GWAS approaches can look for specific gene differences or evaluate CNVs or changes in the epigenome, such as methylation patterns in particular regions of a chromosome. By determining which SNPs, CNVs, or epigenome changes co-occur in individuals with the disease, scientists can calculate the disease risk associated with each variation. Analysis of GWAS results requires statistical analysis to predict the relative potential impact

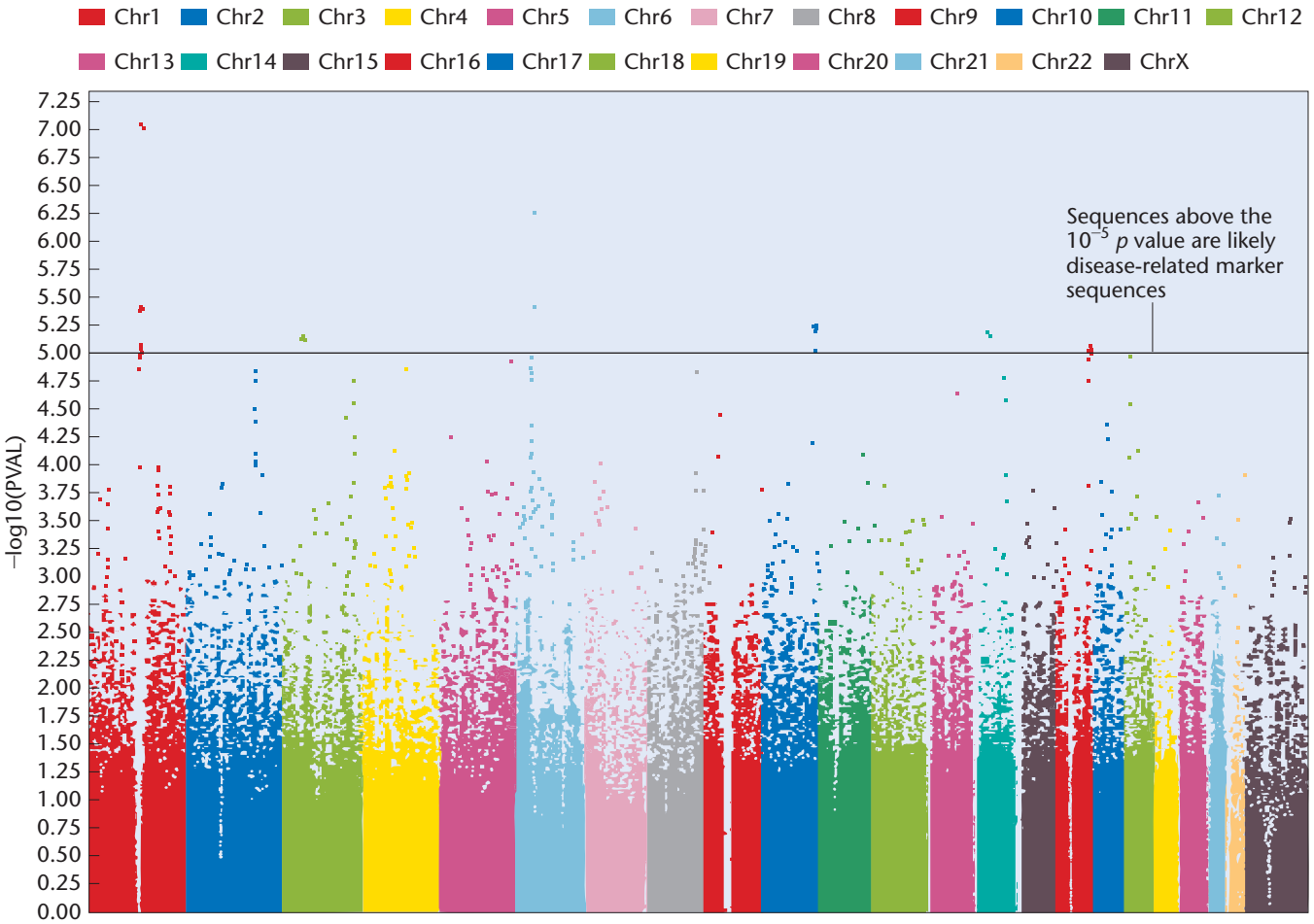


FIGURE 19–12 A GWAS study for Type 2 diabetes revealed 386,371 genetic markers, clustered here by chromosome number. Markers above the black line appeared to be significantly associated with the disease.

(association or risk) of a particular genetic variation on development of a disease phenotype.

Figure 19–12 shows a typical representation of one way that results from GWAS are commonly reported. Called a Manhattan plot, such representations are "scatterplots" that are used to display data with a large number of data points. The *x*-axis typically plots a particular position in the genome; in this case loci on each chromosome are plotted in a different color code. The *y*-axis plots results of a genotypic association test. There are several ways that association can be calculated. Shown here is a negative log of *p* values that shows loci determined to be significantly associated with a particular condition. The top line of this plot establishes a threshold value for significance. Marker sequences with significance levels exceeding 10^{-5}, corresponding to 5.0 on the *y*-axis, are likely disease-related sequences (Figure 19–12).

There are many questions and ethical concerns about patients involved in GWAS and their emotional responses to knowing about genetic risk data. For example,

- What does it mean if an individual has 3, 5, 9, or 30 risk alleles for a particular condition?
- How do we categorize rare, common, and low-frequency risk alleles to determine the overall risk for developing a disease?
- GWAS often reveal dozens of DNA variations, but many variations have only a modest effect on risk. How does one explain to a person that he or she has a gene variation that changes a risk difference for a particular disease from 12 to 16 percent over an individual's lifetime? What does this information mean?
- If the sum total of GWAS for a particular condition reveals about 50 percent of the risk alleles, what are the other missing elements of heritability that may contribute to developing a complex disease?

In some cases, risk data revealed by GWAS may help patients and physicians develop diet and exercise plans designed to minimize the potential for developing a particular disease. But the number of risk genes identified by most GWAS is showing us that, unlike single-gene disorders, complex genetic disease conditions involve a multitude of genetic factors contributing to the total risk for developing a condition. We need such information

to make meaningful progress in disease diagnosis and treatment, which is ultimately a major purpose of GWAS GWAS is another technique that is likely to be replaced by genome sequencing and RNA-seq in the future.

ESSENTIAL POINT

Genome-wide association studies can reveal genetic variations linked with disease conditions within populations. ■

19.8 Genomics Leads to New, More Targeted Medical Treatment Including Personalized Medicine

Genomic technologies are changing medical diagnosis and allowing scientists to manufacture abundant and effective therapeutic proteins. The examples already available today are a strong indication that in the near future, we will see even more transformative medical treatments based on genomics and advanced DNA-based technologies. In this section, we provide brief introductions to pharmacogenomics, rational drug design, and gene therapy, topics that will be considered in greater detail later in the text (see Special Topic Chapter 4—Genomics and Personalized Medicine and Special Topic Chapter 6—Gene Therapy).

Pharmacogenomics and Rational Drug Design

Every year, more than 2 million Americans experience serious side effects of medications, and more than 100,000 die from adverse drug reactions. Until recently, the selection of effective medications for each individual has been a random, trial-and-error process. The new field of **pharmacogenomics** promises to lead to more specific, effective, and personally customized drugs that are designed to complement each person's individual genetic makeup.

In the 1950s, scientists discovered that individual reactions to drugs had a hereditary component. We now know that many genes affect how different individuals react to drugs. Some of these genes encode products such as cell-surface receptors that bind a drug and allow it to enter a cell, as well as enzymes that metabolize drugs. For example, liver enzymes encoded by the cytochrome *P450* gene family affect the metabolism of many modern drugs, including those used to treat cardiovascular and neurological conditions. DNA sequence variations in these genes result in enzymes with different abilities to metabolize and utilize these drugs. Thus, gene variants that encode inactive forms of the cytochrome P450 enzymes are associated with a patient's inability to break down drugs in the body, leading to drug overdoses. A genetic test that recognizes some of these variants is currently being used to screen patients who are recruited into clinical trials for new drugs.

Knowledge from genetics and molecular biology is also contributing to the development of new drugs targeted at specific disease-associated molecules. Most drug development is currently based on trial-and-error testing of chemicals in lab animals, in the hope of finding a chemical that has a useful effect. In contrast, **rational drug design** involves the synthesis of specific chemical substances that affect specific gene products. An example of a rational drug design product is the drug imatinib, trade name **Gleevec,** used to treat chronic myelogenous leukemia (CML). Geneticists had discovered that CML cells contain the Philadelphia chromosome, which results from a reciprocal translocation between chromosomes 9 and 22. Gene cloning revealed that the t(9;22) translocation creates a fusion of the *C-ABL* proto-oncogene with the *BCR* gene. This *BCR-ABL* fusion gene encodes a powerful fusion protein that causes cells to escape cell-cycle control. The fusion protein, which acts as a tyrosine kinase, is not present in noncancer cells from CML patients.

To develop Gleevec, chemists used high-throughput screens of chemical libraries to find a molecule that bound to the BCR-ABL enzyme. After chemical modifications to make the inhibitory molecule bind more tightly, tests showed that it specifically inhibited BCR-ABL activity. Clinical trials revealed that Gleevec was effective against CML, with minimal side effects and a higher remission rate than that seen with conventional therapies. Gleevec is now used to treat CML and several other cancers. With scientists discovering more genes and gene products associated with diseases, rational drug design promises to become a powerful technology within the next decade.

Gene Therapy

Although drug treatments are often effective in controlling symptoms of genetic disorders, the ideal outcome of medical treatment is to cure these diseases. In an effort to cure genetic diseases, scientists are actively investigating **gene therapy**—a therapeutic technique that aims to transfer normal genes into a patient's cells. In theory, the normal genes will be transcribed and translated into functional gene products, which, in turn, will bring about a normal phenotype.

In many ways, gene therapy is the ultimate application of recombinant technology and genomics. Although there have been some successful applications of gene therapy, it has proven to be technically challenging for many reasons.

ESSENTIAL POINT

Pharmacogenomics and gene therapy apply an understanding of the role of genes to develop customized treatments for genetic disorders. ■

19.9 Genetic Engineering, Genomics, and Biotechnology Create Ethical, Social, and Legal Questions

Geneticists use recombinant DNA and genomic technologies to identify genes, diagnose and treat genetic disorders, produce commercial and pharmaceutical products, and solve crimes. However, the applications that arise from these technologies raise important ethical, social, and legal issues that must be identified, debated, and resolved. Here we present a brief overview of some current ethical debates concerning the uses of genetic technologies.

Genetic Testing and Ethical Dilemmas

When the Human Genome Project was first discussed, scientists and the general public raised concerns about how genome information would be used and how the interests of both individuals and society can be protected. To address these concerns, the **Ethical, Legal, and Social Implications (ELSI) Program** was established. The ELSI Program considers a range of issues, including the impact of genetic information on individuals, the privacy and confidentiality of genetic information, and implications for medical practice, genetic counseling, and reproductive decision making. Through research grants, workshops, and public forums, ELSI is formulating policy options to address these issues.

When the Human Genome Project started, ELSI focused on four areas in its deliberations concerning these various issues: (1) privacy and fairness in the use and interpretation of genetic information, (2) ways to transfer genetic knowledge from the research laboratory to clinical practice, (3) ways to ensure that participants in genetic research know and understand the potential risks and benefits of their participation and give informed consent, and (4) public and professional education. The ELSI Program continues to work on a broad range of ethical, societal and policy issues.

A majority of the most widely applied genetic tests that have been used to date have provided patients and physicians with information that improve quality of life. One example involves prenatal testing for phenylketonuria (PKU) and implementing dietary restrictions to diminish the effects of the disease. But many of the potential benefits and consequences of genetic testing are not always clear. For example,

- We have the technologies to test for genetic diseases for which there are no effective treatments. *Should we test people for these disorders?*
- With current genetic tests, a negative result does not necessarily rule out future development of a disease; nor does a positive result always mean that an individual will get the disease. *How can we effectively communicate the results of testing and the actual risks to those being tested?*
- *What information should people have before deciding to have a genome scan or a genetic test for a single disorder or have their whole genome sequenced?*
- Sequencing fetal genomes from the maternal bloodstream has revealed examples of mutations in the fetal genome (for example, a gene involved in Parkinson disease). *How might parents and physicians use this information?*
- *How can we protect the information revealed by such tests?*
- Since sharing patient data through electronic medical records is a significant concern, *what issues of consent need to be considered?*
- *How can we define and prevent genetic discrimination?*

Let's explore a specific example. In 2011, a case in Boston revealed the dangers of misleading results based on genetic testing. A prenatal ultrasound of a pregnant woman revealed a potentially debilitating problem (Noonan syndrome) involving the spinal cord of the woman's developing fetus. Physicians ordered a DNA test, which came back positive for a gene variant in a database that listed the gene as implicated in Noonan syndrome. The parents chose to terminate the pregnancy. Months later it was learned that the locus linked to Noonan was not involved in the disease, yet there was no effective way to inform the research community.

To minimize these kinds of problems in the future, the National Institutes of Health (NIH) National Center for Biotechnology Information (NCBI) has developed a database called ClinVar (see www.ncbi.nlm.nih.gov/clinvar/) which integrates data from clinical genetic testing labs and research literature to provide an updated resource for researchers and physicians.

Disclosure of incidental results is another ethically challenging issue. When someone has his or her genome sequenced or has a test done for a particular locus thought to be involved in a disease condition, the analysis sometimes reveals other mutations that could be of significance to the patient. Researchers and clinicians are divided on whether such information should be disclosed to the patient. What do you think?

Earlier in this chapter we discussed preimplantation genetic diagnosis (PGD), which provides couples with the ability to screen embryos created by *in vitro* fertilization for genetic disorders. As we learn more about genes involved in human traits, will other, nondisease-related genes be screened for by PGD? Will couples be able to select embryos with certain genes encoding desirable traits for height, weight, intellect, and other physical or mental characteristics? What do you think of using genetic testing to purposely select for an embryo with a genetic disorder? There have been several well-publicized

cases of couples seeking to use prenatal diagnosis or PGD to select for embryos with dwarfism and deafness.

As identification of genetic traits becomes more routine in clinical settings, physicians will need to ensure genetic privacy for their patients. There are significant concerns about how genetic information could be used in negative ways by employers, insurance companies, governmental agencies, or the general public. Genetic privacy and prevention of genetic discrimination will be increasingly important in the coming years. In 2008, the **Genetic Information Nondiscrimination Act** was signed into law in the United States. This legislation is designed to prohibit the improper use of genetic information in health insurance and employment.

Direct-to-Consumer Genetic Testing and Regulating the Genetic Test Providers

The past decade has seen dramatic developments in **direct-to-consumer (DTC) genetic tests.** A simple Web search will reveal many companies offering DTC genetic tests. As of 2015, there are over 2000 diseases for which such tests are now available (in 1993 there were about 100 such tests). Most DTC tests require that a person mail a saliva sample, hair sample, or cheek cell swab to the company. For a range of pricing options, DTC companies largely use SNP-based tests such as ASO tests to screen for different mutations. For example, in 2007 Myriad Genetics, Inc. began a major DTC marketing campaign of its tests for *BRCA1* and *BRCA2,* which have been available since 1996. Mutations in these genes increase risk of developing breast and ovarian cancer. DTC testing companies report absolute risk, the probability that an individual will develop a disease, but how such risks results are calculated is highly variable and subject to certain assumptions.

Such tests are controversial for many reasons. For example, the test is purchased online by individual consumers and requires no involvement of a physician or other health-care professionals such as a nurse or genetic counselor to administer or to interpret results. There are significant questions about the quality, effectiveness, and accuracy of such products because currently the DTC industry is largely self-regulated. The FDA does not regulate DTC genetic tests. There is at present no comprehensive way for patients to make comparisons and evaluations about the range of tests available and their relative quality.

Most companies make it clear that they are not trying to diagnose or prevent disease, nor that they are offering health advice, so what is the purpose of the information that test results provide to the consumer?

Web sites and online programs from DTC companies provide information on what advice a person should pursue if positive results are obtained. But is this enough? If results are not understood, might negative tests not provide a false sense of security? Just because a woman is negative for *BRCA1* and *BRCA2* mutations *does not* mean that she cannot develop breast or ovarian cancer. Refer to Problem 15 for an example of a personal decision that actress Angelina Jolie made based on the results of a genetic test.

In 2010, the FDA announced that five genetic test manufacturers would need FDA approval before their tests could be sold to consumers. This action was prompted when Pathway Genomics announced plans to market a DTC kit for "comprehensive genotyping" in the pharmacy chain Walgreens. Pathway Genomics and the other companies had been selling their DTCs through company Web sites for several years. Pathway and others claimed that because their DTC kits are Clinical Laboratory Improvement Amendments (CLIA) approved that no further regulation was required. CLIA regulates certain laboratory tests but is not part of the FDA. Pathway and de-CODE dropped out of the DTC market.

Whether the FDA will oversee DTC genetic tests in the future is unclear. At the time of publication of this edition, the FDA continues to work on plans for regulation of DTC genetic tests. But because some DTC genetic testing companies, such as 23andMe, offer health-related analyses or health reports, they do fall under FDA regulations. The FDA continues to issue warnings to DTC companies to provide what the FDA considers to be appropriate health-related interpretations of genetic tests. There are varying opinions on the regulatory issue. Some believe that the FDA has no business regulating DTC tests and that consumers should be free to purchase products according to their own needs or interests. Others insist that the FDA must regulate DTCs in the interest of protecting consumers.

The National Institutes of Health created the **Genetic Testing Registry (GTR)**, designed to increase transparency by allowing companies to publicly share information about the utility of their genetic testing products, research for the general public, patients, health-care workers, genetic counselors, insurance companies, and others. The GTR is intended to allow individuals and families access to key resources to make more well-informed decisions about their health and genetic tests. But participation in the GTR by DTC companies has not been made mandatory yet, so will companies involved in genetic testing participate?

DNA and Gene Patents

Intellectual property (IP) rights are also being debated as an aspect of the ethical implications of genetic engineering, genomics, and biotechnology. Patents on intellectual property (isolated genes, new gene constructs, recombinant cell types, GMOs) can be potentially lucrative for the patent-holders but may also pose ethical and scientific problems. Why is protecting IP important for companies? Consider this issue. If a company is willing to spend millions or billions of dollars and several years doing research and development

(R&D) to produce a valuable product, then shouldn't it be afforded a period of time to protect its discovery so that it can recover R&D costs and make a profit on its product?

Genes in their natural state as products of nature cannot be patented. Consider the possibilities for a human gene that has been cloned and then patented by the scientists who did the cloning. The person or company holding the patent could require that anyone attempting to do research with the patented gene pay a licensing fee for its use. Should a diagnostic test or therapy result from the research, more fees and royalties may be demanded, and as a result the costs of a genetic test may be too high for many patients to afford. But limiting or preventing the holding of patents for genes or genetic tools could reduce the incentive for pursuing the research that produces such genes and tools, especially for companies that need to profit from their research. Should scientists and companies be allowed to patent DNA sequences from naturally living organisms? And should there be a lower or an upper limit to the size of those sequences? For example, should patents be awarded for small pieces of genes, such as expressed sequence tags (ESTs), just because some individual or company wants to claim a stake in having cloned a piece of DNA first, even if no one knows whether the DNA sequence has a use? Can or should investigators be allowed to patent the entire genome of any organism they have sequenced?

As of 2014, the U.S. Patent and Trademark Office has granted patents for more than 35,000 genes or gene sequences, including an estimated 20 percent of human genes. Incidentally the patenting of human genes has led some to use the term *patentome*! Some scientists are concerned that to award a patent for simply cloning a piece of DNA is awarding a patent for too little work. Given that computers do most of the routine work of genome sequencing, who should get the patent? What about individuals who figure out *what* to do with the gene? What if a gene sequence has a role in a disease for which a genetic therapy may be developed? Many scientists believe that it is more appropriate to patent novel technology and applications that make use of gene sequences than to patent the gene sequences themselves.

In recent years Congress has been considering legislation that would ban the patenting of human genes and any sequences, functions, or correlations to naturally occurring products from a gene. The patenting of genetic tests is also under increased scrutiny in part because of concerns that a patented test can create monopolies in which patients cannot get a second opinion if only one company holds the rights to conduct a particular genetic test. Recent analysis has estimated that as many as 64 percent of patented tests for disease genes make it very difficult or impossible for other groups to propose a different way to test for the same disease.

In 2010 a landmark case brought by the American Civil Liberties Union against Myriad Genetics contended that Myriad could not patent the *BRCA1* and *BRCA2* sequences used to diagnose breast cancer. Myriad's BRACAnalysis product has been used to screen over a million women for *BRCA* 1 and 2 during its period of patent exclusivity. A U.S. District Court judge ruled Myriad's patents invalid on the basis that DNA in an isolated form is not fundamentally different from how it exists in the body. Myriad was essentially accused of having a monopoly on its tests, which have existed for a little over a decade based on its exclusive licenses in the United States.

This case went to the Supreme Court in 2013, which rendered a 9–0 ruling against Myriad, stripping it of five of its patent claims for the *BRCA1* and *BRCA2* genes, largely based on the view that natural genes are a product of nature and just because they are isolated does not mean they can be patented. The Court ruled that cDNA sequences produced in a lab can continue to be patentable. Myriad still holds about 500 valid claims related to *BRCA* gene testing.

Whole-Genome Sequence Analysis Presents Many Questions of Ethics

To date, the majority of genetic testing applications have involved approaches such as amniocentesis and chorionic villus sampling for identifying chromosomal defects and testing for individual genes through methods such as ASO analysis. Even microarray analysis has not been widely used for genetic testing. But in the next decade and beyond, it is expected that whole-genome sequencing of adults and babies will increasingly become common in clinical settings. A Genomic Sequencing and Newborn Screening Disorders Program is underway to sequence the exomes of more than 1500 babies. Both infants with illnesses and babies who are healthy will be part of this screening program. This will allow scientists to carry out comparative genomic analyses of specific sequences to help identify genes involved in disease conditions.

Screening of newborns is important to help prevent or minimize the impacts of certain disorders. Earlier in this section we mentioned the positive impact of early screening for PKU. Each year routine blood tests from a heel prick of newborn babies reveal rare genetic conditions in several thousand infants in the United States alone. A small number of states allow parents to opt out of newborn testing. In the future, should DNA sequencing at the time of birth be required? Do we really know enough about which human genes are involved in disease to help prevent disease in children? Estimates suggest that sequencing can identify approximately 15 to 50 percent of children with diseases that currently cannot be diagnosed by other methods. What is the value of having sequencing data for healthy children?

As exciting as this period of human genetics and medicine is becoming, many of the whole-genome sequencing studies of individuals are happening in a largely unregulated environment, especially with respect to DNA collection, the

variability and quality control of DNA handling protocols, sequence analysis, storage, and confidentiality of genetic information (see the Genetics, Technology, and Society box below), which is raising significant ethical concerns.

Preconception Testing, Destiny Predictions, and Baby-Predicting Patents

Companies are now promoting the ability to do *preconception* testing and thus make "destiny predictions" about the potential phenotypes of hypothetical offspring based on computation methods for analyzing sequence data of parental DNA samples. The company 23andMe has been awarded a U.S. patent for a computational method called the *Family Traits Inheritance Calculator* to use parental DNA samples to predict a baby's traits, including eye color and the risk of certain diseases. This patent includes applications of technologies to screen sperm and ova for *in vitro fertilization* (IVF).

Currently, gender selection of embryos generated by IVF is very common. But could preconception testing lead to the selection of "designer babies"? Fear of *eugenics* surrounds these conversations, particularly as genetic analysis starts moving away from disease conditions to nonmedical traits such as hair color, eye color, other physical traits, and potentially behavioral traits. The patent has been awarded for a process that will compare the genotypic data of an egg provider and a sperm provider to suggest gamete donors that might result in a baby or hypothetical offspring with particular phenotypes of interest to a prospective parent. What do you think about this?

A company called GenePeeks claims to have a patent-pending technology for reducing the risk of inherited disorders by "digitally weaving" together the DNA of prospective parents. GenePeeks plans to sequence the DNA of sperm donors and women who want to get pregnant to inform women about donors who are most genetically compatible for the traits they seek in offspring. Their proprietary computing technology is intended to use sequence data to examine virtual progeny from donor-client pairings to estimate the likelihood of particular diseases from prospective parents. Initially, the company claims that it plans to focus on looking at around 100,000 loci involved in rare disease. Will technologies such as this become widespread and attract consumer demand in the future? What do you think? Would you want this analysis done before deciding whether to have a child with a particular person?

Patents and Synthetic Biology

The J. Craig Venter Institute (JCVI) has filed two patent applications for what is being called "the world's first-ever human-made life form." The patents are intended to cover the minimal genome of *M. genitalium,* which JCVI believes are the genes essential for self-replication. One of these patent applications is designed to claim the rights to synthetically constructed organisms. Another U.S. patent issued to a different group of researchers covers application of a minimal genome for *E. coli,* which has generated even more concern given its relative importance compared with *M. genitalium.* What do you think? Should it be possible to patent a minimal genome or a synthetic organism?

Consider these other ethical issues about synthetic biology. Synthetic biology has the potential to be used for harmful purposes (such as bioterrorism). What regulatory policies and restrictions should be placed on applications of synthetic biology and on patents of these applications? The ability to modify life forms offends some people. How will this issue be addressed by the synthetic biology research community?

ESSENTIAL POINT

Applications of genetic engineering and biotechnology involve a wide range of ethical, legal, and social dilemmas with important scientific and societal implications. ■

GENETICS, TECHNOLOGY, AND SOCIETY

Privacy and Anonymity in the Era of Genomic Big Data

Our lives are surrounded by Big Data. Enormous quantities of personal information are stored on private and public databases, revealing our purchasing preferences, search engine histories, social contacts, and even GPS locations. We allow this information to be collected in exchange for services that we perceive as valuable, or at least innocuous. But often we do not know how this information may be used now, or in the future, and we do not consider how its distribution may affect us, our families, and our relationships.

Perhaps the most personal of all Big Data entries are those obtained from personal genome sequences and genomic analyses. Tens of thousands of individuals are now donating DNA for whole-genome sequencing—to be carried out by both private gene-sequencing companies and

(continued)

Genetics, Technology, and Society, continued

public research projects. Most people who donate their DNA for sequence analysis do so with the assumption, or promise, that these data will remain anonymous and private. Even when we are informed that the data will be available to others, we express little concern. After all, what consequences could possibly come from access to gigabytes of anonymous A's, C's, T's, and G's? Surprisingly, the answer is—quite a lot.

One of the first inklings of genetic privacy problems arose in 2005, when a 15-year-old boy named Ryan Kramer tracked down his anonymous sperm-donor father using his own Y chromosome sequence data and the Internet (Motluk, A., *New Scientist* 2524: November 3, 2005). Ryan submitted a DNA sample to a genealogy company that generates Y chromosome profiles, matches them against entries in their database, and puts people into contact with others who share similar genetic profiles, indicating relatedness. Two men contacted Ryan, and both had the same last name, with different spellings. Ryan combined the last names with the only information that he had about his sperm-donor father—date of birth, birth place, and college degree. Using an Internet people-search service, he obtained the names of everyone born on that date in that place. On the list, there was one man with the same last name as his two Y chromosome relatives. Through another Internet search, Ryan confirmed that the man also had the appropriate college degree. He then contacted his sperm-donor father. Since this report, other children of sperm donors have used DNA genealogy searching to find their paternal parent. In these cases, the sperm donors had not submitted their DNA to genealogy companies, but their identities could be determined indirectly. The implications for sperm donors have been unsettling, as most are promised anonymity. In some cases, donors are troubled to learn that they have fathered dozens of offspring.

More recently, several published reports reveal the ease with which anyone's identity can be traced using whole-genome DNA sequences. For example, a research team led by Dr. Yaniv Erlich of the Whitehead Institute described how anonymous entries in the 1000 *Genomes Project* database could be traced and identified (Gymrek, M. et al. 2013. *Science* 339:321–324). To begin, they converted a number of randomly chosen Y chromosome sequences from the database into STR profiles, and then they used the profiles to search two free public genealogy databases. These searches yielded family surnames and other information such as geographical location and pedigrees. In all, the identities of 50 people were revealed starting from the genome sequences of only five individuals.

Studies similar to Erlich's show that a person's identity and other personal information such as age, sex, body mass index, glucose, insulin, lipid levels, and disease susceptibilities can be revealed, starting with anonymous RNA expression-level data or data from SNP genotyping microarrays.

To many people, the implications of "genomic de-identifications" are disturbing. Genomic information leaks could reveal personal medical information, physical appearance, and racial origins. They could also be used to synthesize DNA to plant at a crime scene or could be used in unforeseen ways in the future as we gain more information that resides in our genome. The consequences of genomic information leaks are not limited to the person whose genome was sequenced, but encompass family members from many generations, who share the person's genetic heritage.

It may be time to address questions revolving around genome data privacy and anonymity before it is too late to put the genome-genie back into the bottle.

Your Turn

1. Would you donate your DNA to a research project designed to discover the genetic links to certain cancers? If not, why not? If so, what privacy assurances would you need to make you comfortable about your donation?

Some of the consequences of genome data leaks as they pertain to research projects are outlined in Brenner, S.E. 2013. Be prepared for the big genome leak. *Nature* 498: 139.

2. Genome data privacy issues are being debated by scientists and regulatory agencies. What are some of their concerns about privacy, and what ideas are being proposed to deal with these concerns?

To begin a discussion about privacy, informed consent, and regulations, see Hayden, E.C. 2012. Informed consent: a broken contract. *Nature* 486:312–314.

3. One way to deal with issues of genome data leaks is to make all personal genome data open and accessible. What are the advantages and disadvantages of this approach?

This approach is being taken by the Personal Genome Project. Read about their nonanonymous approach to DNA data at http://www.personalgenomes.org/non-anonymous.

CASE STUDY | Three-parent babies—the ethical debate

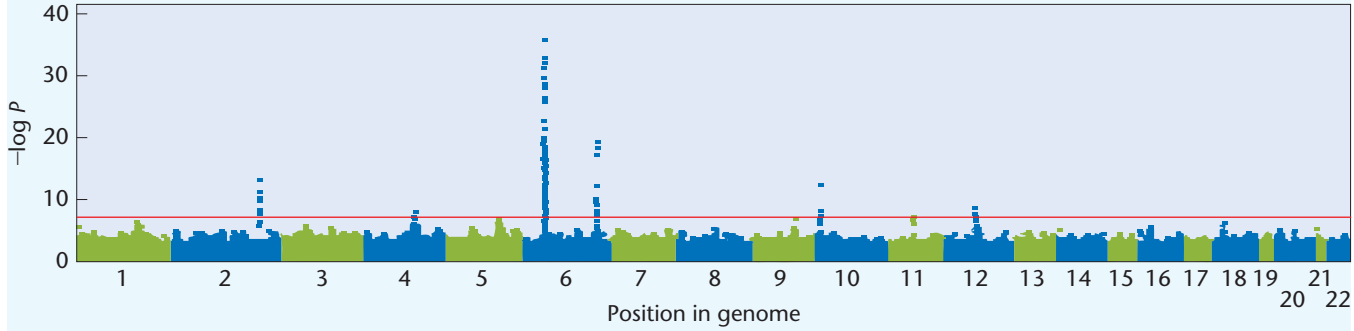
A couple wants to have a baby, but the woman has MERRF (Myoclonic Epilepsy with Ragged Red Fibers) syndrome, caused by a mutation at position 8344 in the mitochondrial genome (mtDNA). Since all mitochondria are inherited solely from the mother, her baby would have her disease. They hear of a new technique called mitochondrial manipulation technology, where the mother's nucleus is transferred to a donor egg and then fertilized with the father's sperm. There are some concerns about the safety and efficiency of the technique, the impacts on the mental well-being of the child, and also about the identity of the child as he or she has three parents. There are also profound ethical implications because the technique introduces permanent changes into the germline.

1. What would you advise the couple to do?
2. Describe the pros and cons of the mitochondrial manipulation technology.
3. The British government legalized mitochondrial manipulation in February 2015, the first government to do so. Should more countries do the same? Justify your answer.

INSIGHTS AND SOLUTIONS

1 Research by Petukhova *et al.* (Nature 466:113–117, 2010) involved a GWAS to analyze 1054 cases of patients with alopecia areata (AA) and 3278 controls. Alopecia areata is a condition that leads to major hair loss and affects approximately 5.3 million people in the United States alone.

(a) A Manhattan plot from this work is shown below:



Based on your interpretation of this plot, which chromosomes were associated with loci that may contribute to AA?

(b) Of the 139 SNPs significantly associated with AA, several genes are involved in controlling the activation and proliferation of regulatory T lymphocytes (Treg cells) and cytotoxic T lymphocytes, genes involved in antigen presentation to immune cells, immune regulatory molecules such as the interleukins, and genes expressed in the hair follicle itself. Speculate how these candidate genes may help scientists understand how AA progresses as a disease.

Solution: (a) Investigators identified eight genomic regions with SNPs that exceed the genome-wide significance value of (red line). These regions were clustered on chromosomes 2, 4, 6, 9, 10, 11, and 12.

(b) AA is an autoimmune disease in which the immune system attacks hair follicles, resulting in hair loss that can permeate across the entire scalp and even the whole body. AA hair follicles are attacked by T cells. The identification of candidate genes involved in T-cell proliferation, immune system regulation, and follicular development may potentially help investigators develop cures for AA.

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we focused on a number of interesting applications of genetic engineering, genomics, and biotechnology. At the same time, we found many opportunities to consider the methods and reasoning by which much of this information was acquired. From the explanations given in the chapter, what answers would you propose to the following fundamental questions:
(a) What experimental evidence confirms that we have introduced a useful gene into a transgenic organism and that it performs as we anticipate?
(b) How can we use DNA analysis to determine that a human fetus has sickle-cell anemia?
(c) How can DNA microarray analysis be used to identify specific genes that are being expressed in a specific tissue?
(d) How are GWAS carried out, and what information do they provide?
(e) What are some of the technical reasons why gene therapy is difficult to carry out effectively?

CONCEPT QUESTION

2. Review the Chapter Concepts list on page 394. Most of these center on applications of genetic technology that are becoming widespread. Write a short essay that summarizes the impacts that genomic applications are having on society and the ethical issues presented by these applications. ■

3. An unapproved form of gene therapy, known as enhancement gene therapy, can create considerable ethical dilemmas. Why?
4. There are more than 1000 cloned farm animals in the United States. In the near future, milk from cloned cows and their offspring (born naturally) may be available in supermarkets. These cloned animals have not been transgenically modified, and they are no different than identical twins. Should milk from such animals and their natural-born offspring be labeled as coming from cloned cows or their descendants? Why?
5. One of the major causes of sickness, death, and economic loss in the cattle industry is *Mannheimia haemolytica*, which causes bovine pasteurellosis, or shipping fever. Noninvasive delivery of a vaccine using transgenic plants expressing immunogens would reduce labor costs and trauma to livestock. An early step toward developing an edible vaccine is to determine whether an injected version of an antigen (usually a derivative of the pathogen) is capable of stimulating the development of antibodies in a test organism. The following table assesses the ability of a transgenic portion of a toxin (Lkt) of *M. haemolytica* to stimulate development of specific antibodies in rabbits.

Immunogen Injected	Antibody Production in Serum
Lkt50*—saline extract	+
Lkt50—column extract	+
Mock injection	–
Pre-injection	–

*Lkt50 is a smaller derivative of Lkt that lacks all hydrophobic regions. + indicates at least 50 percent neutralization of toxicity of Lkt; – indicates no neutralization activity.

Source: Modified from Lee et al. 2001. *Infect. and Immunity* 69: 5786–5793.

(a) What general conclusion can you draw from the data?
(b) With regards to development of a usable edible vaccine, what work remains to be done?

6. Describe how the team from the J. Craig Venter Institute created a synthetic genome. How did they demonstrate that the genome converted the recipient strain of bacteria into a different strain?
7. Sequencing the human genome and the development of microarray technology promise to improve our understanding of normal and abnormal cell behavior. How are microarrays dramatically changing our understanding of complex diseases such as cancer?
8. A couple with European ancestry seeks genetic counseling before having children because of a history of cystic fibrosis (CF) in the husband's family. ASO testing for CF reveals that the husband is heterozygous for the Δ*508* mutation and that the wife is heterozygous for the *R117* mutation. You are the couple's genetic counselor. When consulting with you, they express their conviction that they are not at risk for having an affected child because they each carry different mutations and cannot have a child who is homozygous for either mutation. What would you say to them?
9. As genetic testing becomes widespread, medical records will contain the results of such testing. Who should have access to this information? Should employers, potential employers, or insurance companies be allowed to have this information? Would you favor or oppose having the government establish and maintain a central database containing the results of individuals' genome scans?
10. What limits the use of differences in restriction enzyme sites as a way of detecting point mutations in human genes?
11. Genes in their natural state cannot be patented. This policy allows research and use of natural products for the common good. What argument might be presented in favor of patenting genes or gene products?
12. What are the different genetic markers that genome-wide association studies (GWAS) employ? How can scientists use this data to calculate the disease risk associated with each variation?
13. The family of a sixth-grade boy in Palo Alto, California, was informed by school administrators that he would have to transfer out of his middle school because they believed his mutation of the *CFTR*, which does not produce any symptoms associated with cystic fibrosis, posed a risk to other students at the school who have cystic fibrosis. After missing 11 days of school, a settlement was reached to have the boy return to school. Based on what you know about GINA, the Genetic Information Nondiscrimination Act, what ethical problems might you associate with this example?
14. Dominant mutations can be categorized according to whether they increase or decrease the overall activity of a gene or gene product. Although a loss-of-function mutation (a mutation that inactivates the gene product) is usually recessive, for some genes, one dose of the normal gene product, encoded by the normal allele, is not sufficient to produce a normal phenotype. In this case, a loss-of-function mutation in the gene will be dominant, and the gene is said to be *haploinsufficient*. A second category of dominant mutation is the gain-of-function mutation, which results in a new activity or increased activity or expression of a gene or gene product. The gene therapy technique currently used in clinical trials involves the "addition" to somatic cells of a normal copy of a gene. In other words, a normal copy of the gene is inserted into the genome of the mutant somatic cell, but the mutated copy of the gene is not removed or replaced. Will this strategy work for either of the two aforementioned types of dominant mutations?
15. The first attempts at gene therapy began in 1990 with the treatment of a young girl with a genetic disorder abbreviated SCID. What does SCID stand for? In the context of SCID, what does ADA stand for?
16. The Genetic Testing Registry is intended to provide better information to patients, but companies involved in genetic testing are not required to participate. Should company participation be mandatory? Why or why not? Explain your answers.
17. Once DNA is separated on a gel, it is often desirable to gain some idea of its informational content. How might this be done?
18. Would you have your genome sequenced, if the price was affordable? Why or why not? If you answered yes, would you make your genome sequence publicly available? How might such information be misused?
19. Following the tragic shooting of 20 children at a school in Newtown, Connecticut, in 2012, Connecticut's state medical examiner requested a full genetic analysis of the killer's genome. What do you think investigators might be looking for? What might they expect to find? Might this analysis lead to oversimplified analysis of the cause of the tragedy?
20. Private companies are now offering personal DNA sequencing along with interpretation. What services do they offer? Do you think that these services should be regulated, and if so, in what way? Investigate one such company, 23andMe, at http://www.23andMe.com.
21. Yeager, M., *et al.* (*Nature Genetics* 39: 645–649, 2007) and Sladek, R. *et al.* (*Nature* 445: 881–885, 2007) have used single-nucleotide polymorphisms (SNPs) in genome-wide association studies (GWAS) to identify novel risk loci for prostate cancer and Type 2 diabetes mellitus, respectively. Each study suggests that disease-risk genes can be identified that significantly contribute to the disease state. Given your understanding of such complex diseases, what would you consider as reasonable factors to consider when interpreting the results of GWAS studies?
22. In March 2010 Judge R. Sweet ruled to invalidate Myriad Genetics' patents on the *BRCA1* and *BRCA2* genes. Sweet wrote that since the genes are part of the natural world, they are not patentable. Myriad Genetics also holds patents on the development of a direct-to-consumer test for the *BRCA1* and *BRCA2* genes.
(a) Would you agree with Judge Sweet's ruling to invalidate the patenting of the *BRCA1* and *BRCA2* genes? If you were asked to judge the patenting of the direct-to-consumer test for the *BRCA1* and *BRCA2* genes, how would you rule?
(b) J. Craig Venter has filed a patent application for his "first-ever human made life form." This patent is designed to cover the genome of *M. genitalium.* Would your ruling for Venter's "organism" be different from Judge Sweet's ruling on patenting of the *BRCA1* and *BRCA2* genes?

20 Developmental Genetics

CHAPTER CONCEPTS

- Gene expression during development is based on the differential transcription of selected genes.
- Animals use a small number of shared signaling systems and regulatory networks to construct a wide range of adult body forms from the zygote. These shared properties make it possible to use animal models to study human development.
- Differentiation is controlled by cascades of gene expression that are a consequence of events that specify and determine the developmental fate of cells.
- Plants independently evolved developmental regulatory mechanisms that parallel those of animals.
- In many organisms, cell–cell signaling systems program the developmental fate of adjacent and distant cells.

This unusual four-winged *Drosophila* has developed an extra set of wings as a result of a mutation in a homeotic selector gene.

Over the last two decades, the use of genetic analysis, molecular biology, and genomics showed that, in spite of wide diversity in the size and shape of adult animals and plants, multicellular organisms share many genetic pathways and molecular signaling mechanisms that control developmental events leading from the zygote to the adult. At the cellular level, development is marked by three important events: **specification,** when the first cues confer spatial identity, **determination,** when a specific developmental fate for a cell becomes fixed, and **differentiation,** the process by which a cell achieves its final adult form and function. Now, thanks to newly developed methods of analysis including microarrays, high-throughput sequencing, epigenetics, proteomics, and systems biology, we are beginning to understand how the action and interaction of genes and environmental factors control developmental processes in multicelluar organisms.

In this chapter, the primary emphasis will be on how genetic analysis has been used to study development. This field, called developmental genetics, laid the foundation for our understanding of developmental events at the molecular and cellular levels, which contribute to the continually changing phenotype of the newly formed organism.

20.1 Differentiated States Develop from Coordinated Programs of Gene Expression

Animal genomes contain tens of thousands of genes, but only a small subset of these control the events that shape the adult body (Figure 18–1).

Developmental geneticists study mutant alleles of these genes to ask important questions about development:

- What genes are expressed?
- When are they expressed?
- In what parts of the developing embryo are they expressed?
- How is the expression of these genes regulated?
- What happens when these genes are defective?

These questions provide a foundation for exploring the molecular basis of developmental processes such as determination, induction, cell–cell communication, and cellular differentiation. Genetic analysis of mutant alleles is used to establish a causal relationship between the presence or absence of inducers, receptors, transcriptional events, cell and tissue interactions, and the observable structural changes that accompany development.

A useful way to define development is to say that it is the attainment of a differentiated state by all the cells of an organism (except for stem cells). For example, a cell in a blastula-stage embryo (when the embryo is just a ball of uniform-looking cells) is undifferentiated, whereas a red blood cell synthesizing hemoglobin in the adult body is differentiated. How do cells get from the undifferentiated to the differentiated state? The process involves progressive activation of different groups of gene sets in different cells of the embryo. From a genetics perspective, one way of defining the different cell types that form during development in multicellular organisms is to identify and catalog the genes that are active in each cell type. In other words, development depends on patterns of differential gene expression.

The idea that differentiation is accomplished by activating and inactivating genes at different times and in different cell types is called the **variable gene activity hypothesis.** Its underlying assumptions are, first, that each cell contains an entire genome and, second, that differential transcription of selected genes controls the development and differentiation of each cell. In multicellular organisms, the genes involved with development have been conserved throughout evolution, along with the patterns of differential transcription, and the ensuing developmental mechanisms. As a result, scientists are able to learn about development in complex multicellular organisms including humans by dissecting these mechanisms in a small number of genetically well-characterized model organisms.

20.2 Evolutionary Conservation of Developmental Mechanisms Can Be Studied Using Model Organisms

Genetic analysis of development across a wide range of organisms has shown that the size and shape of all animal bodies are controlled by a common set of genes and developmental mechanisms. For example, most of the differences in anatomical structures between organisms as diverse as zebras and zebrafish result from different patterns of expression in a single gene set, called the homeotic (abbreviated as *Hox*) genes, and not by expression of a host of species-specific genes. Genome-sequencing projects have confirmed that homeotic genes from a wide range of organisms have a common ancestry; this homology means that many aspects of normal human embryonic development and associated genetic disorders can be studied in model organisms such as *Drosophila,* where genetic methods including mutagenesis, genetic crosses, and large-scale experiments involving hundreds of offspring can be conducted (see Chapter 1 for a discussion of model organisms in genetics).

Studies comparing developmental processes in different organisms (a field called evolutionary developmental biology) have revealed that although many developmental mechanisms are common to all animals, over evolutionary time, several new and unique ways of transforming a zygote into an adult have appeared. These evolutionary changes result from mutation, gene duplication and divergence, assignment of new functions to old genes, and the recruitment of genes to new developmental pathways. However, the emphasis in this chapter will be on the similarities in genes and developmental mechanisms among species.

Analysis of Developmental Mechanisms

In the space of this chapter, we cannot survey all aspects of development, nor can we explore in detail how the array of developmental mechanisms triggered by the fusion of sperm and egg were identified. Instead, we will focus on a number of general processes in development:

- how the adult body plan of animals is laid down in the embryo
- the program of gene expression that turns undifferentiated cells into differentiated cells
- the role of cell–cell communication in development

To examine these developmental processes, we will use three model systems—the fruit fly *Drosophila melanogaster,* the flowering plant *Arabidopsis thaliana,* and the nematode *Caenorhabditis elegans*. We will examine how patterns of differential gene expression lead to the progressive restriction of developmental options resulting in the formation of the adult body plan in *Drosophila* and *Arabidopsis.* We will then expand the discussion to consider how our knowledge of events in these organisms has contributed to our understanding of developmental defects in humans. Finally, we will consider the role of cell–cell communication in the development of adult structures in *C. elegans.*

20.3 Genetic Analysis of Embryonic Development in *Drosophila* Reveals How the Body Axis of Animals Is Specified

How does a given cell at a precise position in the embryo switch on or switch off specific genes at timed stages of development? This is a central question in developmental biology. To answer this question, we will examine the sequence of gene expression in the embryo of *Drosophila*. Although development in a fruit fly appears to have little in common with humans, recall that shared genes drive these steps in both species.

Overview of *Drosophila* Development

The life cycle of *Drosophila* is about 10 days long with a number of distinct phases: the embryo, three larval stages, the pupal stage, and the adult stage (**Figure 20–1**). Internally, the cytoplasm of the unfertilized egg is organized into a series of maternally constructed molecular gradients that play a key role in determining the developmental fates of nuclei located in specific regions of the embryo.

Immediately after fertilization, the zygote nucleus undergoes a series of divisions without cytokinesis [**Figure 20–2(a)** and **(b)**]. The resulting cell, which contains multiple nuclei, is called a syncytium. At about the tenth nuclear division, the nuclei move to the periphery of the egg, where the cytoplasm contains localized gradients of maternally derived mRNA transcripts and proteins [**Figure 20–2(c)**]. After several more divisions, the nuclei become enclosed in plasma membranes [**Figure 20–2(d)**], forming a cellular layer on the outside edge of the embryo. Interactions between the nuclei and the cytoplasmic components of these cells initiate and direct the pattern of embryonic gene expression.

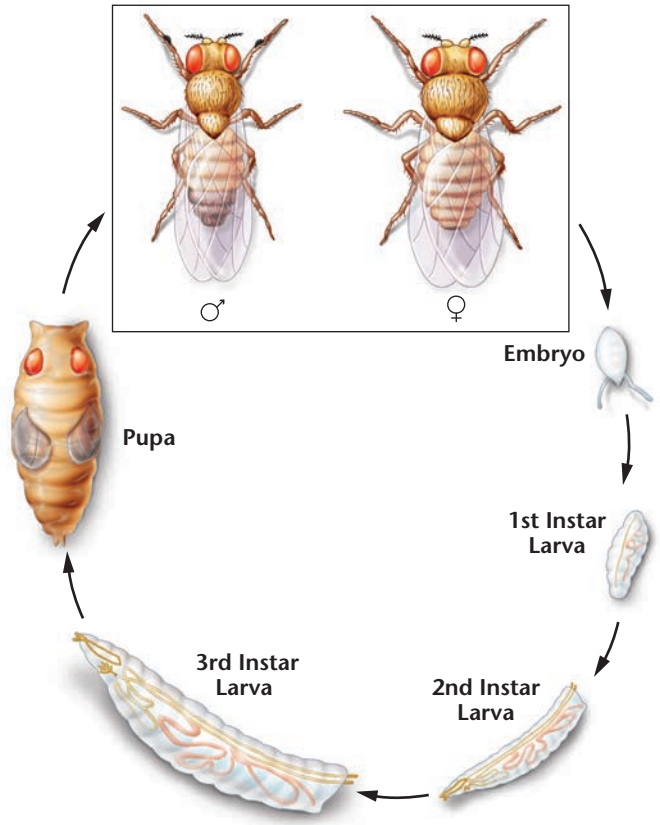


FIGURE 20–1 *Drosophila* life cycle.

Germ cells, which in the adult, are destined to undergo meiosis and produce gametes, form at the posterior pole of the embryo [Figure 20–2(c) and (d)]. Nuclei in other regions of the embryo normally form somatic cells. If nuclei from these regions are transplanted into the posterior cytoplasm, they will form germ cells and not somatic cells, confirming that the cytoplasm at the posterior pole of the embryo contains maternal components that direct nuclei to form germ cells.

Transcriptional programs activated by cytoplasmic components in somatic (non–germ-cell) nuclei form the embryo's anterior–posterior (front to back) and dorsal–ventral (upper to lower) axes of symmetry, leading to the formation of a segmented embryo [**Figure 20–2(e)**]. Under control of the *Hox* gene set (discussed in a later section), these segments will give rise to the differentiated structures of the adult fly [**Figure 20–2(f)**].

ESSENTIAL POINT

In *Drosophila,* both genetic and molecular studies have confirmed that the egg contains gradients of molecular information, which initiates a transcriptional cascade that specifies the body plan of the larva, pupa, and adult. ■

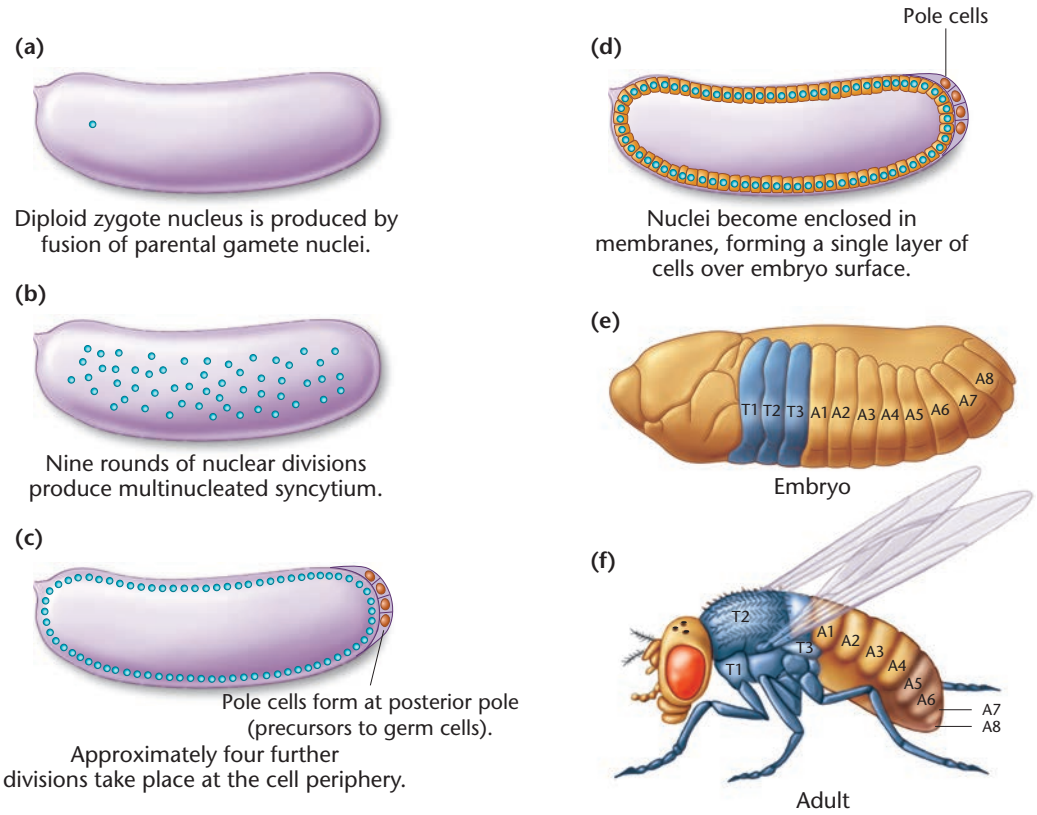


FIGURE 20–2 Early stages of embryonic development in *Drosophila*. (a) Fertilized egg with zygotic nucleus (2*n*), shortly after fertilization. (b) Nuclear divisions occur about every 10 minutes. Nine rounds of division produce a multinucleate cell, the syncytial blastoderm. (c) At the tenth division, the nuclei migrate to the periphery or cortex of the egg, and four additional rounds of nuclear division occur. A small cluster of cells, the pole cells, form at the posterior pole about 2.5 hours after fertilization. These cells will form the germ cells of the adult. (d) About 3 hours after fertilization, the nuclei become enclosed in membranes, forming a single layer of cells over the embryo surface, creating the cellular blastoderm. (e) The embryo at about 10 hours after fertilization. At this stage, the segmentation pattern of the body is clearly established. Behind the segments that will form the head, T1–T3 are thoracic segments, and A1–A8 are abdominal segments. (f) The adult fly showing the structures formed from each segment of the embryo.

Genetic Analysis of Embryogenesis

Two different gene sets control embryonic development in *Drosophila:* **maternal-effect genes** and **zygotic genes** (**Figure 20–3**). Products of maternal gene transcription (mRNA and/or proteins) are placed in the cytopasm of the developing egg. Many of these products are distributed in a gradient or concentrated in specific regions of the egg.

Female flies homozygous for deleterious recessive mutations of maternal-effect genes are sterile. None of their embryos receive wild-type maternal gene products, so all of the embryos develop abnormally and die. Maternal-effect genes encode transcription factors, receptors, and proteins that regulate gene expression. During development, these gene products activate or repress time- and location-specific programs of gene expression in the embryo.

Zygotic genes are those transcribed in the embryonic nuclei formed after fertilization. These products of the embryonic genome are differentially transcribed in specific regions of the embryo in response to the distribution of maternal-effect proteins. Recessive mutations in these genes can lead to embryonic lethality in homozygotes. In a cross between flies heterozygous for a recessive zygotic mutation, one-fourth of the embryos (the recessive homozygotes) fail to develop normally and die.

Much of our knowledge about the genes that regulate *Drosophila* development is based on the work of Ed Lewis, Christiane Nüsslein-Volhard, and Eric Wieschaus, who were awarded the 1995 Nobel Prize for Physiology or Medicine. Ed Lewis initially identified and studied one of these regulatory selector genes. Later, Nüsslein-Volhard and Wieschaus devised a strategy to identify all the genes that control the segmental pattern in *Drosophila* larvae. Their scheme required examining thousands of offspring of mutagenized flies, looking for recessive embryonic lethal mutations that alter external structures. These mutations, called segmentation genes, were grouped into three classes: *gap, pair-rule,* and *segment polarity* genes. In 1980, on the basis of their observations, Nüsslein-Volhard and

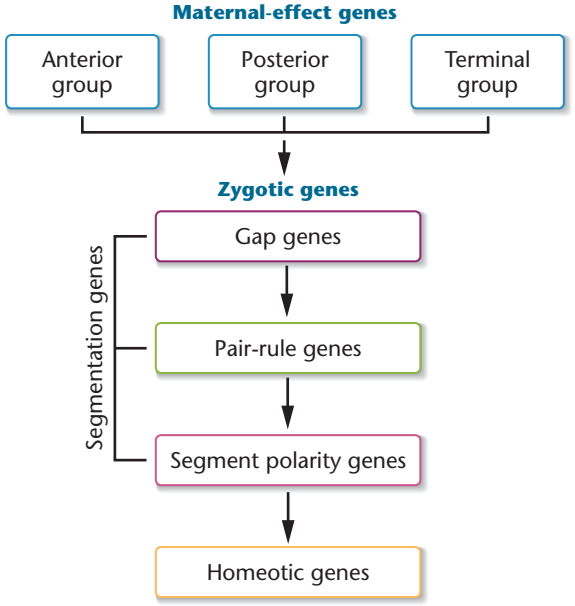


FIGURE 20–3 The hierarchy of genes involved in establishing the segmented body plan in *Drosophila*. Gene products from the maternal genes regulate the expression of the first three groups of zygotic genes (gap, pair-rule, and segment polarity, collectively called the segmentation genes), which in turn control expression of the homeotic genes.

Wieschaus proposed a model in which embryonic development is initiated by gradients of maternal-effect gene products. Then, the positional information laid down by these gradients is interpreted by two sets of zygotic (embryonic) genes: (1) the **segmentation genes** (gap, pair-rule, and segment polarity genes) identified in their search for mutants, and (2) **homeotic selector *(Hox)* genes.** Segmentation genes divide the embryo into a series of stripes or segments and define the number, size, and polarity of each segment. The homeotic genes specify the developmental fate of cells within each segment as well as the adult structures that will be formed from each segment (Figure 20–3).

Their model is shown in **Figure 20–4**. Most maternal-effect gene products placed in the egg during oogenesis are activated immediately after fertilization and help establish the anterior–posterior axis of the embryo by activating position-specific patterns of gene expression in the embryo's nuclei [Figure 20–4(a)]. Many maternal gene products encode transcription factors that activate gap genes, whose expression divides the embryo into regions corresponding to the head, thorax, and abdomen of the adult [Figure 20–4(b)]. The activated gap genes encode other transcription factors that activate pair-rule genes, whose products divide the embryo into smaller regions about two segments wide [Figure 20–4(c)]. In turn, expression of the pair-rule genes activates the segment polarity genes, which divide each segment into anterior and posterior regions [Figure 20–4(d)]. The collective action of the maternal genes and the segmentation genes define the anterior–posterior axis, which is the field of action for the homeotic (*Hox*) genes [Figure 20–4(e)].

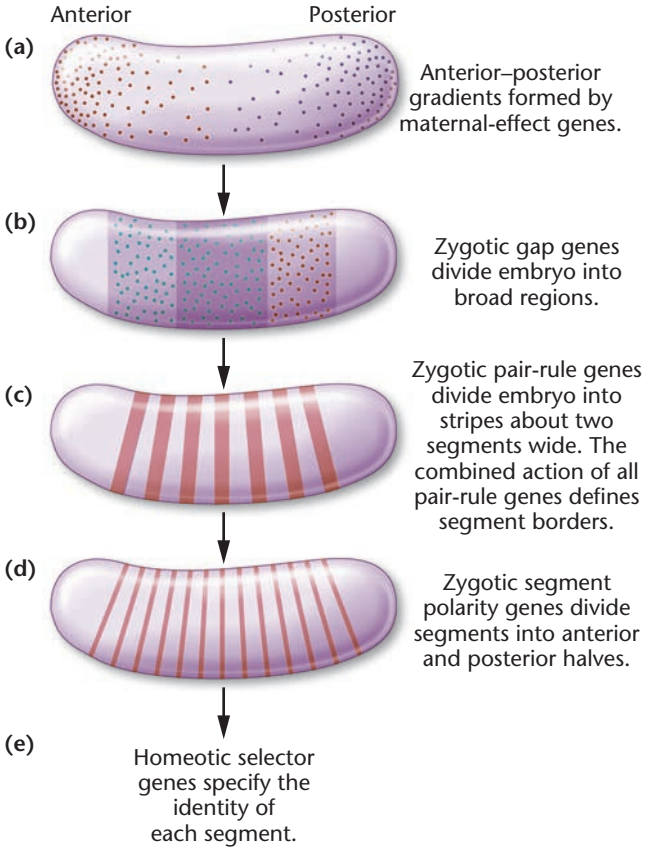


FIGURE 20–4 (a) Progressive restriction of cell fate during development in *Drosophila*. Gradients of maternal proteins are established along the anterior-posterior axis of the embryo. (b), (c), and (d) Three groups of segmentation genes progressively define the body segments. (e) Individual segments are given identity by the homeotic genes.

ESSENTIAL POINT

Maternal-effect gene products activate genes that lay down the anterior–posterior axis of the embryo and specify the location and number of segments, which in turn have their identity determined by homeotic selector genes. ■

NOW SOLVE THIS

20–1 Suppose you initiate a screen for maternal-effect mutations in *Drosophila* affecting external structures of the embryo and you identify more than 100 mutations that affect these structures. From their screenings, other researchers concluded that there are only about 40 maternal-effect genes. How do you reconcile these different results?

■ **HINT:** *This problem involves an understanding of how mutants are identified when adult Drosophila have been exposed to mutagens. The key to its solution is an understanding of the differences between genes and alleles.*

20.4 Zygotic Genes Program Segment Formation in *Drosophila*

To summarize, certain genes in the zygote's genome are activated or repressed according to a positional gradient of maternal gene products. Expression of three sets of segmentation genes divides the embryo into a series of segments along its anterior–posterior axis. These segmentation genes are normally transcribed in the developing embryo, and mutations of these genes have embryo-lethal phenotypes.

More than 20 segmentation genes (Table 20.1) have been identified, and they are classified on the basis of their mutant phenotypes: (1) mutations in gap genes delete a group of adjacent segments, causing gaps in the normal body plan of the embryo, (2) mutations in pair-rule genes affect every other segment and eliminate a specific part of each affected segment, and (3) mutations in segment polarity genes cause defects in portions of each segment.

In addition to these three sets of genes that determine the anterior–posterior axis of the developing embryo, another set of genes determines the dorsal–ventral axis of the embryo. Our discussion will be limited to the gene sets involved in the anterior–posterior axis. Let us now examine members of each group in greater detail.

Gap Genes

Transcription of **gap genes** in the embryo is controlled by maternal gene products laid down in gradients in the egg. Gap genes also cross-regulate each other to define the early stage of the body plan. Mutant alleles of these genes produce large gaps in the embryo's segmentation pattern. *Hunchback* mutants lose head and thorax structures, *Krüppel* mutants lose thoracic and abdominal structures, and *knirps* mutants lose most abdominal structures. Transcription of wild-type gap genes (which encode transcription factors) divides the embryo into a series of broad regions that become the head, thorax, and abdomen. Within these regions, different combinations of gene activity eventually specify both the type of segment that forms and the proper order of segments in the body of the larva, pupa, and adult. Expression domains of the gap genes in different parts of the embryo correlate roughly with the location of their mutant phenotypes: *hunchback* at the anterior, *Krüppel* in the middle **(Figure 20–5)**, and *knirps* at the posterior. As mentioned earlier, gap genes encode transcription factors that bind to enhancer regions that control the expression of pair-rule genes.

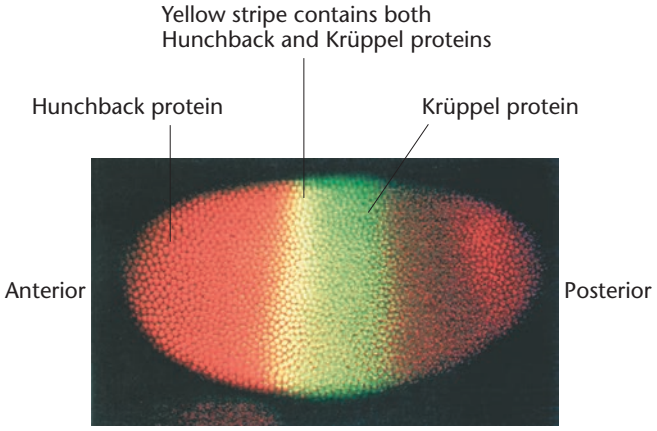


FIGURE 20–5 Expression of gap genes in a *Drosophila* embryo. The hunchback protein is shown in orange, and Krüppel is indicated in green. The yellow stripe is created when cells contain both hunchback and Krüppel proteins. Each dot in the embryo is a nucleus.

TABLE 20.1 Segmentation Genes in *Drosophila*

Gap Genes	Pair-Rule Genes	Segment Polarity Genes
Krüppel	*hairy*	*engrailed*
knirps	*even-skipped*	*wingless*
hunchback	*runt*	*cubitis*
giant	*fushi-tarazu*	*hedgehog*
tailless	*paired*	*fused*
buckebein	*odd-paired*	*armadillo*
caudal	*odd-skipped*	*patched*
	sloppy-paired	*gooseberry*
		paired
		naked
		disheveled

Pair-Rule Genes

Pair-rule genes are expressed in a series of seven narrow bands or stripes of nuclei extending around the circumference of the embryo. The expression of this gene set does two things: first it establishes the boundaries of segments, and then it programs the developmental fate of the cells within each segment by controlling expression of the segment polarity genes. Mutations in pair-rule genes eliminate segment-size sections of the embryo at every other segment. At least eight pair-rule genes act to divide the embryo into a series of stripes. The transcription of the pair-rule genes is mediated by the action of maternal gene products and gap gene products. Initially, the boundaries of these stripes overlap, so that in each area of overlap, cells express a different combination of pair-rule genes (Figure 20–6). The resolution of boundaries in this segmentation pattern results from the interaction among the gene products of the pair-rule genes themselves (Figure 20–7).

Segment Polarity Genes

Expression of **segment polarity genes** is controlled by transcription factors encoded by pair-rule genes. Within

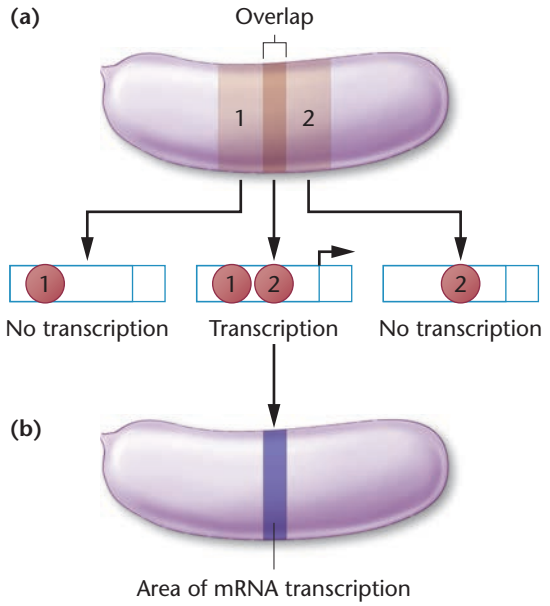
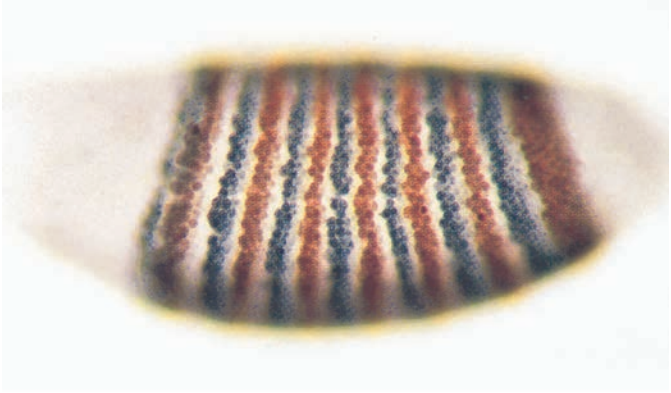


FIGURE 20–6 New patterns of gene expression can be generated by overlapping regions containing two different gene products. (a) Transcription factors 1 and 2 are present in an overlapping region of expression. If both transcription factors must bind to the promoter of a target gene to trigger expression, the gene will be active only in cells containing both factors (most likely in the zone of overlap). (b) The expression of the target gene in the restricted region of the embryo.

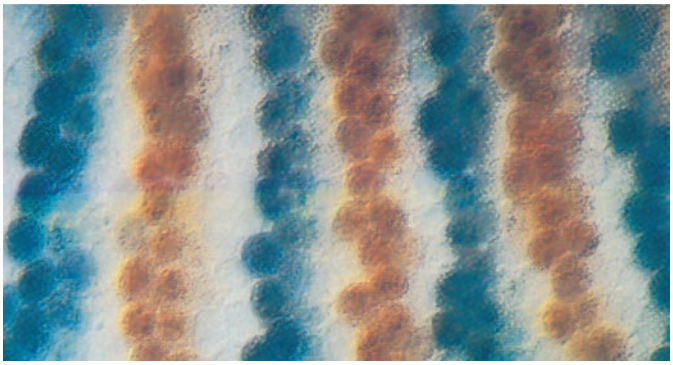
each of the segments created by pair-rule genes, segment polarity genes become active in a single band of cells that extends around the embryo's circumference (**Figure 20–8**). This divides the embryo into 14 segments. The products of the segment polarity genes control the cellular identity within each of them and establish the anterior–posterior pattern (the polarity) within each segment.

Segmentation Genes in Mice and Humans

We have seen that segment formation in *Drosophila* depends on the action of three sets of segmentation genes. Are these genes found in humans and other mammals, and do they control aspects of embryonic development in these organisms? To answer this question, let's examine *runt,* one of the pair-rule genes in *Drosophila.* Later in development, it controls aspects of sex determination and formation of the nervous system. The gene encodes a protein that regulates transcription of its target genes. Runt contains a 128-amino-acid DNA-binding region (called the runt domain) that is highly conserved in *Drosophila*, mouse, and human proteins. In fact, *in vitro* experiments show that the *Drosophila* and mouse runt proteins are functionally interchangeable. In mice, *runt* is expressed early in development and controls formation of blood cells, bone, and the genital system. Although the target gene sets controlled by *runt* are different in *Drosophila* and the mouse, in both organisms,



(a)



(b)

FIGURE 20–7 Stripe pattern of pair-rule gene expression in a *Drosophila* embryo. This embryo is stained to show patterns of expression of the genes *even-skipped* and *fushi-tarazu;* (a) low-power view and (b) high-power view of the same embryo.

expression of *runt* specifies the fate of uncommitted cells in the embryo by regulating transcription of target genes.

In humans, mutations in *RUNX2,* a human homolog of *runt,* causes cleidocranial dysplasia (CCD), an autosomal dominantly inherited trait. Those affected with CCD have a hole in the top of their skull because bone does not form in the membranous gap known as the fontanel. Their collar bones (clavicles) do not develop, enabling affected individuals to fold their shoulders across their chest. Mice with

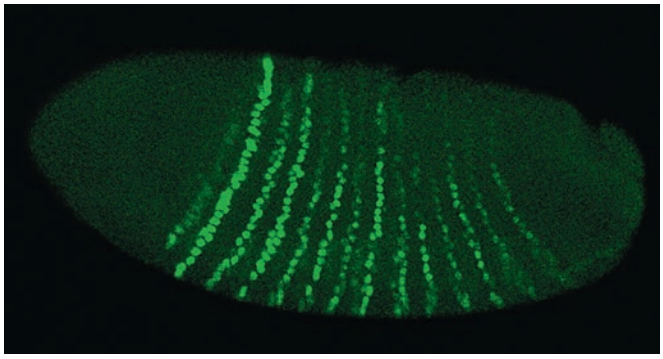


FIGURE 20–8 The 14 stripes of expression of the segment polarity gene *engrailed* in a *Drosophila* embryo.

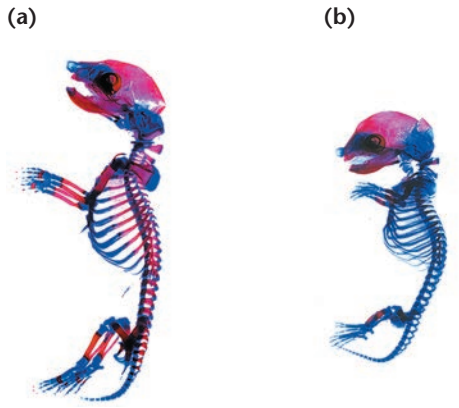


FIGURE 20–9 (a) The skeletal system of a normal mouse. Bone is stained red and cartilage is stained blue. (b) The skeletal system of a mouse carrying a mutation of the *Cbfa1* gene. Most of the skeleton contains only cartilage and not bone. Expression of the normal allele of *Cbfa1* is required for bone formation.

one mutant copy of the *runt* homolog have a phenotype similar to that seen in humans; mice with two mutant copies of the gene have no bones at all. Their skeletons contain only cartilage (**Figure 20–9**), much like sharks, emphasizing the role of *runt* as an important gene controlling the initiation of bone formation in both mice and humans.

The sequence similarity of the *runt* domain in *Drosophila*, mice, and humans and the ability of the mouse *runt* gene to replace the *Drosophila* version in fly development all indicate that the same segmentation genes are found in organisms separated from a common ancestor by millions of years.

20.5 Homeotic Selector Genes Specify Body Parts of the Adult

As segment boundaries are established by expression of segmentation genes, the homeotic (from the Greek word for "same") genes are activated. Expression of homeotic selector genes determines which adult structures will be formed by each body segment. In *Drosophila,* this includes the antennae, mouth parts, legs, wings, thorax, and abdomen. Mutant alleles of these genes are called **homeotic mutations** because one segment is transformed so that it forms the same structure as another segment. For example, the wild-type allele of the homeotic selector gene *Antennapedia (Antp)* specifies formation of a leg on the second segment of the thorax. Dominant gain-of-function *Antp* mutations cause this gene to be expressed in the head *and* the thorax. The result is that mutant flies have a leg on their head instead of an antenna (**Figure 20–10**).

Hox Genes in *Drosophila*

The *Drosophila* genome contains two clusters of homeotic selector genes (called *Hox* genes) on chromosome 3 that encode transcription factors (**Table 20.2**). The *Antennapedia (ANT-C)* cluster contains five genes that specify structures in the head and first two segments of the thorax [**Figure 20–11(a)**]. The second cluster, the *bithorax (BX-C)* complex, contains three genes that specify structures in the second and third segments of the thorax, and the abdominal segments [**Figure 20–11(b)**].

Hox genes (listed in Table 20.2) from a wide range of species have two properties in common. First, each contains a highly conserved 180-bp nucleotide sequence known as a **homeobox.** (*Hox* is a contraction of homeobox.) The homeobox encodes a DNA-binding region of 60 amino acids known

(a)

(b)

FIGURE 20–10 *Antennapedia (Antp)* mutation in *Drosophila.* (a) Head from wild-type *Drosophila*, showing the antenna and other head parts. (b) Head from an *Antp* mutant, showing the replacement of normal antenna structures with legs. This is caused by activation of the *Antp* gene in the head region.

TABLE 20.2 ***Hox* Genes of *Drosophila***

***Antennapedia* Complex**	***Bithorax* Complex**
labial	*Ultrabithorax*
Antennapedia	*abdominal A*
Sex combs reduced	*Abdominal B*
Deformed	
proboscipedia	

as a **homeodomain.** Second, in most species, expression of *Hox* genes is colinear with the anterior to posterior organization of the body. In other words, genes at the beginning of the cluster (the 3′-end) are expressed at the anterior end of the embryo, those in the middle are expressed in the middle of the embryo, and genes at the end of a cluster (the 5′-end) are expressed at the embryo's posterior region (**Figure 20–12**). Although first identified in *Drosophila, Hox* genes are found in the genomes of most eukaryotes with segmented body plans, including nematodes, sea urchins, zebrafish, frogs, mice, and humans (**Figure 20–13**).

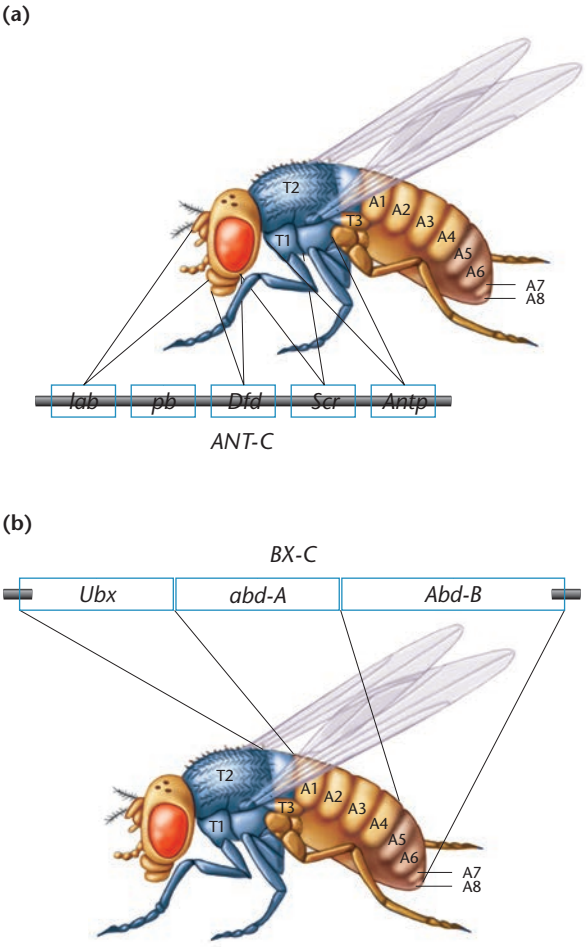


FIGURE 20–11 Genes of the *Antennapedia* complex and the adult structures they specify. (a) In the *ANT-C* complex, the *labial* (*lab*) and *Deformed* (*Dfd*) genes control the formation of head segments. The *Sex combs reduced* (*Scr*) and *Antennapedia* (*Antp*) genes specify the identity of the first two thoracic segments, T1 and T2. The remaining gene in the complex, *proboscipedia* (*pb*), may not act during embryogenesis but may be required to maintain the differentiated state in adults. In mutants, the labial palps are transformed into legs. (b) In the *BX-C* complex, *Ultrabithorax* (*Ubx*) controls formation of structures in the posterior compartment of T2 and structures in T3. The two other genes, *abdominal A* (*abdA*) and *Abdominal B* (*AbdB*), specify the segmental identities of the eight abdominal segments (A1–A8).

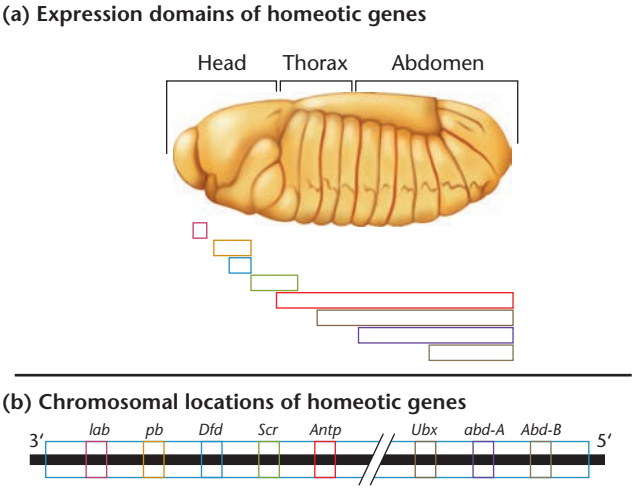


FIGURE 20–12 The colinear relationship between the spatial pattern of expression and chromosomal locations of homeotic genes in *Drosophila.* (a) *Drosophila* embryo and the domains of homeotic gene expression in the embryonic epidermis and central nervous system. (b) Chromosomal location of homeotic selector genes. Note that the order of genes on the chromosome correlates with the sequential anterior borders of their expression domains.

To summarize, genes that control development in *Drosophila* act in a temporally and spatially ordered cascade, beginning with the genes that establish the anterior–posterior (and dorsal–ventral) axis of the early embryo. Gradients of maternal mRNAs and proteins along the anterior–posterior axis activate gap genes, which subdivide the embryo into broad bands. Gap genes in turn activate pair-rule genes, which divide the embryo into segments. The final group of segmentation genes, the segment polarity genes, divides each segment into anterior and posterior regions arranged linearly along the anterior–posterior axis. The segments are then given identity by action of the *Hox* genes. Therefore, this progressive restriction of developmental potential of the *Drosophila* embryo's cells (all of which occurs during the first third of embryogenesis) involves a cascade of gene action, with regulatory proteins acting on transcription, translation, and signal transduction.

Hox Genes and Human Genetic Disorders

Hox genes are found in the genomes of all animals where they play a fundamental role in shaping the body and its appendages. In vertebrates, the conservation of sequence, the order of genes in the *Hox* clusters, and their pattern of expression in vertebrates suggest that, as in *Drosophila,*

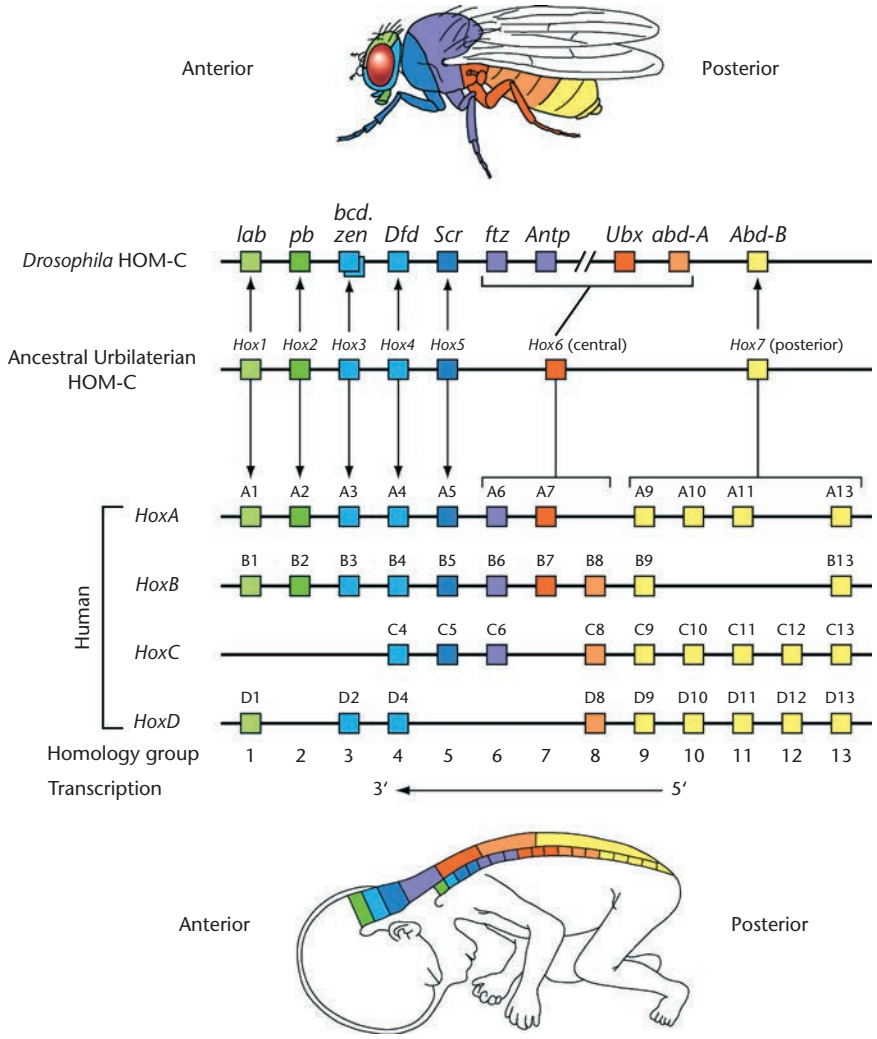


FIGURE 20–13 Conservation of organization and patterns of expression in *Hox* genes. (Top) The structures formed in adult *Drosophila* are shown, with the colors corresponding to members of the *Hox* cluster that control their formation. (Bottom) The arrangement of the *Hox* genes in an early human embryo. As in *Drosophila,* genes at the 3′-end of the cluster form anterior structures, and genes at the 5′-end of the cluster form posterior structures.

these genes control development along the anterior–posterior and the formation of appendages. However, in vertebrates, there are four clusters of *Hox* genes: *HOXA*, *HOXB*, *HOXC,* and *HOXD* instead of a single cluster as in *Drosophila*. This means that in vertebrates, not just one, but a combination of 2 to 4 *Hox* genes, is involved in forming specific structures. As a result, in vertebrates, mutations in individual *Hox* genes do not produce complete transformation as in *Drosophila*, where mutation of a single *Hox* gene can transform a haltere into a wing (see the photo at the beginning of this chapter). The role of *HOXD* genes in human development was confirmed by the discovery that several inherited limb malformations are caused by mutations in *HOXD* genes. For example, mutations in *HOXD13* cause synpolydactyly (SPD), a malformation characterized by extra fingers and toes, and abnormalities in bones of the hands and feet (**Figure 20–14**).

ESSENTIAL POINT

Once the boundaries of body segments have been established by segmentation genes, the homeotic selector genes act to specify which body structures will be formed by each segment. ■

NOW SOLVE THIS

20–2 In *Drosophila*, both *fushi tarazu* (*ftz*) and *engrailed* (*eng*) genes encode homeobox transcription factors and are capable of eliciting the expression of other genes. Both genes work at about the same time during development and in the same region to specify cell fate in body segments. To discover if *ftz* regulates the expression of *engrailed;* if *engrailed* regulates *ftz;* or if both are regulated by another gene, you perform a mutant analysis. In *ftz* embryos (*ftz/ftz*) engrailed protein is absent; in *engrailed* embryos (*eng/eng*) *ftz* expression is normal. What does this tell you about the regulation of these two genes—does the *engrailed* gene regulate *ftz,* or does the *ftz* gene regulate *engrailed*?

■ **HINT:** *This problem involves an understanding of how genes are regulated at different stages of preadult development in Drosophila. The key to its solution lies in using the results of the mutant analysis to determine the timing of expression of the two genes being examined.*

FIGURE 20–14 Mutations in posterior *Hox* genes (*HOXD13* in this case) in humans result in malformations of the limbs, shown here as extra toes. This condition is known as synpolydactyly. Mutations in *HOXD13* are also associated with abnormalities of the bones in the hands and feet.

20.6 Binary Switch Genes and Regulatory Pathways Program Organ Formation

The *Hox* genes that determine which adult structures will be formed by each body segment in *Drosophila* act as switches, selecting alternative developmental pathways for cells to follow. Each pathway decision point is usually binary—that is, there are two alternative developmental fates for a cell at a given time—and these **binary switch genes** program a cell to follow only one of these pathways. These genes, which encode transcription factors, are defined by their ability to initiate complete development of an organ or a tissue type and are part of **gene-regulatory networks** (**GRNs**) and subnetworks that program transcription of gene sets at specific times and specific stages of tissue and organ formation.

Here we will describe how a binary switch gene controls the formation of the eye in *Drosophila* and how its regulatory network is used in all organisms with eyes.

The Control of Eye Formation

Drosophila adults have compound eyes [**Figure 20–15(a)**]. Action of the wild-type allele of the binary switch gene *eyeless* programs cells to follow the developmental pathway for eye formation instead of the pathway for antenna formation. In flies homozygous for recessive mutant alleles of the *eyeless* gene, the eyes are reduced in size and have irregular facets [**Figure 20–15(b)**]. In developmental pathways that normally specify the formation of other organs such as legs, wings, and antennae, abnormal expression of *eyeless* may result in eye formation in other parts of the body [**Figure 20–15(c)**]. This indicates that switching on the *eyeless* gene at the wrong time or in the wrong cells can override normal programs of determination and differentiation, causing cells to follow the developmental program for eye formation instead of their normal pathway.

Gene Networks in Eye Formation

The wild-type allele of the *eyeless* gene is part of a GRN that is the master regulator of eye formation. Five cross-regulating genes that encode transcription factors are the core of this network. However, much less is known about events in downstream regulation and the number of target genes that control eye development. However, recent work combining

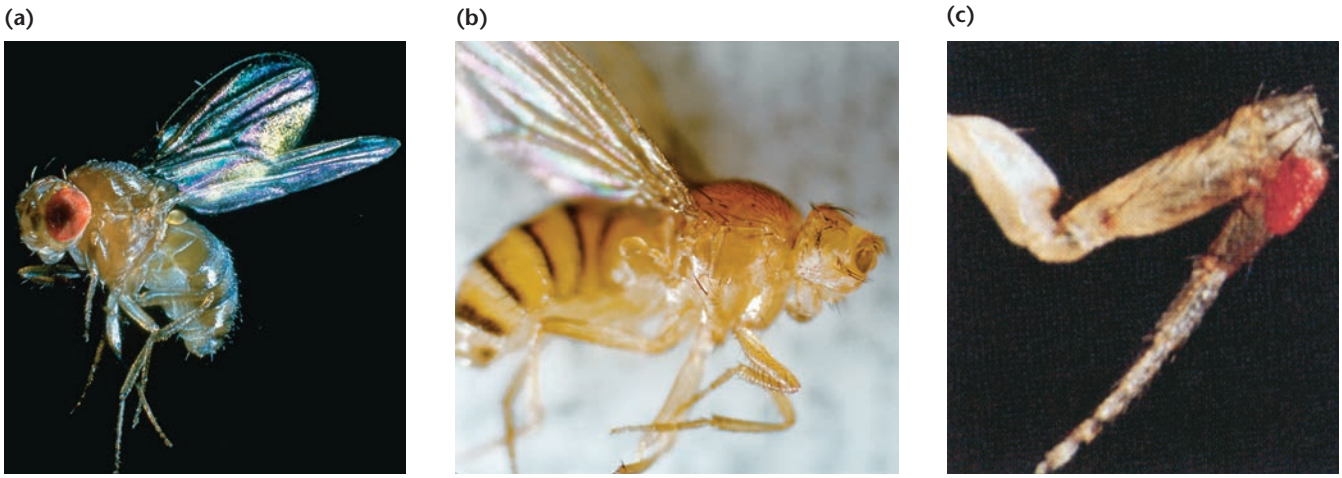


FIGURE 20–15 (a)The normal compound eye of adult *Drosophila*. (b) In flies homozygous for the *eyeless* (*ey*) mutation, eye development is abnormal, and adults have reduced or absent eyes. The *ey* gene is a binary switch gene in a gene-regulatory network that controls eye formation in all animals. (c) An eye formed on the leg of a fly. This abnormal location is the result of *eyeless* expression in cells normally destined to form leg structures.

genomics and classic reverse genetics has revealed that the *eyeless* GRN is large and complex, and contains 241 genes encoding transcription factors and has more than 5,600 target genes. Eye formation in *Drosophila* (and in other animals) is obviously an extremely complex event. In spite of its size and complexity, this GRN has been highly conserved during evolution, and its core components are used by all animals, including humans, to make eyes. The discovery that the *eyeless* gene and its vertebrate equivalent, *Pax6*, have a high degree of DNA sequence homology and are expressed during eye development forced reevaluation of the long-held belief that the compound eyes of insects and the single-lens eyes of vertebrates evolved independently. In addition, copies of the mouse *Pax6* gene, when transferred to *Drosophila* as transgenes, control eye formation in these flies, demonstrating that the eyes of *Drosophila* and the mouse are homologous structures. The downstream targets of these binary switch genes are also conserved, indicating that steps in the genetic control of eye development are shared between species that diverged over half a billion years ago from a common ancestor. This evolutionary conservation makes it possible to use genetic analysis in *Drosophila* to study the development of eyes and to explore the molecular basis for inherited eye defects in humans.

20.7 Plants Have Evolved Developmental Regulatory Systems That Parallel Those of Animals

Plants and animals diverged from a common ancestor about 1.6 billion years ago, after the origin of eukaryotes and probably before the rise of multicellular organisms. Genetic analysis of mutants and genome sequencing in plants and animals indicate that basic mechanisms of pattern formation evolved independently in animals and plants. We have already examined the genetic systems that control development and pattern formation in animals, using *Drosophila* as a model organism.

In plants, pattern formation has been extensively studied using flower development in *Arabidopsis thaliana* (**Figure 20–16**), a small plant in the mustard family, as a model organism. A cluster of undifferentiated cells, called the *floral meristem,* gives rise to flowers (**Figure 20–17**). Each flower consists of four organs—sepals, petals, stamens, and carpels—that develop from concentric rings of cells within the meristem [**Figure 20–18(a)**]. Each organ develops from a different concentric ring, or whorl of cells.

Homeotic Genes in *Arabidopsis*

Three classes of floral homeotic genes control the development of these organs (**Table 20.3**). Acting alone, class A genes specify sepals; class A and class B genes acting together specify petals. Expression of both class B and class C genes controls stamen formation. Class C genes acting alone specify carpels. During flower development [**Figure 20–18(b)**], class A genes are active in whorls 1 and 2 (sepals and petals), class B genes are expressed in whorls 2 and 3 (petals and stamens), and class C genes are expressed in whorls 3 and 4 (stamens and carpels). Which organs are formed depends on the expression pattern of the three gene classes. In whorl 1, expression of class A genes alone causes sepals to form. Expression of class A *and* class B genes in whorl 2 leads to petal formation. Expression of class B and class C genes in whorl 3 leads to stamen formation. In whorl 4, expression of class C genes alone causes carpel formation.

As in *Drosophila,* mutations in homeotic genes cause organs to form in abnormal locations. For example, in *APETALA2* mutants (*ap2*), the order of organs is carpel, stamen, stamen, and carpel instead of the normal order, sepal, petal, stamen, and carpel [**Figure 20–19(a)** and **(b)**]. In class B loss-of-function mutants (*ap3*, *pi*), petals become sepals, and stamens are transformed into carpels [**Figure 20–19(c)**], and the order of organs becomes sepal,

FIGURE 20.16 The flowering plant *Arabidopsis thaliana,* used as a model organism in plant genetics.

TABLE 20.3 Homeotic Selector Genes in *Arabidopsis**

Class A	*APETALA1 (AP1)*
	APETALA2 (AP2)
Class B	*APETALA3 (AP3)*
	PISTILLATA (P1)
Class C	*AGAMOUS (AG)*

*By convention, wild-type genes in *Arabidopsis* use capital letters.

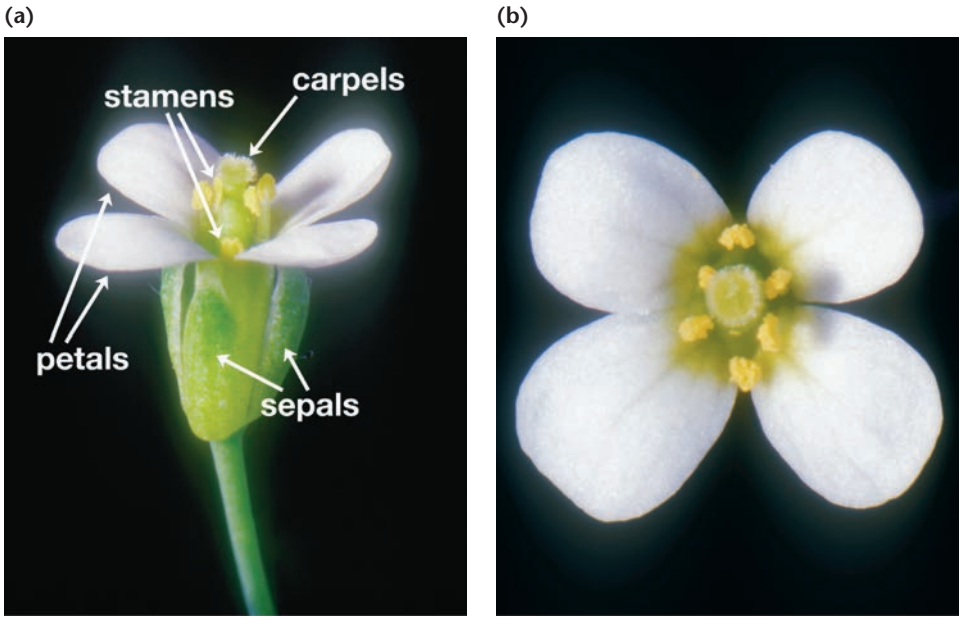


FIGURE 20–17 (a) Parts of the *Arabidopsis* flower. The floral organs are arranged concentrically. The sepals form the outermost ring, followed by petals and stamens, with carpels on the inside. (b) View of the flower from above.

sepal, carpel, carpel. Plants carrying a mutation for the Class 3 gene *AGAMOUS* will have petals in whorl 3 (instead of stamens) and sepals in whorl 4 (instead of carpels), and the order of organs will be sepal, petal, petal, and sepal [**Figure 20–19(d)**].

Evolutionary Divergence in Homeotic Genes

Drosophila and *Arabidopsis* use different sets of master regulatory genes to establish the body axis and specify the identity of structures along the axis. In *Drosophila,* this task is accomplished in part by the *Hox* genes, which encode a set of transcription factors sharing a homeobox domain. In *Arabidopsis,* the floral homeotic genes belong to a different family of transcription factors, called the **MADS-box proteins,** characterized by a shared domain of 58 amino acids with no similarity in amino acid sequence or protein structure with the *Hox* genes. Both gene sets encode transcription factors, both sets are master regulators of development expressed in a pattern of overlapping domains, and both specify identity of structures.

Reflecting their evolutionary origin from a common ancestor, the genomes of *Drosophila* and *Arabidopsis* both contain members of the homeobox and MADS-box genes, but these genes have been adapted for different uses in the plant and animal kingdoms. This indicates that developmental mechanisms evolved independently in each group.

In both plants and animals, the action of transcription factors depends on changes in chromatin structure that make genes available for expression. Mechanisms of transcription initiation are conserved in plants and animals, as is reflected in the homology of genes in *Drosophila* and *Arabidopsis* that maintain patterns of expression initiated by regulatory gene sets. Action of the floral homeotic genes is controlled by a gene called *CURLY LEAF*. This gene shares significant homology with members of the *Polycomb* gene family in *Drosophila*. This family of regulatory genes controls expression of homeobox genes during development. Both *CURLY LEAF* and *Polycomb* encode proteins that alter chromatin conformation and shut off gene expression. Thus, although different genes are used

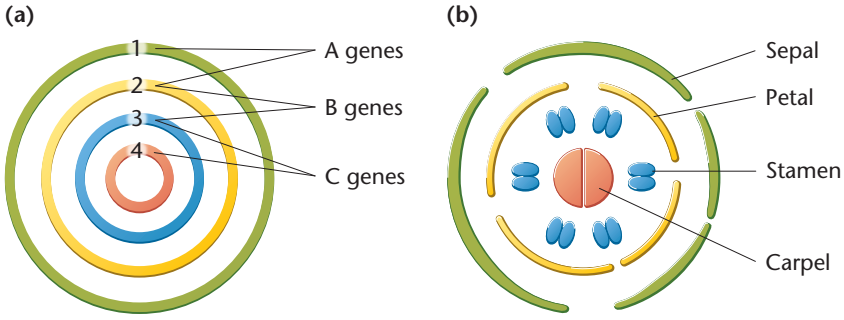


FIGURE 20–18 Cell arrangement in the floral meristem. (a) The four concentric rings, or whorls, labeled 1–4, give rise to (b) arrangement of the sepals, petals, stamens, and carpels, respectively, in the mature flower.

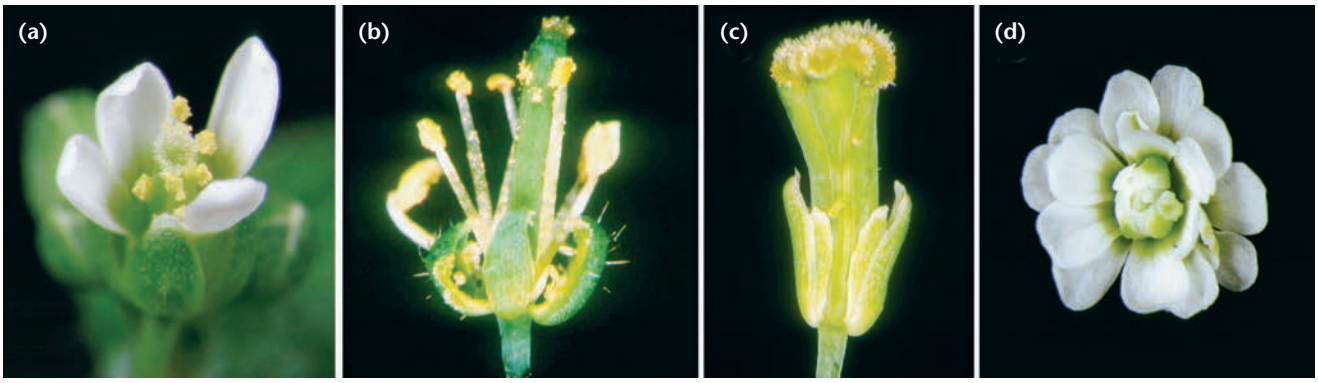


FIGURE 20–19 (a) Wild-type flowers of *Arabidopsis* have (from outside to inside) sepals, petals, stamens, and carpels. (b) A homeotic *APETALA2* mutant flower has carpels, stamens, stamens, and carpels. (c) *PISTILLATA* mutants have sepals, sepals, carpels, and carpels. (d) *AGAMOUS* mutants have petals and sepals at places where stamens and carpels should form.

to control development, both plants and animals use an evolutionarily conserved mechanism to regulate expression of these gene sets.

ESSENTIAL POINT

Flower formation in *Arabidopsis* is controlled by homeotic genes, but these gene sets are from a different gene family than the homeotic selector genes of *Drosophila* and other animals. ■

20.8 *C. elegans* Serves as a Model for Cell–Cell Interactions in Development

During development in multicellular organisms, cell–cell interactions influence the transcriptional programs and developmental fate of the interacting cells and surrounding cells. Cell–cell interaction is an important process in the embryonic development of most eukaryotic organisms, including *Drosophila,* mice, and humans.

Signaling Pathways in Development

In early development, animals use a number of signaling pathways to regulate development; after organ formation begins, additional pathways are added to those already in use. These newly activated pathways act both independently and in coordinated networks to generate specific transcriptional patterns. Signal networks establish anterior–posterior polarity and body axes, coordinate pattern formation, and direct the differentiation of tissues and organs. The signaling pathways used in early development and some of the developmental processes they control are listed in **Table 20.4**. After an introduction to the components and interactions of one of these systems—the **Notch signaling pathway**—we will briefly examine its role in the development of the vulva in the nematode, *Caenorhabditis elegans.*

The Notch Signaling Pathway

The genes in the Notch pathway are named after the *Drosophila* mutants that were used to identify components of this signal transduction system (*Notch* mutants have an indentation or notch in their wings). Notch signaling works through direct cell–cell contact to control the developmental fate of interacting cells. The *Notch* gene (and the equivalent gene in other organisms) encodes a receptor protein embedded in the plasma membrane (**Figure 20–20**). The signal is another membrane protein encoded by the *Delta* gene (and its equivalents). Because both the signal and receptor are membrane proteins, the Notch signal system works only when adjacent cells come into physical contact. When the Delta protein from one cell binds to the Notch

TABLE 20.4 Signaling Pathways Used in Early Embryonic Development

Wnt Pathway
Dorsalization of body
Female reproductive development
Dorsal–ventral differences
TGF-β Pathway
Mesoderm induction
Left–right asymmetry
Bone development
Hedgehog Pathway
Notochord induction
Somitogenesis
Gut/visceral mesoderm
Receptor Tyrosine Kinase Pathway
Mesoderm maintenance
Notch Signaling Pathway
Blood cell development
Neurogenesis
Retina development

*Source: *Taken from* Gerhart, J. 1999. 1998 Warkany lecture: Signaling pathways in development. *Teratology* 60: 226–239.

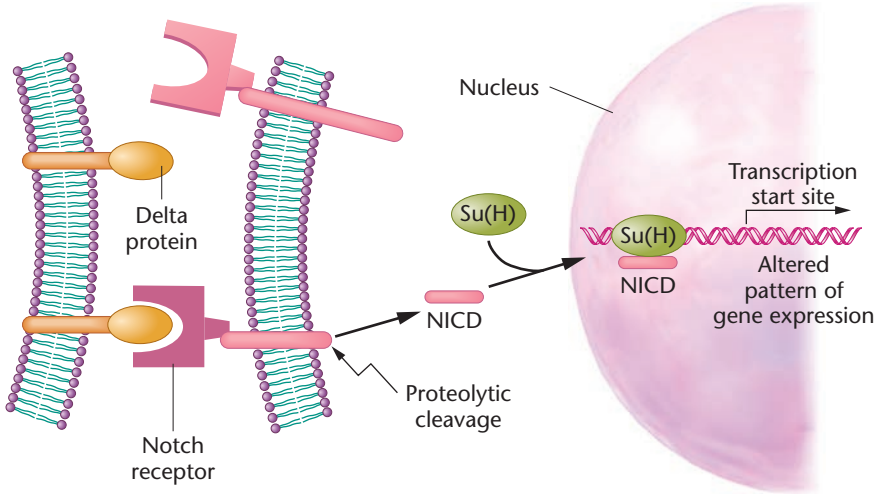


FIGURE 20–20 Components of the Notch signaling pathway in *Drosophila*. The cell carrying the Delta transmembrane protein is the sending cell; the cell carrying the transmembrane Notch protein receives the signal. Binding of Delta to Notch triggers a proteolytic-mediated activation of transcription. The fragment cleaved from the cytoplasmic side of the Notch protein, called the Notch intracellular domain (NICD), combines with the Su(H) protein and moves to the nucleus where it activates a program of gene transcription.

receptor protein on a neighboring cell, the cytoplasmic tail of the Notch protein is cleaved off and binds to a cytoplasmic protein encoded by the *Su(H)* (suppressor of *Hairless*) gene. This protein complex moves into the nucleus and binds to transcriptional cofactors, activating transcription of a gene set that controls a specific developmental pathway (Figure 20–20).

One of the main roles of the Notch signal system is to specify different developmental fates for equivalent cells in a population. In its simplest form, this interaction involves two neighboring cells that are developmentally equivalent but become specified to form different adult structures. We will explore the role of the Notch signaling system in development of the vulva in *C. elegans,* after a brief introduction to nematode embryogenesis.

Overview of *C. elegans* Development

The nematode *C. elegans* is widely used as a model organism to study the genetic control of development. There are several advantages in using this organism: (1) its genetics are well known, (2) its genome has been sequenced, and (3) adults contain a small number of cells that follow a highly deterministic developmental program. Adult nematodes are about 1 mm long and develop from a fertilized egg in about two days (**Figure 20–21**). The life cycle includes an embryonic stage (about 16 hours), four larval stages (L1 through L4), and the adult stage. Adults are of two sexes: XX self-fertilizing hermaphrodites that can make both eggs and sperm, and XO males. Self-fertilization of mutagen-treated hermaphrodites is used to develop homozygous

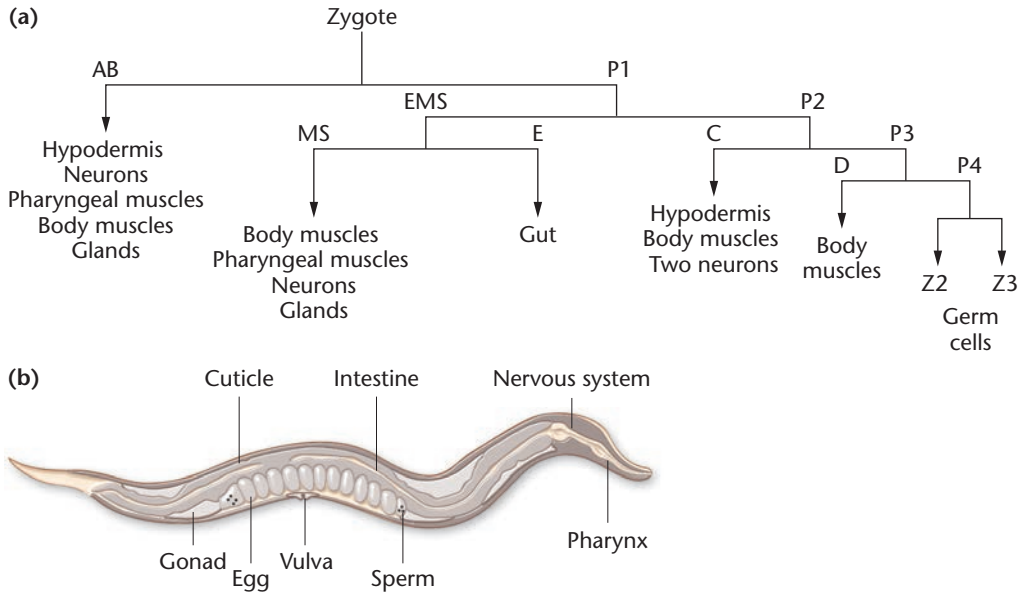


FIGURE 20–21 (a) A truncated cell lineage chart for *C. elegans,* showing early divisions and the tissues and organs formed from these lineages. Each vertical line represents a cell division, and horizontal lines connect the two cells produced. For example, the first division of the zygote creates two new cells, AB and P1. During embryogenesis, cell divisions will produce the 959 somatic cells of the adult hermaphrodite worm. (b) An adult *C. elegans* hermaphrodite. This nematode, about 1 mm in length, consists of 959 cells and is widely used as a model organism to study the genetic control of development.

stocks of mutant strains, and hundreds of such mutants have been generated, cataloged, and mapped.

Adult hermaphrodites have 959 somatic cells (and about 2000 germ cells). The lineage of each cell, from fertilized egg to adult, has been mapped (Figure 20–21) and is invariant from individual to individual. Knowing the lineage of each cell, we can easily follow altered cell fates generated by mutations or by killing specific cells with laser microbeams or ultraviolet irradiation. In hermaphrodites, the developmental fate of cells in the reproductive system is determined by cell–cell interaction, illustrating how gene expression and cell–cell interaction work together to specify developmental outcomes.

NOW SOLVE THIS

20–3 The identification and characterization of genes that control sex determination have been another focus of investigators working with *C. elegans*. As with *Drosophila*, sex in this organism is determined by the ratio of X chromosomes to sets of autosomes. A diploid wild-type male has one X chromosome, and a diploid wild-type hermaphrodite has two X chromosomes. Many different mutations have been identified that affect sex determination. Loss-of-function mutations in a gene called *her-1* cause an XO nematode to develop into a hermaphrodite and have no effect on XX development. (That is, XX nematodes are normal hermaphrodites.) In contrast, loss-of-function mutations in a gene called *tra-1* cause an XX nematode to develop into a male. Deduce the roles of these genes in wild-type sex determination from this information.

■ **HINT:** *This problem involves an understanding of the mechanism of sex determination by the ratio of X chromosomes to sets of autosomes. The key to its solution is an understanding of the effect of loss-of-function mutations on expression of other genes or the action of other proteins.*

Genetic Analysis of Vulva Formation

Adult *C. elegans* hermaphrodites lay eggs through the vulva, an opening near the middle of the body (Figure 20–21). The vulva is formed in stages during larval development and involves several rounds of cell–cell interactions.

In *C. elegans,* interaction between two neighboring cells, Z1.ppp and Z4.aaa, determines which will become the gonadal anchor cell (from which the vulva forms) and which will become a precursor to the uterus (**Figure 20–22**). The determination of which cell becomes which occurs during the second larval stage (L2) and is controlled by the Notch receptor gene, *lin-12*. In recessive *lin-12(0)* mutants (a loss-of-function mutant), no functional receptor protein is present, and both cells become anchor cells. The

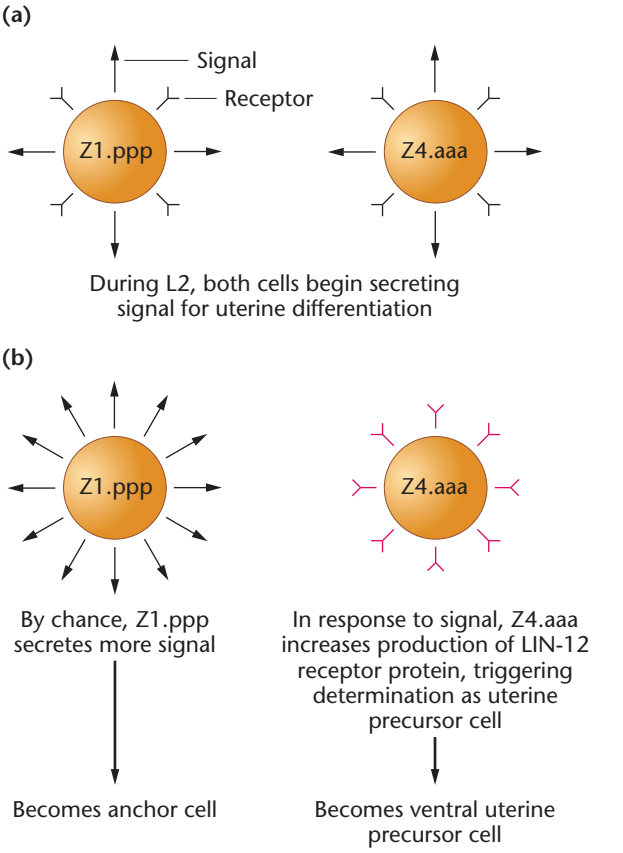


FIGURE 20–22 Cell–cell interaction in anchor cell determination. (a) During L2, two neighboring cells begin the secretion of chemical signals for the induction of uterine differentiation. (b) By chance, cell Z1.ppp produces more of these signals, causing cell Z4.aaa to increase production of the receptor for signals. The action of increased signals causes Z4.aaa to become the ventral uterine precursor cell and allows Z1.ppp to become the anchor cell.

dominant mutation *lin-12(d)* (a gain-of-function mutation) causes both to become uterine precursors. Based on the phenotypes of these two mutant alleles, we can conclude that normally, expression of *lin-12* directs selection of the uterine pathway because in the absence of the LIN-12 (Notch) receptor, both cells become anchor cells.

However, the situation is more complex than it first appears. Initially, the two neighboring cells are developmentally equivalent. Each synthesizes low levels of the Notch signal protein (encoded by the *lag-2* gene) *and* the Notch receptor protein (encoded by the *lin-12* gene). By chance, one cell ends up secreting more of the signal (LAG-2 or Delta protein) than the other cell. This causes the neighboring cell to increase production of the receptor (LIN-12 protein). The cell producing more of the receptor protein becomes the uterine precursor, and the cell, producing more signal protein, becomes the anchor cell. The critical factor in this first round of cell–cell interaction is the balance between the LAG-2 (Delta) signal gene product and the LIN-12 (Notch) receptor gene product.

Once the gonadal anchor cell has been determined, a second round of cell–cell interaction leads to formation of the vulva. This interaction involves the anchor cell (located in the gonad) and six neighboring cells (called precursor cells) located in the skin. The precursor cells, named P3.p to P8.p, are called Pn.p cells. The developmental fate of each Pn.p cell is specified by its position relative to the anchor cell.

During vulval development, the LIN-3 signal protein is synthesized by the anchor cell; this signal is received and processed by three adjacent Pn.p precursor cells (Pn.p 5–7). The cell closest to the anchor cell (usually Pn.p 6) becomes the primary vulval precursor cell, and the adjacent cells (Pn.p 5 and 7) become secondary precursor cells. A signal protein from the primary vulval precursor cell activates the *lin-12* receptor gene in the secondary cells, preventing them from becoming primary precursor cells. The other precursor cells (Pn.p 3, 4, and 8) receive no signal from the anchor cell and become skin cells.

ESSENTIAL POINT

In *C. elegans,* the well-studied pathway of cell lineage during embryonic development allows developmental biologists to study the cell–cell signaling required for organogenesis. ■

GENETICS, TECHNOLOGY, AND SOCIETY

Stem Cell Wars

Stem cell research may be the most controversial research area since the beginning of recombinant DNA technology in the 1970s. Although stem cell research is the focus of presidential proclamations, media campaigns, and ethical debates, few people understand it sufficiently to evaluate its pros and cons.

Stem cells are primitive cells that replicate indefinitely and have the capacity to differentiate into cells with specialized functions, such as the cells of heart, brain, liver, and muscle tissue. Some types of stem cells are defined as *totipotent,* meaning that they have the ability to differentiate into any mature cell type in the body, as well as tissues associated with the developing embryo, such as placenta. Other types of stem cells are *pluripotent,* meaning that they are able to differentiate into any of a smaller number of mature cell types. In contrast, mature, fully differentiated cells do not replicate or undergo transformations into different cell types.

In the last few years, several research teams have isolated and cultured human pluripotent stem cells. When treated with growth factors or hormones, these pluripotent stem cells differentiate into cells that have the characteristics of neural, bone, kidney, liver, heart, or pancreatic cells.

The fact that pluripotent stem cells grow prolifically in culture and differentiate into more specialized cells has created great excitement. Some foresee a day when stem cells may be a cornucopia from which to harvest unlimited numbers of specialized cells to replace cells in damaged and diseased tissues. Hence, stem cells could be used to treat Parkinson disease, type 1 diabetes, chronic heart disease, Alzheimer disease, genetic defects, and even cancers. The excitement about stem cell therapies has been fueled by reports of dramatically successful experiments in animals. For example, mice with spinal cord injuries regained their mobility and bowel and bladder control after they were injected with human stem cells. Given the potential for such beneficial treatments, why should stem cell research be so contentious?

The answer to that question lies in the source of the pluripotent stem cells. Until recently, all pluripotent stem cell lines were derived from five-day-old embryonic blastocysts. Blastocysts at this stage consist of 50–150 cells, most of which will develop into placental and supporting tissues for the early embryo. The inner cell mass of the blastocyst consists of about 30 to 40 pluripotent stem cells that can develop into all the embryo's tissues. *In vitro* fertilization clinics grow fertilized eggs to the five-day blastocyst stage prior to uterine transfer. Pluripotent embryonic stem cell (ESC) lines are created by taking the inner cell mass out of five-day blastocysts and growing the cells in culture dishes.

The fact that early embryos are destroyed in the process of establishing human ESC lines disturbs people who believe that preimplantation embryos are persons with rights; however, it does not disturb people who believe that these embryos are too primitive to have the status of a human being. Both sides in the debate invoke fundamental questions of what constitutes a human being.

Recently, scientists have developed several types of pluripotent stem cells without using embryos. One of the most promising types—known as *induced pluripotent stem (iPS) cells*—uses adult somatic cells as the source of pluripotent stem cell lines. To prepare iPS cells, scientists isolate somatic cells (such as cells from skin) and infect them with engineered retroviruses that integrate into the cells' DNA. These retroviruses contain several cloned human genes that encode products responsible for converting the somatic cells into immortal, pluripotent stem cells.

The development of iPS cell lines has the potential to bypass the ethical problems associated with the use of human embryos. In addition, they may become sources of patient-specific pluripotent stem cell lines that can be used for transplantation, without immune system rejection.

(continued)

Genetics, Technology, and Society, continued

Your Turn

Take time, individually or in groups, to answer the following questions. Investigate the references and links to help you understand the technologies and controversies surrounding stem cell research.

1. In 2013, a research group reported that they had created a human blastocyst after transferring a somatic cell nucleus into an enucleated human egg and activating the egg. They used the blastocyst to isolate ESCs that were a genetic match to the nucleus donor. How did these researchers do this, and why did they use this method rather than derive iPS cells from the donor?

   Read about this study in Tachibana, M., et al. 2013. Human embryonic stem cells derived by somatic cell nuclear transfer. Cell 153: 1–11.

2. The technologies that give rise to iPS cells and ESCs using nuclear transfer are based on the work of Drs. John Gurdon and Shinya Yamanaka. In 2012, these scientists were awarded the Nobel Prize in Physiology or Medicine for their work in reprogramming adult cells to stem cells. Describe their key discoveries on which stem cell therapies are based.

   A summary of their work can be found on the Nobel Prize Web site: http://www.nobelprize.org/nobel_prizes/medicine/laureates/2012/advanced.html.

3. Although several stem cell therapies are in clinical trials, some unregulated clinics currently offer these therapies to patients. Which clinical trials are presently being offered, and what is the status of unregulated stem cell therapy?

   Begin your search with information from the International Society for Stem Cell Research: http://www.closerlookatstemcells.org.

CASE STUDY | A case of short thumbs and toes

A doctor received a female patient with unusually short thumbs and great toes, small feet, short fifth fingers with clinodactyly (bending towards fourth fingers), and a duplication of the genital tract that led to urinary problems. He diagnosed her with hand-foot-genital syndrome (HFGS). The sequencing of her DNA revealed a polyalanine expansion in exon 1 of *HOXA13*, the most 5′ gene in the *HOXA* cluster on chromosome 7p15.2. The same mutation was found in her father, who had short thumbs and great toes but no genital abnormalities. HFGS is inherited in an autosomal dominant fashion, and is caused by the haploinsufficiency of *HOXA13* transcription factor.

1. How do you think a haploinsufficiency of *HOXA13* transcription factor during embryonic development would lead to the phenotypes observed in the patient with HFGS?
2. What is the equivalent gene of *HOXA13* in *Drosophila*? What does it do?
3. In humans, what other homeobox genes have spatial expression patterns similar to those of *HOXA13*?

INSIGHTS AND SOLUTIONS

1. In the slime mold *Dictyostelium*, experimental evidence suggests that cyclic AMP (cAMP) plays a central role in the developmental program leading to spore formation. The genes encoding the cAMP cell-surface receptor have been cloned, and the amino acid sequence of the protein components is known. To form reproductive structures, free-living individual cells aggregate together and then differentiate into one of two cell types, prespore cells or prestalk cells. Aggregating cells secrete waves or oscillations of cAMP to foster the aggregation of cells and then continuously secrete cAMP to activate genes in the aggregated cells at later stages of development. It has been proposed that cAMP controls cell—cell interaction and gene expression. It is important to test this hypothesis by using several experimental techniques. What different approaches can you devise to test this hypothesis, and what specific experimental systems would you employ to test them?

   Solution: Two of the most powerful forms of analysis in biology involve the use of biochemical analogs (or inhibitors) to block gene transcription or the action of gene products in a predictable way, and the use of mutations to alter genes and their products. These two approaches can be used to study the role of cAMP in the developmental program of *Dictyostelium*. First, compounds chemically related to cAMP, such as GTP and GDP, can be used to test whether they have any effect on the processes controlled by cAMP. In fact, both GTP and GDP lower the affinity of cell-surface receptors for cAMP, effectively blocking the action of cAMP.

   Mutational analysis can be used to dissect components of the cAMP receptor system. One approach is to use transformation with wild-type genes to restore mutant function. Similarly, because the genes for the receptor proteins have been cloned, it is possible to construct mutants with known alterations in the component proteins and transform them into cells to assess their effects.

2. In the sea urchin, early development may occur even in the presence of actinomycin D, which inhibits RNA synthesis. However, if actinomycin D is present early in development but is removed a few hours later, all development stops. In fact, if actinomycin D is present only between the sixth and eleventh hours of development, events that normally occur at the fifteenth hour are arrested. What conclusions can be drawn concerning the role of gene transcription between hours 6 and 15?

Solution: Maternal mRNAs are present in the fertilized sea urchin egg. Thus, a considerable amount of development can take place without transcription of the embryo's genome. Because development past 15 hours is inhibited by prior treatment with actinomycin D, it appears that transcripts from the embryo's genome are required to initiate or maintain these events. This transcription must take place between the sixth and fifteenth hours of development.

3. If it were possible to introduce one of the homeotic genes from *Drosophila* into an *Arabidopsis* embryo homozygous for a homeotic flowering gene, would you expect any of the *Drosophila* genes to negate (rescue) the *Arabidopsis* mutant phenotype? Why or why not?

Solution: The *Drosophila* homeotic genes belong to the *Hox* gene family, whereas *Arabidopsis* homeotic genes belong to the MADS-box protein family. Both gene families are present in *Drosophila* and *Arabidopsis*, but they have evolved different functions in the animal and the plant kingdoms. As a result, it is unlikely that a transferred *Drosophila Hox* gene would rescue the phenotype of a MADS-box mutant, but only an actual experiment would confirm this.

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we have focused on large-scale as well as the inter- and intracellular events that take place during embryogenesis and the formation of adult structures. In particular, we discussed how the adult body plan is laid down by a cascade of gene expression, and the role of cell-cell communication in development. Based on your knowledge of these topics, answer several fundamental questions:
(a) How do we know how many genes control development in an organism like *Drosophila*?
(b) What experimental evidence demonstrates that molecular gradients in the egg control development?
(c) How did we discover that selector genes specify which adult structures will be formed by body segments?
(d) How did we learn about the levels of gene regulation involved in vulval development in *C. elegans*?

CONCEPT QUESTION

2. Review the Chapter Concepts list on page 419. Most of these are concerned with the cascade of gene transcription that converts a zygote into an adult organism. Write a short essay outlining the differences and similarities in the gene families used by plants and animals to establish the body axis and to regulate gene expression of these gene sets. ■
3. Nuclei from almost any source may be injected into *Xenopus* oocytes. Studies have shown that these nuclei remain active in transcription and translation. How can such an experimental system be useful in developmental genetic studies?
4. Distinguish between the syncytial blastoderm stage and the cellular blastoderm stage in *Drosophila* embryogenesis.
5. (a) What are maternal-effect genes? (b) When are gene products from these genes made, and where are they located? (c) What aspects of development do maternal-effect genes control? (d) What is the phenotype of maternal-effect mutations?
6. How are the zygotic genes influenced by the maternal genes? What would happen if you mutate the zygotic genes?
7. List the main classes of zygotic genes. What is the function of each class of these genes?
8. Experiments have shown that any nuclei placed in the polar cytoplasm at the posterior pole of the *Drosophila* egg will differentiate into germ cells. If polar cytoplasm is transplanted into the anterior end of the egg just after fertilization, what will happen to nuclei that migrate into this cytoplasm at the anterior pole?
9. How can you determine whether a particular gene is being transcribed in different cell types?
10. You observe that a particular gene is being transcribed in different cell types during development. You now want to assess whether protein is being made in these cells or not. Suggest an experiment for this.
11. What is the primary function of *Hox* genes?
12. The homeotic mutation *Antennapedia* causes mutant *Drosophila* to have legs in place of antennae and is a dominant gain-of-function mutation. What are the properties of such mutations? How does the *Antennapedia* gene change antennae into legs?
13. The *Drosophila* homeotic mutation *spineless aristapedia* (ss^a) results in the formation of a miniature tarsal structure (normally part of the leg) on the end of the antenna. What insight is provided by (ss^a) concerning the role of genes during determination?
14. A number of genes that control expression of *Hox* genes in *Drosophila* have been identified. One of these homozygous mutants is *extra sex combs,* where some of the head and all of the thorax and abdominal segments develop as the last abdominal segment. In other words, all affected segments develop as posterior segments. What does this phenotype tell you about which set of *Hox* genes is controlled by the *extra sex combs* gene?
15. In *Arabidopsis,* flower development is controlled by sets of homeotic genes. How many classes of these genes are there, and what structures are formed by their individual and combined expression?
16. The floral homeotic genes of *Arabidopsis* belong to the MADS-box gene family, while in *Drosophila,* homeotic genes belong to the homeobox gene family. In both *Arabidopsis* and *Drosophila,* members of the *Polycomb* gene family control expression of these divergent homeotic genes. How do *Polycomb* genes control expression of two very different sets of homeotic genes?
17. Dominguez et al. (2004) suggest that by studying genes that determine growth and tissue specification in the eye of *Drosophila*, much can be learned about human eye development.
(a) What evidence suggests that genetic eye determinants in *Drosophila* are also found in humans? Include a discussion of orthologous genes in your answer.
(b) What evidence indicates that the *eyeless* gene is part of a developmental network?
(c) Are genetic networks likely to specify developmental processes in general? Explain fully and provide an example.

21 Quantitative Genetics and Multifactorial Traits

CHAPTER CONCEPTS

- Quantitative inheritance results in a range of measurable phenotypes for a polygenic trait.
- Polygenic traits most often demonstrate continuous variation.
- Quantitative inheritance can be explained in Mendelian terms whereby certain alleles have an additive effect on the traits under study.
- The study of polygenic traits relies on statistical analysis.
- Heritability values estimate the genetic contribution to phenotypic variability under specific environmental conditions.
- Twin studies allow an estimation of heritability in humans.
- Quantitative trait loci (QTLs) can be mapped and identified.

A field of pumpkins, where size is under the influence of quantitative inheritance.

Up to this point in the text, most of our examples of phenotypic variation have been those that have been assigned to distinct and separate categories; for example, human blood type was A, B, AB, or O; squash fruit shape was spherical, disc-shaped, or elongated; and fruit fly eye color was red or white (see Chapter 4). Typically in these traits, a genotype will produce a single identifiable phenotype, although phenomena such as variable penetrance and expressivity, pleiotropy, and epistasis can obscure the relationship between genotype and phenotype.

However, many traits are not as distinct and clear cut, including many that are of medical or agricultural importance. They show much more variation, often falling into a continuous range of multiple phenotypes. Most show what we call *continuous variation,* including, for example, height in humans, milk and meat production in cattle, and yield and seed protein content in various crops. Continuous variation across a range of phenotypes can be measured and described in quantitative terms, so this genetic phenomenon is known as **quantitative inheritance.** And because the varying phenotypes result from the input of genes at more than one, and often many, loci, the traits are said to be **polygenic** (literally "of many genes"). The genes involved are often referred to as **polygenes.**

To further complicate the link between the genotype and phenotype, the genotype generated at fertilization establishes a quantitative range within which a particular individual can fall. However, the final phenotype is often also influenced by environmental factors to which that individual is

exposed. Human height, for example, is genetically influenced but is also affected by environmental factors such as nutrition. Quantitative (polygenic) traits whose phenotypes result from both gene action and environmental influences are termed **multifactorial,** or **complex traits.** Often these terms are used interchangeably. For consistency throughout the chapter, we will utilize the term *multifactorial* in our discussions.

In this chapter, we will examine examples of quantitative inheritance, multifactorial traits, and some of the statistical techniques used to study them. We will also consider how geneticists assess the relative importance of genetic versus environmental factors contributing to continuous phenotypic variation, and we will discuss approaches to identifying and mapping genes that influence quantitative traits.

21.1 Quantitative Traits Can Be Explained in Mendelian Terms

The question of whether continuous phenotypic variation could be explained in Mendelian terms caused considerable controversy in the early 1900s. Some scientists argued that, although Mendel's unit factors, or genes, explained patterns of discontinuous variation with discrete phenotypic classes, they could not account for the range of phenotypes seen in quantitative patterns of inheritance. However, geneticists William Bateson and G. Udny Yule, adhering to a Mendelian explanation, proposed the **multiple-factor** or **multiple-gene hypothesis,** in which many genes, each individually behaving in a Mendelian fashion, contribute to the phenotype in a *cumulative* or *quantitative* way.

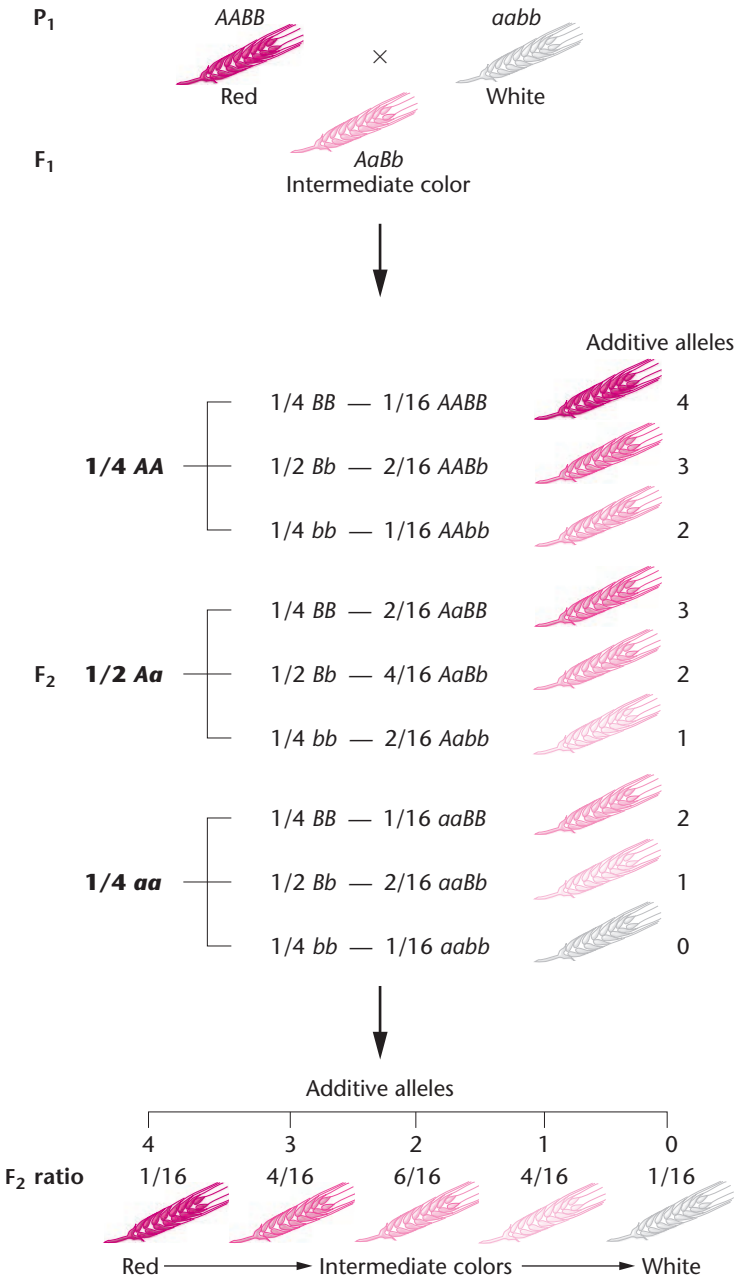


FIGURE 21–1 How the multiple-factor hypothesis accounts for the 1:4:6:4:1 phenotypic ratio of grain color when all alleles designated by an uppercase letter are additive and contribute an equal amount of pigment to the phenotype.

The Multiple-Gene Hypothesis for Quantitative Inheritance

The **multiple-gene hypothesis** was initially based on a key set of experimental results published by Hermann Nilsson-Ehle in 1909. Nilsson-Ehle used grain color in wheat to test the concept that the cumulative effects of alleles at multiple loci produce the range of phenotypes seen in quantitative traits. In one set of experiments, wheat with red grain was crossed to wheat with white grain (**Figure 21–1**). The F_1 generation demonstrated an intermediate pink color, which at first sight suggested incomplete dominance of two alleles at a single locus. However, in the F_2 generation, Nilsson-Ehle did not observe the typical segregation of a monohybrid cross. Instead, approximately 15/16 of the plants showed some degree of red grain color, while 1/16 of the plants showed white grain color. Careful examination of the F_2 revealed that grain with color could be classified into four different shades of red. Because the F_2 ratio occurred in sixteenths, it appears that two genes, each with two alleles, control the phenotype and that they segregate independently from one another in a Mendelian fashion.

If each gene has one potential **additive allele** that contributes approximately equally to the red grain color and one potential **nonadditive allele** that fails to produce any red pigment, we can see how the multiple-factor hypothesis could account for the various grain color phenotypes. In the P_1 both parents are homozygous; the red parent contains only additive alleles (*AABB* in Figure 21–1), while the white parent contains only nonadditive alleles (*aabb*). The F_1 plants are heterozygous (*AaBb*), contain two additive (*A* and *B*) and two nonadditive (*a* and *b*) alleles, and express the intermediate pink phenotype. Each of the F_2 plants has 4, 3, 2, 1, or 0 additive alleles. F_2 plants with no additive alleles are white (*aabb*) like one of the P_1 parents, while F_2 plants with 4 additive alleles are red (*AABB*) like the other P_1 parent. Plants with 3, 2, or 1 additive alleles constitute the other three categories of red color observed in the F_2 generation. The greater the number of additive alleles in the genotype, the more intense the red color expressed in the phenotype, as each additive allele present contributes equally to the cumulative amount of pigment produced in the grain.

Nilsson-Ehle's results showed how continuous variation could still be explained in a Mendelian fashion, with additive alleles at multiple loci influencing the phenotype in a quantitative manner, but each individual allele segregating according to Mendelian rules. As we saw in Nilsson-Ehle's initial cross, if two loci, each with two alleles, were involved, then five F_2 phenotypic categories in a 1:4:6:4:1 ratio would be expected. However, there is no reason why three, four, or more loci cannot function in a similar fashion in controlling various quantitative phenotypes. As more quantitative loci become involved, greater and greater numbers of classes appear in the F_2 generation in more complex ratios. The number of phenotypes and the expected F_2 ratios for crosses involving up to five gene pairs are illustrated in **Figure 21–2**.

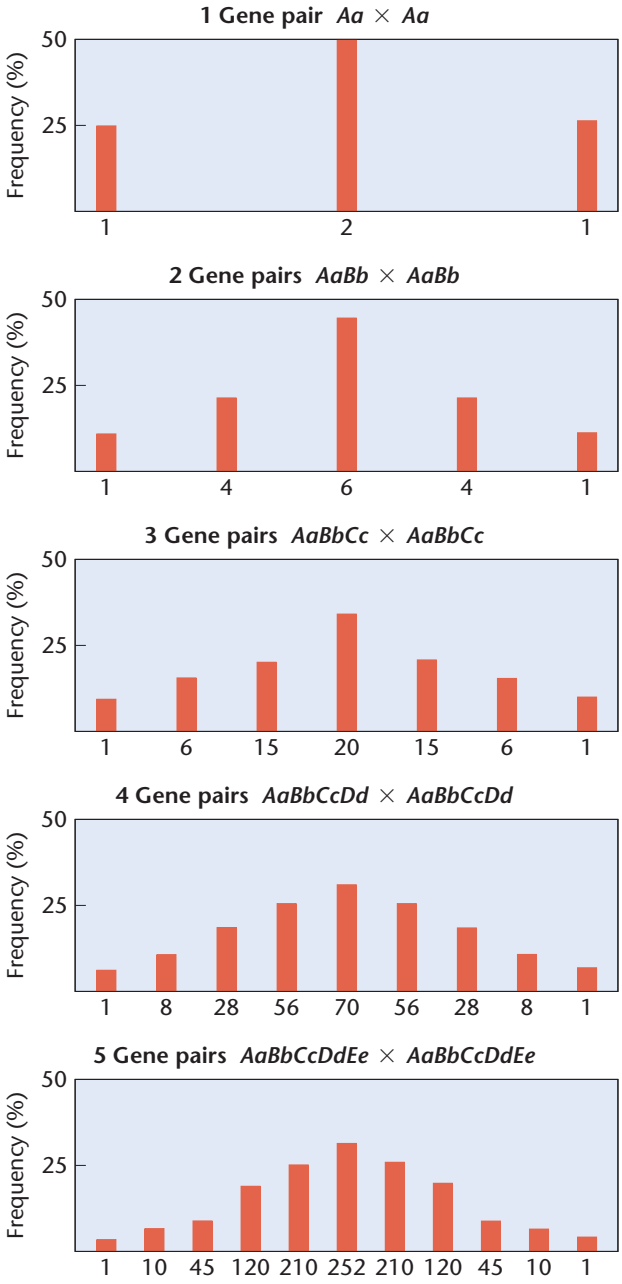


FIGURE 21–2 The genetic ratios (on the X-axis) resulting from crossing two heterozygotes when polygenic inheritance is in operation with 1 to 5 gene pairs. The histogram bars indicate the distinct F_2 phenotypic classes, ranging from one extreme (left end) to the other extreme (right end). Each phenotype results from a different number of additive alleles.

Additive Alleles: The Basis of Continuous Variation

The multiple-gene hypothesis consists of the following major points:

1. Phenotypic traits showing continuous variation can be quantified by measuring, weighing, counting, and so on.
2. Two or more gene loci, often scattered throughout the genome, account for the hereditary influence on the phenotype in an *additive way.* Because many genes may be involved, inheritance of this type is called *polygenic.*
3. Each gene locus may be occupied by either an *additive* allele, which contributes a constant amount to the phenotype, or a *nonadditive* allele, which does not contribute quantitatively to the phenotype.
4. The contribution to the phenotype of each additive allele, though often small, is approximately equal. While we now know this is not always true, we have made this assumption in the above discussion.
5. Together, the additive alleles contributing to a single quantitative character produce substantial phenotypic variation.

Calculating the Number of Polygenes

Various formulas have been developed for estimating the number of polygenes contributing to a quantitative trait. For example, if the ratio of F_2 individuals resembling *either* of the two extreme P_1 phenotypes can be determined, the number of polygenes (loci) involved (n) may be calculated as

$$1/4^n = \text{ratio of } F_2 \text{ individuals expressing either extreme phenotype}$$

In the example of the red and white wheat grain color summarized in Figure 21–1, 1/16 of the progeny are either red *or* white like the P_1 phenotypes. This ratio can be substituted on the right side of the equation to solve for n:

$$\frac{1}{4^n} = \frac{1}{16}$$
$$\frac{1}{4^2} = \frac{1}{16}$$
$$n = 2$$

Table 21.1 lists the ratio and the number of F_2 phenotypic classes produced in crosses involving up to five gene pairs.

For low numbers of polygenes (n), it is sometimes easier to use the equation

$$(2n + 1) = \text{the number of distinct phenotypic categories observed}$$

For example, when there are two polygenes involved ($n = 2$), then $(2n + 1) = 5$ and each phenotype is the result of 4, 3, 2, 1, or 0 additive alleles. If $n = 3$, $2n + 1 = 7$ and each phenotype is the result of 6, 5, 4, 3, 2, 1, or 0 additive alleles. Thus, working backwards with this rule and knowing the number of phenotypes, we can calculate the number of polygenes controlling them.

It should be noted, however, that both of these simple methods for estimating the number of polygenes involved in a quantitative trait assume not only that all the relevant alleles contribute equally and additively, but also that phenotypic expression in the F_2 is not affected significantly by environmental factors. As we will see later, for many quantitative traits, these assumptions may not be true.

TABLE 21.1 Determination of the Number of Polygenes (n) Involved in a Quantitative Trait

n	Individuals Expressing Either Extreme Phenotype	Distinct Phenotypic Classes
1	1/4	3
2	1/16	5
3	1/64	7
4	1/256	9
5	1/1024	11

ESSENTIAL POINT

Quantitative inheritance results in a range of phenotypes due to the action of additive alleles from two or more genes, as influenced by environmental factors. ■

NOW SOLVE THIS

21–1 A homozygous plant with 20-cm diameter flowers is crossed with a homozygous plant of the same species that has 40-cm diameter flowers. The F_1 plants all have flowers 30 cm in diameter. In the F_2 generation of 512 plants, 2 plants have flowers 20 cm in diameter, 2 plants have flowers 40 cm in diameter, and the remaining 508 plants have flowers of a range of sizes in between.

(a) Assuming that all alleles involved act additively, how many genes control flower size in this plant?

(b) What frequency distribution of flower diameter would you expect to see in the progeny of a backcross between an F_1 plant and the large-flowered parent?

■ **HINT:** *This problem provides F_1 and F_2 data for a cross involving a quantitative trait and asks you to calculate the number of genes controlling the trait. The key to its solution is to remember that unless you know the total number of distinct F_2 phenotypes involved, then the ratio (not the number) of parental phenotypes reappearing in the F_2 must be used in your determination of the number of genes involved.*

21.2 The Study of Polygenic Traits Relies on Statistical Analysis

Before considering the approaches that geneticists use to dissect how much of the phenotypic variation observed in a population is due to genotypic differences among individuals and how much is due to environmental factors, we need to consider the basic statistical tools they use for the task. It is not usually feasible to measure expression of a polygenic trait in every individual in a population, so a random subset of individuals is usually selected for measurement to provide a *sample*. It is important to remember that the accuracy of the final results of the measurements depends on whether the sample is truly random and representative of the population from which it was drawn. Suppose, for example, that a student wants to determine the average height of the 100 students in his genetics class, and for his sample he measures the two students sitting next to him, both of whom happen to be centers on the college basketball team. It is unlikely that this sample will provide a good estimate of the average height of the class, for two reasons: First, it is too small; second, it is not a representative subset of the class (unless all 100 students are centers on the basketball team).

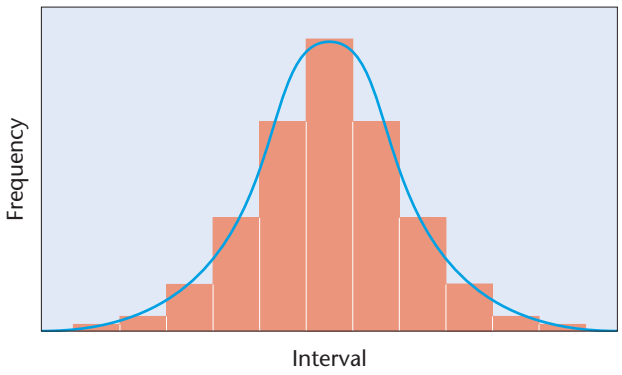


FIGURE 21–3 Normal frequency distribution, characterized by a bell-shaped curve.

If the sample measured for expression of a quantitative trait is sufficiently large and also representative of the population from which it is drawn, we often find that the data form a **normal distribution;** that is, they produce a characteristic bell-shaped curve when plotted as a frequency histogram (**Figure 21–3**). Several statistical concepts are useful in the analysis of traits that exhibit a normal distribution, including the mean, variance, standard deviation, standard error of the mean, covariance, and correlation coefficient.

The Mean

The mean provides information about where the central point lies along a range of measurements for a quantitative trait. **Figure 21–4** shows the distribution curves for two different sets of phenotypic measurements. Each of these sets of measurements clusters around a central value (as it happens, they both cluster around the same value). This clustering is called a *central tendency*, and the central point is the *mean*.

Specifically, the **mean** ($\overline{X}$) is the arithmetic average of a set of measurements and is calculated as

$$\overline{X} = \frac{\Sigma X_i}{n}$$

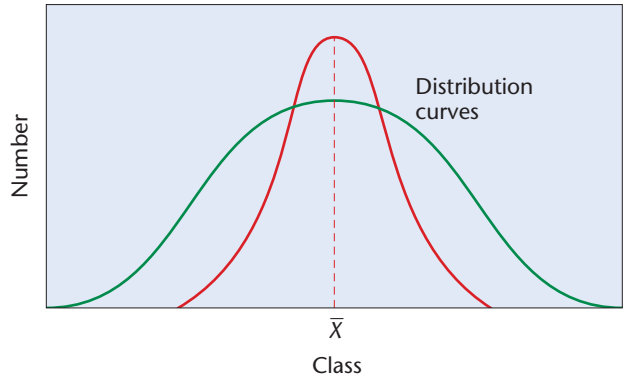


FIGURE 21–4 Two normal frequency distributions with the same mean but different amounts of variation.

where $\overline{X}$ is the mean, ΣX_i represents the sum of all individual values in the sample, and n is the number of individual values.

The mean provides a useful descriptive summary of the sample, but it tells us nothing about the range or spread of the data. As illustrated in Figure 21–4, a symmetrical distribution of values in the sample may, in one case, be clustered near the mean. Or a set of measurements may have the same mean but be distributed more widely around it. A second statistic, the variance, provides information about the spread of data around the mean.

Variance

The **variance** (s^2) for a sample is the average squared distance of all measurements from the mean. It is calculated as

$$s^2 = \frac{\Sigma\,(X_i - \overline{X})^2}{n - 1}$$

where the sum (Σ) of the squared differences between each measured value (X_i) and the mean ($\overline{X}$) is divided by one less than the total sample size ($n - 1$).

As Figure 21–4 shows, it is possible for two sets of sample measurements for a quantitative trait to have the same mean but a different distribution of values around it. This range will be reflected in different variances. Estimation of variance can be useful in determining the degree of genetic control of traits when the immediate environment also influences the phenotype.

Standard Deviation

Because the variance is a squared value, its unit of measurement is also squared (m^2, g^2, etc.). To express variation around the mean in the original units of measurement, we can use the square root of the variance, a term called the **standard deviation** (s)

$$s = \sqrt{s^2}$$

Table 21.2 shows the percentage of individual values within a normal distribution that fall within different multiples of the standard deviation. The values that fall within one standard deviation to either side of the mean represent

TABLE 21.2 Sample Inclusion for Various *s* Values

Multiples of *s*	Sample Included (%)
$\overline{X} \pm 1s$	68.3
$\overline{X} \pm 1.96s$	95.0
$\overline{X} \pm 2s$	95.5
$\overline{X} \pm 3s$	99.7

68 percent of all values in the sample. More than 95 percent of all values are found within two standard deviations to either side of the mean. This indicates that the standard deviation (s) can also be interpreted in the form of a probability. For example, a sample measurement picked at random has a 68 percent probability of falling within the range of one standard deviation.

Standard Error of the Mean

If multiple samples are taken from a population and measured for the same quantitative trait, we might find that their means vary. Theoretically, larger, truly random samples will represent the population more accurately, and their means will be closer to each other. To measure the accuracy of the sample mean, we use the **standard error of the mean** ($S_{\overline{X}}$), calculated as

$$S_{\overline{X}} = \frac{s}{\sqrt{n}}$$

where s is the standard deviation and $\sqrt{n}$ is the square root of the sample size. Because the standard error of the mean is computed by dividing s by $\sqrt{n}$, it is always a smaller value than the standard deviation.

Covariance and Correlation Coefficient

Often geneticists working with quantitative traits find they have to consider two phenotypic characters simultaneously. For example, a poultry breeder might investigate the correlation between body weight and egg production in hens: Do heavier birds tend to lay more eggs? The covariance statistic measures how much variation is common to both quantitative traits. It is calculated by taking the deviations from the mean for each trait (just as we did for estimating variance) for each individual in the sample. This gives a pair of values for each individual. The two values are multiplied together, and the sum of all these individual products is then divided by one fewer than the number in the sample. Thus, the **covariance** ($\mathbf{cov}_{XY}$) of two sets of trait measurements, X and Y, is calculated as

$$\text{cov}_{XY} = \frac{\Sigma\left[(X_i - \overline{X})(Y_i - \overline{Y})\right]}{n - 1}$$

The covariance can then be standardized as yet another statistic, the **correlation coefficient (r).** The calculation is

$$r = \text{cov}_{XY}/S_X S_Y$$

where S_X is the standard deviation of the first set of quantitative measurements X, and S_Y is the standard deviation of the second set of quantitative measurements Y. Values for the correlation coefficient r can range from -1 to $+1$. Positive r values mean that an increase in measurement for one trait tends to be associated with an increase in measurement for the other, while negative r values mean that increases in one trait are associated with decreases in the other. Therefore, if heavier hens do tend to lay more eggs, a positive r value can be expected. A negative r value, on the other hand, suggests that greater egg production is more likely from less heavy birds. One important point to note about correlation coefficients is that even significant r values—close to $+1$ or -1—do not prove that a cause-and-effect relationship exists between two traits. Correlation analysis simply tells us the extent to which variation in one quantitative trait is associated with variation in another, not what causes that variation.

Analysis of a Quantitative Character

To apply these statistical concepts, let's consider a genetic experiment that crossed two different homozygous varieties of tomato. One of the tomato varieties produces fruit averaging 18 oz in weight, whereas fruit from the other averages 6 oz. The F_1 obtained by crossing these two varieties has fruit weights ranging from 10 to 14 oz. The F_2 population contains individuals that produce fruit ranging from 6 to 18 oz. The results characterizing both generations are shown in Table 21.3.

The mean value for the fruit weight in the F_1 generation can be calculated as

$$\overline{X} = \frac{\Sigma X_i}{n} = \frac{626}{52} = 12.04$$

The mean value for fruit weight in the F_2 generation is calculated as

$$\overline{X} = \frac{\Sigma X_i}{n} = \frac{872}{72} = 12.11$$

Although these mean values are similar, the frequency distributions in Table 21.3 show more variation in the

TABLE 21.3 Distribution of F_1 and F_2 Progeny Derived from a Theoretical Cross Involving Tomatoes

		Weight (oz)												
		6	7	8	9	10	11	12	13	14	15	16	17	18
Number of	F_1					4	14	16	12	6				
Individuals	F_2	1	1	2	0	9	13	17	14	7	4	3	0	1

F_2 generation. The range of variation can be quantified as the sample variance s^2, calculated, as we saw on page 442, as the sum of the squared differences between each value and the mean, divided by one less than the total number of observations:

$$s^2 = \frac{\Sigma (X_i - \overline{X})^2}{n - 1}$$

When the above calculation is made, the variance is found to be 1.29 for the F_1 generation and 4.27 for the F_2 generation. When converted to the standard deviation ($s = \sqrt{s^2}$), the values become 1.13 and 2.06, respectively. Therefore, the distribution of tomato weight in the F_1 generation can be described as 12.04 ± 1.13, and in the F_2 generation it can be described as 12.11 ± 2.06.

Assuming that both parental varieties are homozygous at the loci of interest and that the alleles controlling fruit weight act additively, we can estimate the number of loci involved in this trait. Since 1/72 of the F_2 offspring have a phenotype that overlaps one of the parental strains (72 total F_2 offspring; one weighs 6 oz, one weighs 18 oz; see Table 21.3), the use of the formula $1/4^n = 1/72$ indicates that *n* is between 3 and 4, providing evidence of the number of genes that control fruit weight in these tomato strains.

NOW SOLVE THIS

21–2 The following table shows measurements for fiber lengths and fleece weight in a small flock of eight sheep.

	Sheep Fiber Length (cm)	Fleece Weight (kg)
1	9.7	7.9
2	5.6	4.5
3	10.7	8.3
4	6.8	5.4
5	11.0	9.1
6	4.5	4.9
7	7.4	6.0
8	5.9	5.1

(a) What are the mean, variance, and standard deviation for each trait in this flock?
(b) What is the covariance of the two traits?
(c) What is the correlation coefficient for fiber length and fleece weight?
(d) Do you think greater fleece weight is correlated with an increase in fiber length? Why or why not?

■ **HINT:** *This problem provides data for two quantitative traits and asks you to make numerous statistical calculations, ultimately determining if the traits are correlated. The key to its solution is that once the calculation of the correlation coefficient (r) is completed, you must interpret that value—whether it is positive or negative, and how close to zero it is.*

ESSENTIAL POINT

Numerous statistical methods are essential during the analysis of quantitative traits, including the mean, variance, standard deviation, standard error, covariance, and the correlation coefficient. ■

21.3 Heritability Values Estimate the Genetic Contribution to Phenotypic Variability

The question most often asked by geneticists working with multifactorial traits and diseases is how much of the observed phenotypic variation in a population is due to genotypic differences among individuals and how much is due to environment. The term **heritability** is used to describe *the proportion of total phenotypic variation in a population that is due to genetic factors.* For a multifactorial trait in a given population, a high heritability estimate indicates that much of the variation can be attributed to genetic factors, with the environment having less impact on expression of the trait. With a low heritability estimate, environmental factors are likely to have a greater impact on phenotypic variation within the population.

The concept of heritability is frequently misunderstood and misused. It should be emphasized that *heritability indicates neither how much of a trait is genetically determined nor the extent to which an individual's phenotype is due to genotype.* In recent years, such misinterpretations of heritability for human quantitative traits have led to controversy, notably in relation to measurements such as intelligence quotients, or IQs. Variation in heritability estimates for IQ among different racial groups led to incorrect suggestions that unalterable genetic factors control differences in intelligence levels among humans of different ancestries. Such suggestions misrepresented the meaning of heritability and ignored the contribution of genotype-by-environment interaction variance (see p. 445) to phenotypic variation in a population. Moreover, heritability is not fixed for a trait. For example, a heritability estimate for egg production in a flock of chickens kept in individual cages might be high, indicating that differences in egg output among individual birds are largely due to genetic differences, as they all have very similar environments. For a different flock kept outdoors, heritability for egg production might be much lower, as variation among different birds may also reflect differences in their individual environments. Such differences could include how much food each bird manages to find and whether it competes successfully for a good roosting spot at night.

Thus, a heritability estimate tells us the proportion of *phenotypic variation* that can be attributed to *genetic*

variation within a certain population in a particular environment. If we measure heritability for the same trait among different populations in a range of environments, we frequently find that the calculated heritability values have large standard errors. This is an important point to remember when considering heritability estimates. Parallel studies using different population bases are likely to yield different heritability estimates. For example, a mean heritability estimate of 0.65 for human height does not mean that your height is 65 percent due to your genes, but rather that in the populations sampled, on average, *65 percent of the overall variation in height could be explained by genotypic differences among individuals.*

With this subtle but important distinction in mind, we will now consider how geneticists divide the phenotypic variation observed in a population into genetic and environmental components. As we saw in the previous section, variation can be quantified as a sample variance: taking measurements of the trait in question from a representative sample of the population and determining the extent of the spread of those measurements around the sample mean. This gives us an estimate of the total **phenotypic variance** in the population (V_P). Heritability estimates are obtained by using different experimental and statistical techniques to partition V_P into **genotypic variance** (V_G) and **environmental variance** (V_E) components.

An important factor contributing to overall levels of phenotypic variation is the extent to which individual genotypes affect the phenotype differently depending on the environment. For example, wheat variety A may yield an average of 20 bushels an acre on poor soil, while variety B yields an average of 17 bushels. On good soil, variety A yields 22 bushels, while variety B averages 25 bushels an acre. There are differences in yield between the two genotypically distinct varieties, so variation in wheat yield has a genetic component. Both varieties yield more on good soil, so yield is also affected by environment. However, we also see that the two varieties do not respond to better soil conditions equally: The genotype of wheat variety B achieves a greater increase in yield on good soil than does variety A. Thus, we have differences in the interaction of genotype, with environment contributing to variation for yield in populations of wheat plants. This third component of phenotypic variation is **genotype-by-environment interaction variance ($V_{G\times E}$)** (Figure 21–5).

We can now summarize all the components of total phenotypic variance V_P using the following equation:

$$V_P = V_G + V_E + V_{G\times E}$$

In other words, total phenotypic variance can be subdivided into genotypic variance, environmental variance, and genotype-by-environment interaction variance. When obtaining heritability estimates for a multifactorial trait, researchers often assume that the genotype-by-environment interaction variance is small enough that it can be ignored or combined with the environmental variance. However, it is worth remembering that this kind of approximation is another reason heritability values are *estimates* for a given population in a particular context, not a *fixed attribute* for a trait.

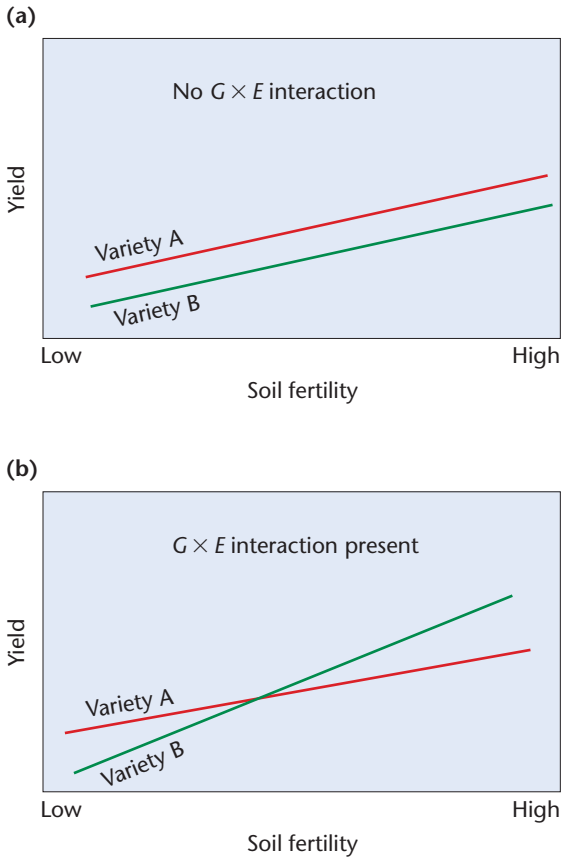


FIGURE 21–5 Differences in yield between two wheat varieties at different soil fertility levels. (a) No genotype-by-environment, or $G \times E$, interaction: The varieties show genetic differences in yield but respond equally to increasing soil fertility. (b) $G \times E$ interaction present: Variety A outyields B at low soil fertility, but B yields more than A at high-fertility levels.

Animal and plant breeders use a range of experimental techniques to estimate heritabilities by partitioning measurements of phenotypic variance into genotypic and environmental components. One approach uses inbred strains containing genetically homogeneous individuals with highly homozygous genotypes. Experiments are then designed to test the effects of a range of environmental conditions on phenotypic variability. Variation *between* different inbred strains reared in a constant environment is due predominantly to genetic factors. Variation *among* members of the same inbred strain reared under different conditions is more likely to be due to environmental factors. Other approaches involve analysis of variance for a

quantitative trait among offspring from different crosses, or comparing expression of a trait among offspring and parents reared in the same environment.

Broad-Sense Heritability

Broad-sense heritability (represented by the term H^2) measures the contribution of the genotypic variance to the total phenotypic variance. It is estimated as a proportion:

$$H^2 = \frac{V_G}{V_P}$$

Heritability values for a trait in a population range from 0.0 to 1.0. A value approaching 1.0 indicates that the environmental conditions have little impact on phenotypic variance, which is therefore largely due to genotypic differences among individuals in the population. Low values close to 0.0 indicate that environmental factors, not genotypic differences, are largely responsible for the observed phenotypic variation within the population studied. Few quantitative traits have very high or very low heritability estimates, suggesting that both genetics and environment play a part in the expression of most phenotypes for the trait.

The genotypic variance component V_G used in broad-sense heritability estimates includes all types of genetic variation in the population. It does not distinguish between quantitative trait loci with alleles acting additively as opposed to those with epistatic or dominance effects. Broad-sense heritability estimates also assume that the genotype-by-environment variance component is negligible. While broad-sense heritability estimates for a trait are of general genetic interest, these limitations mean this kind of heritability is not very useful in breeding programs. Animal or plant breeders wishing to develop improved strains of livestock or higher-yielding crop varieties need more precise heritability estimates for the traits they wish to manipulate in a population. Therefore, another type of estimate, narrow-sense heritability, has been devised that is of more practical use.

Narrow-Sense Heritability

Narrow-sense heritability (h^2) is the proportion of phenotypic variance due to additive genotypic variance alone. Genotypic variance can be divided into subcomponents representing the different modes of action of alleles at quantitative trait loci. As not all the genes involved in a quantitative trait affect the phenotype in the same way, this partitioning distinguishes among three different kinds of gene action contributing to genotypic variance. **Additive variance, V_A,** is the genotypic variance due to the additive action of alleles at quantitative trait loci. **Dominance variance, V_D,** is the deviation from the additive components that results when phenotypic expression in heterozygotes is not precisely intermediate between the two homozygotes. **Interactive variance, V_I,** is the deviation from the additive components that occurs when two or more loci behave epistatically. The amount of interactive variance is often negligible, and so this component is often excluded from calculations of total genotypic variance.

The partitioning of the total genotypic variance V_G is summarized in the equation

$$V_G = V_A + V_D + V_I$$

and a narrow-sense heritability estimate based only on that portion of the genotypic variance due to additive gene action becomes

$$h^2 = \frac{V_A}{V_P}$$

Omitting V_I and separating V_P into genotypic and environmental variance components, we obtain

$$h^2 = \frac{V_A}{V_E + V_A + V_D}$$

Heritability estimates are used in animal and plant breeding to indicate the potential response of a population to artificial selection for a quantitative trait. Narrow-sense heritability, h^2, provides a more accurate prediction of selection response than broad-sense heritability, H^2, and therefore h^2 is more widely used by breeders.

ESSENTIAL POINT

Heritability is an estimate of the relative contribution of genetic versus environmental factors to the range of phenotypic variation seen in a quantitative trait in a particular population and environment. ■

Artificial Selection

Artificial selection is the process of choosing specific individuals with preferred phenotypes from an initially heterogeneous population for future breeding purposes. Theoretically, if artificial selection based on the same trait preferences is repeated over multiple generations, a population can be developed containing a high frequency of individuals with the desired characteristics. If selection is for a simple trait controlled by just one or two genes subject to little environmental influence, generating the desired population of plants or animals is relatively fast and easy. However, many traits of economic importance in crops and livestock, such as grain yield in plants, weight gain or milk yield in cattle, and speed or stamina in horses, are polygenic and frequently multifactorial. Artificial selection for such traits is slower and more complex. Narrow-sense heritability estimates are valuable to the plant or animal breeder because, as we have just seen, they estimate the

proportion of total phenotypic variance for the trait that is due to additive genetic variance. Quantitative trait alleles with additive impact are those most easily manipulated by the breeder. Alleles at quantitative trait loci that generate dominance effects or interact epistatically (and therefore contribute to V_D or V_I) are less responsive to artificial selection. Thus, narrow-sense heritability, h^2, can be used to predict the impact of selection. The higher the estimated value for h^2 in a population, the more likely the breeder will observe a change in phenotypic range for the trait in the next generation after artificial selection.

Partitioning the genetic variance components to calculate h^2 and predict response to selection is a complex task requiring careful experimental design and analysis. The simplest approach is to select individuals with superior phenotypes for the desired quantitative trait from a heterogeneous population and breed offspring from those individuals. The mean score for the trait of those offspring ($M2$) can then be compared to that of: (1) the original population's mean score (M) and (2) the selected individuals used as parents ($M1$). The relationship between these means and h^2 is

$$h^2 = \frac{M2 - M}{M1 - M}$$

This equation can be further simplified by defining $M2 - M$ as the **selection response** (R)—the degree of response to mating the selected parents—and $M1 - M$ as the **selection differential** (S)—the difference between the mean for the whole population and the mean for the selected population—so h^2 reflects the ratio of the response observed to the total response possible. Thus,

$$h^2 = \frac{R}{S}$$

A narrow-sense heritability value obtained in this way by selective breeding and measuring the response in the offspring is referred to as an estimate of **realized heritability.**

As an example of a realized heritability estimate, suppose that we measure the diameter of corn kernels in a population where the mean diameter M is 20 mm. From this population, we select a group with the smallest diameters, for which the mean $M1$ equals 10 mm. The selected plants are interbred, and the mean diameter $M2$ of the progeny kernels is 13 mm. We can calculate the realized heritability h^2 to estimate the potential for artificial selection on kernel size:

$$h^2 = \frac{M2 - M}{M1 - M}$$

$$h^2 = \frac{13 - 20}{10 - 20}$$

$$= \frac{-7}{-10}$$

$$= 0.70$$

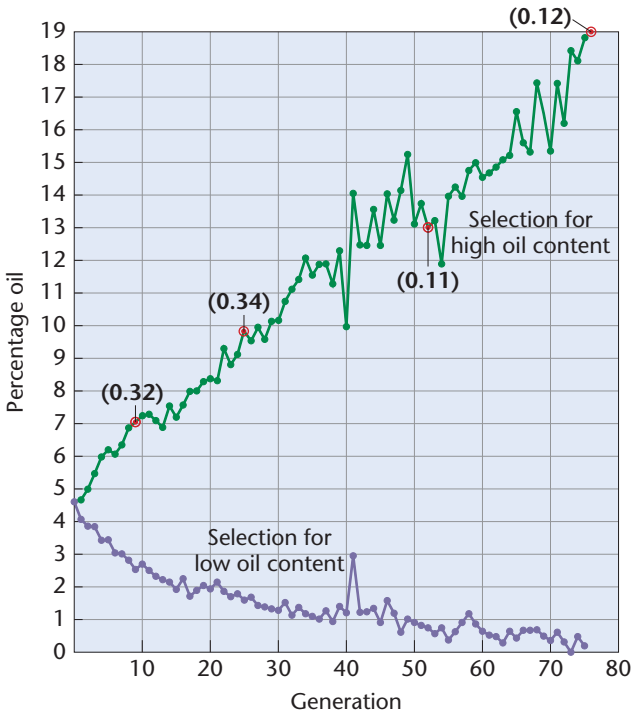


FIGURE 21–6 Response of corn selected for high and low oil content over 76 generations. The numbers in parentheses at generations 9, 25, 52, and 76 for the "high oil" line indicate the calculation of heritability at these points in the continuing experiment.

This value for narrow-sense heritability indicates that the selection potential for kernel size is relatively high.

The longest running artificial selection experiment known is still being conducted at the State Agricultural Laboratory in Illinois. Corn has been selected for both high and low oil content. After 76 generations, selection continues to result in increased oil content (**Figure 21–6**). With each cycle of successful selection, more of the corn plants accumulate a higher percentage of additive alleles involved in oil production. Consequently, the narrow-sense heritability h^2 of increased oil content in succeeding generations has declined (see parenthetical values at generations 9, 25, 52, and 76 in Figure 21–6) as artificial selection comes closer and closer to optimizing the genetic potential for oil production. Theoretically, the process will continue until all individuals in the population possess a uniform genotype that includes all the additive alleles responsible for high oil content. At that point, h^2 will be reduced to zero, and response to artificial selection will cease. The decrease in response to selection for low oil content shows that heritability for low oil content is approaching this point.

Table 21.4 lists narrow-sense heritability estimates expressed as percentage values for a variety of quantitative traits in different organisms. As you can see, these h^2 values

TABLE 21.4 Estimates of Heritability for Traits in Different Organisms

Trait	Heritability (h^2)
Mice	
Tail length	60%
Body weight	37
Litter size	15
Chickens	
Body weight	50
Egg production	20
Egg hatchability	15
Cattle	
Birth weight	45
Milk yield	44
Conception rate	3

vary, but heritability tends to be low for quantitative traits that are essential to an organism's survival. Remember, this does not indicate the absence of a genetic contribution to the observed phenotypes for such traits. Instead, the low h^2 values show that natural selection has already largely optimized the genetic component of these traits during evolution. Egg production, litter size, and conception rate are examples of how such physiological limitations on selection have already been reached. Traits that are less critical to survival, such as body weight, tail length, and wing length, have higher heritabilities because more genotypic variation for such traits is still present in the population. Remember also that any single heritability estimate can only provide information about one population in a specific environment. Studies involving the same trait in differing environments most often yield different results. Therefore, narrow-sense heritability is a more valuable predictor of response to selection when estimates are calculated for many populations and environments and show the presence of a clear trend.

Limitations of Heritability Studies

While the above discussion makes clear that heritability studies are valuable in estimating the genetic contribution to phenotypic variance, the knowledge gained about heritability of traits must be balanced by awareness of some of the constraints inherent in such estimates:

- Heritability values provide no information about what genes are involved in traits.
- Heritability is measured in populations, and has only limited application to individuals.
- Measured heritability depends on the environmental variation present in the population being studied, and cannot be used to evaluate differences between populations.
- Future changes in environmental factors can affect heritability.

21.4 Twin Studies Allow an Estimation of Heritability in Humans

Human twins are useful subjects for examining how much phenotypic variance for a multifactorial trait is due to the genotype as opposed to the environment. In these studies, the underlying principle has been that **monozygotic** (**MZ**), or **identical, twins** are derived from a single zygote that divides mitotically and then spontaneously splits into two separate cells. Both cells give rise to a genotypically identical embryo. **Dizygotic** (**DZ**), or **fraternal, twins,** on the other hand, originate from two separate fertilization events and are only as genetically similar as any two siblings, with an average of 50 percent of their alleles in common. For a given trait, therefore, phenotypic differences between pairs of MZ twins will be equivalent to the environmental variance (V_E) (because the genotypic variance is zero). Phenotypic differences between DZ twins, however, display both environmental variance (V_E) and approximately half the genotypic variance (V_G). Comparing the extent of phenotypic variance for the same trait in MZ and DZ sets of twins provides an estimate of broad-sense heritability for the trait.

Twins are said to be **concordant** for a given trait if both express it or neither expresses it. If one expresses the trait and the other does not, the pair is said to be **discordant.** Comparison of concordance values of MZ versus DZ twins reared together illustrates the potential value for heritability assessment. (See the Now Solve This feature on page 450, for example.)

Before any conclusions can be drawn from twin studies, the data must be examined carefully. For example, if concordance values approach 90 to 100 percent in MZ twins, we might be inclined to interpret that as a large genetic contribution to the phenotype of the trait. In some cases—for example, blood types and eye color—we know that this is indeed true. In the case of contracting measles, however, a high concordance value merely indicates that the trait is almost always induced by a factor in the environment—in this case, a virus.

It is more meaningful to compare the *difference* between the concordance values of MZ and DZ twins. If concordance values are significantly higher in MZ twins, we suspect a strong genetic component in the determination of the trait. In the case of measles, where concordance is high in both types of twins, the environment is assumed to be the major contributing factor. Such an analysis is useful because

phenotypic characteristics that remain similar in different environments are likely to have a strong genetic component.

Large Scale Analysis of Twin Studies

For decades, researchers have used twin studies to examine the relative contributions of genotype and environment to the phenotypic variation observed in complex traits in humans. These traits involve the interplay of multiple genes with a network of environmental factors, and the genetic components of the resulting phenotypic variance can be difficult to study. The simplest way to assess the genetic contribution is to assume that the effect of each gene on a trait is independent of the effects of other genes. Because the effects of all genes are added together, this is called the *additive model*. However, in recent years, some geneticists have proposed that non-additive factors such as dominance and epistasis are more important than additive genetic effects. As a result, the relative roles of additive and non-additive factors are a subject of active debate.

In an attempt to resolve this issue, an international project recently examined the results of all twin studies performed in the last 50 years. This study, published in 2015, involved the compilation and analysis of the data for over 17,000 traits studied in more than 14 million twin pairs drawn from more than 2,700 published papers.

Several important general conclusions can be drawn from this landmark study. First, based on correlations between MZ and DZ twin pairs, which can be used to draw conclusions about how likely it is that genetic influences on a trait are mostly additive or non-additive, researchers concluded that the vast majority of traits follow a simple additive model, providing strong support for one of the foundations of heritability studies. This does not exclude the role of non-additive factors such as dominance and epistasis, but these factors most likely play a secondary role in heritability. Second, the results are consistent with the findings from genome wide association studies (GWAS) that many complex traits are controlled by many genes, each with a small effects. Third, genetic variance is an important component of the individual variations observed in populations. In addition, the relative effects of genotypes and environmental factors are non-randomly distributed, making their contributions somewhat trait-specific.

The data from this study are available in a web-based application, Meta-analysis of Twin Correlations and Heritability (MaTCH), which can be used as a resource for the study of complex traits and the genetic and environmental components of heritability.

Twin Studies Have Several Limitations

Interesting as they are, human twin studies contain some unavoidable sources of error. For example, MZ twins are often treated more similarly by parents and teachers than are DZ twins, especially when the DZ siblings are of different sex. This circumstance may inflate the environmental variance for DZ twins. Another possible error source is interactions between the genotype and the environment that produce variability in the phenotype. These interactions can increase the total phenotypic variance for DZ twins compared to MZ twins raised in the same environment, influencing heritability calculations. Overall, heritability estimates for human traits based on twin studies should therefore be considered approximations and examined very carefully before any conclusions are drawn.

Although they must often be viewed with caution, classical twin studies, based on the assumption that MZ twins share the same genome, have been valuable for estimating heritability over a wide range of traits including multifactorial disorders such as cardiovascular disease, diabetes, and mental illness, for example. These disorders clearly have genetic components, and twin studies provide a foundation for studying interactions between genes and environmental factors. However, results from genomics research have challenged the view that MZ twins are truly identical and have forced a reevaluation of both the methodology and the results of twin studies. Such research has also opened the way to new approaches to the study of interactions between the genotype and environmental factors.

The most relevant genomic discoveries about twins include the following:

- By the time they are born, MZ twins do not necessarily have identical genomes.
- Gene-expression patterns in MZ twins change with age, leading to phenotypic differences.

We will address these points in order. First, MZ twins develop from a single fertilized egg, where sometime early in development the resulting cell mass separates into two distinct populations creating two independent embryos. Until that time, MZ twins have identical genotypes. Subsequently, however, the genotypes can diverge slightly. For example, differences in *copy number variation* (*CNV*)—variation in the number of copies of numerous large DNA sequences (usually 1000 bp or more)—may arise, differentially producing genetically distinct populations of cells in each embryo (see Chapter 6 for a discussion of CNV). This creates a condition called *somatic mosaicism,* which may result in a milder disease phenotype in some disorders and may play a similar role in phenotypic discordance observed in some pairs of MZ twins.

At this point, it is difficult to know for certain how often CNV arises after MZ twinning, but one estimate suggests that such differences occur in 10 percent of all twin

pairs. In those pairs where it does occur, one estimate is that such divergence takes place in 15 to 70 percent of the somatic cells. In one case, a CNV difference between MZ twins has been associated with chronic lymphocytic leukemia in one twin, but not the other.

The second genomic difference between MZ twins involves **epigenetics**—the chemical modification of their DNA and associated histones. An international study of epigenetic modifications in adult European MZ twins showed that MZ twin pairs are epigenetically identical at birth, but adult MZ twins show significant differences in the *methylation patterns* of both DNA and histones. Such epigenetic changes in turn affect patterns of gene expression. The accumulation of epigenetic changes and gene-expression profiles may explain some of the observed phenotypic discordance and susceptibility to diseases in adult MZ twins. For example, a clear difference in DNA methylation patterns is observed in MZ twins discordant for Beckwith–Wiedemann syndrome, a genetic disorder associated with variable developmental overgrowth of certain tissues and organs and an increased risk of developing cancerous and noncancerous tumors. Affected infants are often larger than normal, and one in five dies early in life.

Other complex disorders displaying a genetic component are similarly being investigated using epigenetic analysis in twin studies. These include susceptibility to several neurobiological disorders, including schizophrenia and autism, as well as to the development of Type 1 diabetes, breast cancer, and autoimmune disease.

Progressive, age-related genomic modifications may be the result of MZ twins being exposed to different environmental factors, or from failure of epigenetic marking following DNA replication. These findings also indicate that concordance studies in DZ twins must take into account genetic as well as *epigenetic differences* that contribute to discordance in these twin pairs.

The realization that epigenetics may play an important role in the development of phenotypes promises to make twin studies an especially valuable tool in dissecting the interactions among genes and the role of environmental factors in the production of phenotypes. Once the degree of epigenetic differences between MZ and DZ twin pairs has been defined, molecular studies on DNA and histone modification can link changes in gene expression with differences in the concordance rates between MZ and DZ twins.

We will discuss the most recent findings involving epigenetics and summarize its many forms and functions later in the text (see Special Topic Chapter 1—Epigenetics).

ESSENTIAL POINT

Twin studies, while having some limitations, are useful in assessing heritabilities for polygenic traits in humans. ■

NOW SOLVE THIS

21–3 The following table gives the percentage of twin pairs studied in which both twins expressed the same phenotype for a trait (concordance). Percentages listed are for concordance for each trait in monozygotic (MZ) and dizygotic (DZ) twins. Assuming that both twins in each pair were raised together in the same environment, what do you conclude about the relative importance of genetic versus environmental factors for each trait?

Trait	MZ %	DZ %
Blood types	100	66
Eye color	99	28
Mental retardation	97	37
Measles	95	87
Hair color	89	22
Handedness	79	77
Idiopathic epilepsy	72	15
Schizophrenia	69	10
Diabetes	65	18
Identical allergy	59	5
Cleft lip	42	5
Club foot	32	3
Mammary cancer	6	3

■ **HINT:** *This problem asks you to evaluate the relative importance of genetic versus environmental contributions to specific traits by examining concordance values in MZ versus DZ twins. The key to its solution is to examine the difference in concordance values and to factor in what you have learned about the genetic differences between MZ and DZ twins.*

21.5 Quantitative Trait Loci Are Useful in Studying Multifactorial Phenotypes

Environmental effects, interaction among segregating alleles, and the large number of genes that may contribute to the phenotype of polygenically controlled complex traits make it difficult to: (1) identify all genes that are involved; and (2) determine the effect of each gene on the phenotype. However, because many quantitative traits are of economic or medical relevance, it is often desirable to obtain this information. In such studies, a chromosome region is identified as containing one or more genes contributing to a quantitative trait known as a **quantitative trait locus (QTL)**.* When possible, the relevant gene or genes contained within a QTL are isolated and studied.

The modern approach used to find and map QTLs involves looking for associations between DNA markers and phenotypes. One way to do this is to begin with

*We utilize QTLs to designate the plural form, quantitative trait loci.

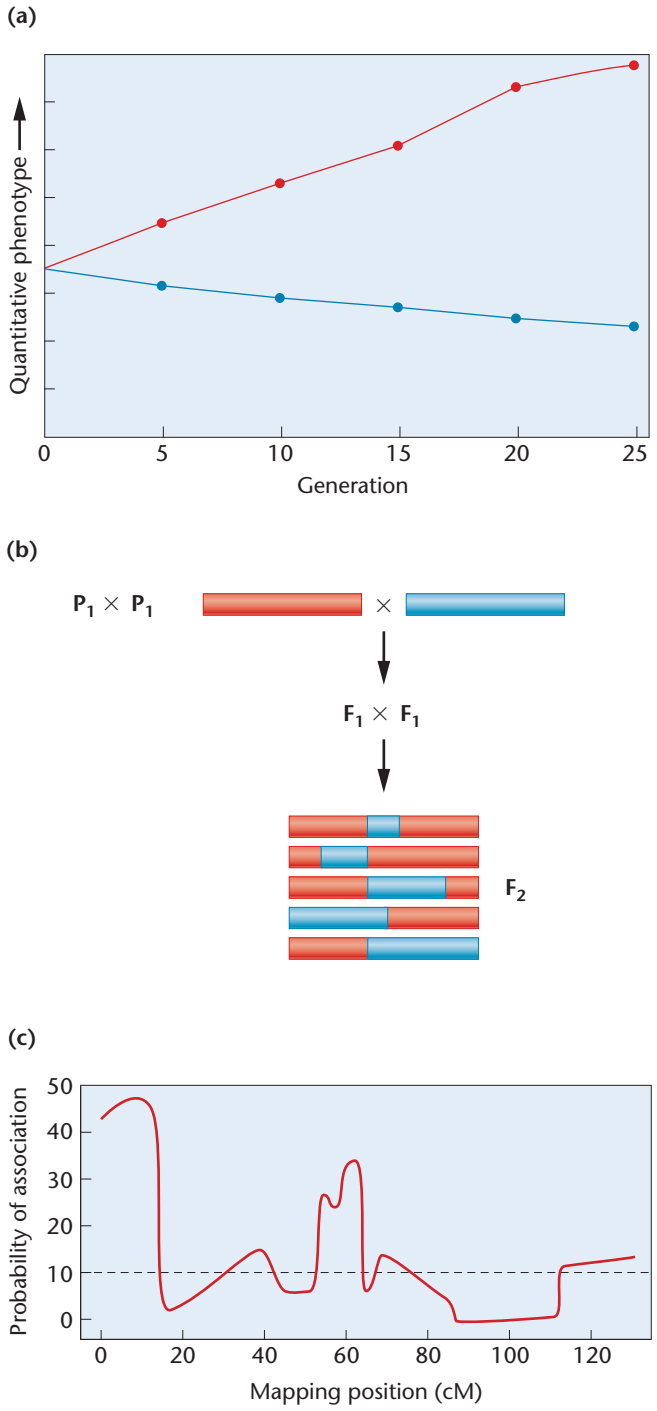


FIGURE 21–7 (a) Individuals from highly divergent lines created by artificial selection are chosen from generation 25 as parents. (b) The thick bars represent the genomes of individuals selected from the divergent lines as parents. These individuals are crossed to produce an F_1 generation (not shown). An F_2 generation is produced by crossing members of the F_1. As a result of crossing over, individual members of the F_2 generation carry different portions of the P_1 genome, as shown by the colored segments of the thick bars. DNA markers and phenotypes in individuals of the F_2 generation are analyzed. (c) Statistical methods are used to determine the probability that a DNA marker is associated with a QTL that affects the phenotype. The results are plotted as the likelihood of association against chromosomal location. Units on genetic maps are measured in centimorgans (cM), determined by crossover frequencies. Peaks above the horizontal line represent significant results. The data show five possible QTLs, with the most significant findings at about 10 cM and 60 cM.

individuals from two lines created by artificial selection that are highly divergent for a phenotype (fruit weight, oil content, bristle number, etc.). For example, **Figure 21–7** illustrates a generic case of QTL mapping. Over many generations of artificial selection, two divergent lines become highly homozygous, which facilitates their use in QTL mapping. Individuals from each of the lines with divergent phenotypes [generation 25 in **Figure 21–7(a)**] are used as parents to create an F_1 generation whose members will be heterozygous at most of the loci contributing to the trait. Additional crosses, either among F_1 individuals or between the F_1 and the inbred parent lines, result in F_2 generations that carry different portions of the parental genomes [**Figure 21–7(b)**] with different QTL genotypes and associated phenotypes. This segregating F_2 is known as the **QTL mapping population.**

Researchers then measure phenotypic expression of the trait among individuals in the mapping population and identify genomic differences among individuals by using chromosome-specific DNA markers such as *restriction fragment length polymorphisms* (*RFLPs*), *microsatellites,* and *single-nucleotide polymorphisms* (*SNPs*) (see Chapter 18). Computer-based statistical analysis is used to search for linkage between the markers and a component of phenotypic variation associated with the trait. If a DNA marker (such as those markers described above) *is not* linked to a QTL, then the phenotypic mean score for the trait will not vary among individuals with different genotypes at that marker locus. However, if a DNA marker *is* linked to a QTL, then different genotypes at that marker locus will also differ in their phenotypic expression of the trait. When this occurs, the marker locus and the QTL are said to *cosegregate.* Consistent cosegregation establishes the presence of a QTL at or near the DNA marker along the chromosome—in other words, the marker and QTL are linked. When numerous QTLs for a given trait have been located, a genetic map is created, showing the probability that specific chromosomal regions are associated with the phenotype of interest [**Figure 21–7(c)**]. Further research using genomic techniques identifies genes in these regions that contribute to the phenotype.

QTL mapping has been used extensively in agriculture, including plants such as corn, rice, wheat, and tomatoes (**Table 21.5**), and livestock such as cattle, pigs, sheep, and chickens.

Tomatoes are one of the world's major vegetable crops, and hundreds of varieties are grown and harvested each year. To aid in the creation of new varieties, hundreds of QTLs for traits including fruit size, shape, soluble solid content, and acidity have been identified and mapped to all 12 chromosomes in the tomato haploid genome. In addition,

TABLE 21.5 QTLs for Quantitative Phenotypes

Organism	Quantitative Phenotype	QTLs Identified
Tomato	Soluble solids	7
	Fruit mass	13
	Fruit pH	9
	Growth	5
	Leaflet shape	9
	Height	9
Maize	Height	11
	Leaf length	7
	Grain yield	18
	Number of ears	10

Source: Used with permission of Annual Reviews of Genetics, from "Mapping Polygenes" by S.D. Tanksley, *Annual Review of Genetics*, Vol. 27:205–233, Table 1, December 1993. © Annual Reviews, Inc.

the genomes of several tomato varieties have been sequenced. We will describe studies focused on quantitative traits controlling fruit shape and weight as an example of QTL research.

While the cultivated tomato can weigh up to 1000 grams, fruit from the related wild species thought to be the ancestor of the modern tomato weighs only a few grams [**Figure 21–8(a)**]. In a study by Steven Tanksley, QTL mapping has identified more than 28 QTLs related to this thousand-fold variation in fruit weight. More than ten years of work was required to localize, identify, and clone one of these QTLs, called *fw2.2* (on chromosome 2). Within this QTL, a specific gene, *ORFX,* has been identified, and alleles at this locus are responsible for about 30 percent of the variation in fruit weight.

The *ORFX* gene has been isolated, cloned, and transferred between plants, with interesting results. One allele of *ORFX* is present in all wild small-fruited varieties of tomatoes investigated, while another allele is present in all domesticated large-fruited varieties. When a cloned *ORFX* gene from small-fruited varieties is transferred to a plant that normally produces large tomatoes, the transformed plant produces fruits that are greatly reduced in weight. In the varieties studied by Tanksley's group, the reduction averaged 17 grams, a statistically significant phenotypic change caused by the action of a gene found within a single QTL.

FIGURE 21–8 A theoretical wild species of tomato, similar in size to the tomato on the left, is regarded as the ancestor of all modern tomatoes, including the beefsteak tomato shown at the right.

Further analysis of *ORFX* revealed that this gene encodes a protein that negatively regulates cell division during fruit development. Differences in the time of gene expression and differences in the amount of transcript produced lead to small or large fruit. Higher levels of expression mediated by transferred *ORFX* alleles exert a negative control over cell division, resulting in smaller tomatoes.

Yet *ORFX* and other related genes cannot account for all the observed variation in tomato size. Analysis of another QTL, *fas* (located on chromosome 11), indicates that the development of extreme differences in fruit size resulting from artificial selection also involves an increase in the number of compartments in the mature fruit. The small, ancestral stocks produce fruit with two to four seed compartments, but the large-fruited present-day strains have eight or more compartments. Thus, the QTLs that affect fruit size in tomatoes work by controlling at least two developmental processes: cell division early in development and the determination of the number of ovarian compartments.

The discovery that QTLs can control levels of gene expression has led to new, molecular definitions of phenotypes associated with quantitative traits. For example, the phenotype investigated may be the amount of an RNA transcript produced by a gene (**expression QTLs,** or **eQTLs**), or the amount of protein produced (**protein QTLs,** or **pQTLs**). These molecular phenotypes are polygenically controlled in the same way as more conventional phenotypes, such as fruit weight. Gene expression, for example, is controlled by *cis* factors, including promoters, and by *trans*-acting transcription factors (see Chapter 15 for a discussion of gene regulation in eukaryotes).

eQTLs and other genomic techniques such as genome editing may be necessary tools to develop new varieties. Studies indicate that intense artificial selection over the last 300 years has resulted in fixation of large portions of the genome and the accompanying loss of 95% of the tomato's genetic diversity.

Expression QTLs (eQTLs) and Genetic Disorders

We conclude this chapter by discussing the role that variation in the levels of gene expression plays in the phenotypic variation observed in complex disorders. The ability to study gene expression (eQTLs) and gene variability in the same individual helps identify gene/disease associations and the network of genes controlling those disorders. This approach has identified genes responsible for complex diseases such as asthma, cleft lip, Type 2 diabetes, and coronary artery disease. Asthma cases have risen dramatically over the last three decades and it is now a major public health concern. Genome-wide association studies (GWAS) have identified loci that confer susceptibility to asthma; however, the functions of many of these genes are unknown, and GWAS alone are unable to establish which alleles of these loci are responsible for susceptibility or the mechanism of their action.

To identify genes directly involved in asthma susceptibility, researchers collected lung specimens from over 1000

individuals and used lung-specific gene expression as a phenotype to study how genetic variants (DNA polymorphisms) are linked to both gene expression (eQTLs) and the asthma phenotype. Integration of the GWAS and the eQTL data identified a network of 34 genes that constitute the most likely gene set that causes asthma. In addition, six other genes were identified as drivers that control the other genes in the network. These driver genes are candidates for drug discovery studies to develop therapies for this chronic and sometimes fatal disease.

Similar approaches have identified candidate susceptibility genes and the outline of gene networks in schizophrenia and in Parkinson disease. The expanded use of eQTL analysis is expected to rapidly advance our knowledge of the genes responsible for other complex genetic disorders.

ESSENTIAL POINT

Quantitative trait loci, or QTLs, may be identified and mapped using DNA markers. ■

GENETICS, TECHNOLOGY, AND SOCIETY

The Green Revolution Revisited: Genetic Research with Rice

Of the more than 7 billion people now living on Earth, over 800 million do not have enough to eat. That number is expected to grow by an additional 1 million people each year for the next several decades. How will we be able to feed the estimated 8 billion people on Earth by 2025?

The past gives us some reasons to be optimistic. In the 1950s and 1960s, plant scientists set about to increase the production of crop plants, including the three most important grains—rice, wheat, and maize. These efforts became known as the *Green Revolution.* The approach was three-fold: (1) to increase the use of fertilizers, pesticides, and irrigation; (2) to bring more land under cultivation; and (3) to develop improved varieties of crop plants by intensive plant breeding.

The results were dramatic. Developing nations more than doubled their production of rice, wheat, and maize between 1961 and 1985.

The Green Revolution saved millions of people from starvation and improved the quality of life for millions more; however, its effects may be diminishing. The rate of increase in grain yields has slowed since the 1980s. If food production is to keep pace with the projected increase in the world's population, we will have to depend more on the genetic improvement of crop plants to provide higher yields.

About half of the Earth's population depends on rice for basic nourishment. The Green Revolution for rice began in 1960, aided by the establishment of the International Rice Research Institute (IRRI). One of their major developments was the breeding of a rice variety with improved disease resistance and higher yield. The IRRI research team crossed a Chinese rice variety (*Dee-geo-woo-gen*) and an Indonesian variety (*Peta*) to create a new cultivar known as IR8. IR8 produced a greater number of rice kernels per plant. However, IR8 plants were so top-heavy with grain that they tended to fall over—a trait called "lodging." To reduce lodging, IRRI breeders crossed IR8 with a dwarf native variety to create semi-dwarf lines. Due in part to the adoption of the semi-dwarf IR8 lines, the world production of rice doubled in 25 years.

Predictions suggest that a 40 percent increase in the annual rice harvest may be necessary to keep pace with anticipated population growth during the next 30 years. Greater emphasis will be placed on creating new rice varieties that have even higher yields and greater disease resistance.

In addition to conventional hybridization and selection techniques, dozens of quantitative trait loci (QTLs) from wild rice appear to contribute to increased yields, and scientists are attempting to introduce these traits into current dwarf varieties of domestic rice. Genomics and genetic engineering are also contributing to the new Green Revolution for rice. In 2002, the rice genome was the first cereal crop genome to be sequenced. As useful genes are identified, they may be transferred into rice plants with the goal of improving disease resistance, tolerance to drought and salinity, and nutritional content.

Your Turn

Take time, individually or in groups, to answer the following questions. Investigate the references and links to help you understand some of the technologies and issues surrounding the new Green Revolution.

1. Almost half of the world's rice is grown in soils that are poor in nutrients, subject to flooding or drought, or farmed by people unable to afford fertilizers. Can genetic research address these limitations?

One recent study indicates that a QTL in a wild rice variety may confer tolerance to several of these factors. Read about the promising research on the PSTOL1 *gene in* Gamuyao, R. et al. 2012. The protein kinase Pstol1 from traditional rice confers tolerance of phosphorus deficiency. *Nature* 488: 535–539.

2. Despite its benefits, the Green Revolution has been the subject of controversy. What are the main criticisms of the Green Revolution, and how can we mitigate some of the Green Revolution's negative aspects?

A discussion of this topic can be found in Pingali, P.L. 2012. Green Revolution: Impacts, limits, and the path ahead. *Proc. Natl. Acad. Sci. USA* 109 (31): 12302–12308.

CASE STUDY | Tissue-specific eQTLs

An eQTL study was carried out using colon and rectal biopsies, collected endoscopically from 65 controls and patients with inflammatory bowel disease (IBD) and colorectal cancer. RNA was extracted from each biopsy, and the gene expression was measured using microarray chipsets. Genomic DNA was genotyped with 730,525 SNPs to measure genetic variation throughout the genome. The linkage between gene expression and genetic variation was calculated in the patients and controls. The study identified 1312 independent eQTLs associated with the differential expression of 1222 genes in rectal tissues. 26% of these were novel and unique, compared to the previous GWAS and eQTL studies for IBD carried out on general tissues, either lymphoblastoid cell lines or blood. An examination of 163 IBD risk loci identified 11 SNPs that were rectal eQTLs. A colorectal cancer locus at 11q23 contained an eQTL associated with COLCA2, a protein implicated in colon cancer.

1. What is an eQTL, and what does it mean in the context of this study?
2. Why do you think it is important to measure gene expression in the tissues relevant to the disease, not just in general tissues like blood, in eQTL studies?
3. How can knowing which gene networks show differential levels of expression in IBD and colorectal cancer patients help us to treat these complex diseases, both now and in the future?

INSIGHTS AND SOLUTIONS

1. In a certain plant, height varies from 6 to 36 cm. When 6-cm and 36-cm plants were crossed, all F_1 plants were 21 cm. In the F_2 generation, a continuous range of heights was observed. Most were around 21 cm, and 3 of 200 were as short as the 6-cm P_1 parent.

   (a) What mode of inheritance does this illustrate, and how many gene pairs are involved?

   (b) How much does each additive allele contribute to height?

   (c) List all genotypes that give rise to plants that are 31 cm.

   Solution:
   (a) Polygenic inheritance is illustrated when a trait is continuous and when alleles contribute additively to the phenotype. The 3/200 ratio of F_2 plants is the key to determining the number of gene pairs. This reduces to a ratio of 1/66.7, very close to 1/64. Using the formula $1/4^n = 1/64$ (where 1/64 is equal to the proportion of F_2 phenotypes as extreme as either P_1 parent), $n = 3$. Therefore, three gene pairs are involved.

   (b) The variation between the two extreme phenotypes is

   $$36 - 6 = 30 \text{ cm}$$

   Because there are six potential additive alleles (*AABBCC*), each contributes

   $$30/6 = 5 \text{ cm}$$

   to the base height of 6 cm, which results when no additive alleles (*aabbcc*) are part of the genotype.

   (c) All genotypes that include five additive alleles will be 31 cm (5 alleles × 5 cm/allele + 6 cm base height = 31 cm). Therefore, *AABBCc*, *AABbCC*, and *AaBBCC* are the genotypes that will result in plants that are 31 cm.

2. A plant of unknown phenotype and genotype from the population described above (1.) was testcrossed, with the following results

   1/4 11 cm
   2/4 16 cm
   1/4 21 cm

   An astute genetics student realized that the unknown plant could be only one phenotype but could be any of three genotypes. What were they?

   Solution: When testcrossed (with *aabbcc*), the unknown plant must be able to contribute either one, two, or three additive alleles in its gametes in order to yield the three phenotypes in the offspring. Since no 6-cm offspring are observed, the unknown plant never contributes all nonadditive alleles (*abc*). Only plants that are homozygous at one locus and heterozygous at the other two loci will meet these criteria. Therefore, the unknown parent can be any of three genotypes, all of which have a phenotype of 26 cm:

   AABbCc
   AaBbCC
   AaBBCc

   For example, in the first genotype (*AABbCc*),

   AABbCc × *aabbcc*
   ↓
   1/4 *AaBbCc* 21 cm
   1/4 *AaBbcc* 16 cm
   1/4 *AabbCc* 16 cm
   1/4 *Aabbcc* 11 cm

   which is the ratio of phenotypes observed.

3. The mean and variance of corolla length in two highly inbred strains of *Nicotiana* and their progeny are shown in the following table. One parent (P_1) has a short corolla, and the other parent (P_2) has a long corolla. Calculate the broad-sense heritability (H^2) of corolla length in this plant.

Strain	Mean (mm)	Variance (mm)
P_1 short	40.47	3.12
P_2 long	93.75	3.87
F_1 ($P_1 \times P_2$)	63.90	4.74
F_2 ($F_1 \times F_1$)	68.72	47.70

   Solution: The formula for estimating heritability is $H^2 = V_G/V_P$, where V_G and V_P are the genetic and phenotypic components of variation, respectively. The main issue in this problem is obtaining some estimate of two components of

Insights and Solutions—continued

phenotypic variation: genetic and environmental factors. V_P is the combination of genetic and environmental variance. Because the two parental strains are true breeding, they are assumed to be homozygous, and the variance of 3.12 and 3.87 is considered to be the result of environmental influences. The average of these two values is 3.50. The F_1 is also genetically homogeneous and gives us an additional estimate of the impact of environmental factors. By averaging this value along with that of the parents,

$$\frac{4.74 + 3.50}{2} = 4.12$$

we obtain a relatively good idea of environmental impact on the phenotype. The phenotypic variance in the F_2 is the sum of the genetic (V_G) and environmental (V_E) components. We have estimated the environmental input as 4.12, so 47.70 minus 4.12 gives us an estimate of V_G of 43.58. Heritability then becomes 43.58/47.70, or 0.91. This value, when interpreted as a percentage, indicates that about 91 percent of the variation in corolla length is due to genetic influences.

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. In this chapter, we focused on a mode of inheritance referred to as quantitative genetics, as well as many of the statistical parameters utilized to study quantitative traits. Along the way, we found opportunities to consider the methods and reasoning by which geneticists acquired much of their understanding of quantitative genetics. From the explanations given in the chapter, what answers would you propose to the following fundamental questions:
   (a) How can we ascertain the number of polygenes involved in the inheritance of a quantitative trait?
   (b) What findings led geneticists to postulate the multiple-factor hypothesis that invoked the idea of additive alleles to explain inheritance patterns?
   (c) How do we assess environmental factors to determine if they impact the phenotype of a quantitatively inherited trait?
   (d) How do we know that monozygotic twins are not identical genotypically as adults?

CONCEPT QUESTION

2. Review the Chapter Concepts list on page 438. These all center on quantitative inheritance and the study and analysis of polygenic traits. Write a short essay that discusses the difference between the more traditional Mendelian and Neomendelian modes of inheritance (qualitative inheritance) and quantitative inheritance. ■
3. Define the following: (a) polygenic, (b) additive alleles, (c) monozygotic and dizygotic twins, (d) heritability, and (e) QTL.
4. A dark-red strain and a white strain of wheat are crossed and produce an intermediate, medium-red F_1. When the F_1 plants are interbred, an F_2 generation is produced in a ratio of 1 dark-red: 4 medium-dark-red: 6 medium-red: 4 light-red: 1 white. Further crosses reveal that the dark-red and white F_2 plants are true breeding.
   (a) Based on the ratios in the F_2 population, how many genes are involved in the production of color?
   (b) How many additive alleles are needed to produce each possible phenotype?
   (c) Assign symbols to these alleles and list possible genotypes that give rise to the medium-red and light-red phenotypes.
   (d) Predict the outcome of the F_1 and F_2 generations in a cross between a true-breeding medium-red plant and a white plant.
5. Height in humans depends on the additive action of genes. Assume that this trait is controlled by the four loci *R, S, T,* and *U* and that environmental effects are negligible. Instead of additive versus nonadditive alleles, assume that additive and partially additive alleles exist. Additive alleles contribute two units, and partially additive alleles contribute one unit to height.
   (a) Can two individuals of moderate height produce offspring that are much taller or shorter than either parent? If so, how?
   (b) If an individual with the minimum height specified by these genes marries an individual of intermediate or moderate height, will any of their children be taller than the tall parent? Why or why not?
6. An inbred strain of plants has a mean height of 24 cm. A second strain of the same species from a different geographical region also has a mean height of 24 cm. When plants from the two strains are crossed together, the F_1 plants are the same height as the parent plants. However, the F_2 generation shows a wide range of heights; the majority are like the P_1 and F_1 plants, but approximately 4 of 1000 are only 12 cm high, and about 4 of 1000 are 36 cm high.
   (a) What mode of inheritance is occurring here?
   (b) How many gene pairs are involved?
   (c) How much does each gene contribute to plant height?
   (d) Indicate one possible set of genotypes for the original P_1 parents and the F_1 plants that could account for these results.
   (e) Indicate three possible genotypes that could account for F_2 plants that are 18 cm high and three that account for F_2 plants that are 33 cm high.
7. Erma and Harvey were a compatible barnyard pair, but a curious sight. Harvey's tail was only 6 cm long, while Erma's was 30 cm. Their F_1 piglet offspring all grew tails that were 18 cm. When inbred, an F_2 generation resulted in many piglets (Erma and Harvey's grandpigs), whose tails ranged in 4-cm intervals from 6 to 30 cm (6, 10, 14, 18, 22, 26, and 30). Most had 18-cm tails, while 1/64 had 6-cm tails and 1/64 had 30-cm tails.
   (a) Explain how these tail lengths were inherited by describing the mode of inheritance, indicating how many gene pairs were at work, and designating the genotypes of Harvey, Erma, and their 18-cm-tail offspring.
   (b) If one of the 18-cm F_1 pigs is mated with one of the 6-cm F_2 pigs, what phenotypic ratio would be predicted if many offspring resulted? Diagram the cross.

8. In the following table, average differences of height, weight, and fingerprint ridge count between monozygotic twins (reared together and apart), dizygotic twins, and nontwin siblings are compared:

Trait	MZ Reared Together	MZ Reared Apart	DZ Reared Together	Sibs Reared Together
Height (cm)	1.7	1.8	4.4	4.5
Weight (kg)	1.9	4.5	4.5	4.7
Ridge count	0.7	0.6	2.4	2.7

Based on the data in this table, which of these quantitative traits has the highest heritability values?

9. Define the term *broad-sense heritability* (*H2*). What is implied by a relatively high value of *H2*? Express aspects of broad-sense heritability in equation form.
10. Describe the value of using twins in the study of questions relating to the relative impact of heredity versus environment.
11. Corn plants from a test plot are measured, and the distribution of heights at 10-cm intervals is recorded in the following table:

Height (cm)	Plants (no.)
100	20
110	60
120	90
130	130
140	180
150	120
160	70
170	50
180	40

Calculate (a) the mean height, (b) the variance, (c) the standard deviation, and (d) the standard error of the mean. Plot a rough graph of plant height against frequency. Do the values represent a normal distribution? Based on your calculations, how would you assess the variation within this population?

12. The following variances were calculated for two traits in a herd of hogs.

Trait	V_P	V_G	V_A
Back fat	30.6	12.2	8.44
Body length	52.4	26.4	11.70

(a) Calculate broad-sense (H^2) and narrow-sense (h^2) heritabilities for each trait in this herd.
(b) Which of the two traits will respond best to selection by a breeder? Why?

13. The mean and variance of plant height of two highly inbred strains (P_1 and P_2) and their progeny (F_1 and F_2) are shown here.

Strain	Mean (cm)	Variance
P_1	34.2	4.2
P_2	55.3	3.8
F_1	44.2	5.6
F_2	46.3	10.3

Calculate the broad-sense heritability (H^2) of plant height in this species.

14. A hypothetical study investigated the vitamin A content and the cholesterol content of eggs from a large population of chickens. The variances (V) were calculated, as shown below:

Variance	Vitamin A	Cholesterol
V_P	123.5	862.0
V_E	96.2	484.6
V_A	12.0	192.1
V_D	15.3	185.3

(a) Calculate the narrow-sense heritability (h^2) for both traits.
(b) Which trait, if either, is likely to respond to selection?

15. If one is attempting to determine the influence of genes or the environment on phenotypic variation, inbred strains with individuals of a relatively homogeneous or constant genetic background are often used. Variation observed between different inbred strains reared in a constant or homogeneous environment would likely be caused by genetic factors. What would be the source of variation observed among members of the same inbred strain reared under varying environmental conditions?
16. A population of laboratory mice was weighed at the age of six weeks (full adult weight) and found to have a mean weight of 20 g. The narrow heritability of weight gain (h^2) is known to be 0.25 in this laboratory strain. If mice weighing 24 g are selected and mated at random, what is the expected mean weight of the next generation?
17. If the experiment was repeated by mating mice 4 g lighter than the mean (16 g), what would be the result? If repeated experiments were carried out always selecting for mice that were 4 g lighter than the mean in the current generation, what would eventually happen?
18. In a herd of Texas Longhorn cattle, the mean horn length from tip to tip is 52" and h^2 is 0.2. Predict the mean horn length if cattle with horns 61" long are interbred.
19. In a population of 100 inbred, genotypically identical rice plants, variance for grain yield is 4.67. What is the heritability for yield? Would you advise a rice breeder to improve yield in this strain of rice plants by selection?
20. A 3-inch plant was crossed with a 15-inch plant, and all F_1 plants were 9 inches. The F_2 plants exhibited a "normal distribution," with heights of 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, and 15 inches.
(a) What ratio will constitute the "normal distribution" in the F_2?
(b) What will be the outcome if the F_1 plants are testcrossed with plants that are homozygous for all nonadditive alleles?
21. Two different crosses were set up between carrots (*Daucus carota*) of different colors and carotenoid content (Santos, Carlos A. F. and Simon, Philipp W. 2002. *Horticultura Brasileira* 20). Analyses of the F_2 generations showed that four loci are associated with the α carotene content of carrots, with a broad-sense heritability of 90%. How many distinct phenotypic categories and genotypes would be seen in each F_2 generation, and what does a broad-sense heritability of 90% mean for carrot horticulture?
22. While most quantitative traits display continuous variation, there are others referred to as "threshold traits" that are distinguished by having a small number of discrete phenotypic classes. For example, Type 2 diabetes (adult-onset diabetes) is considered to be a polygenic trait, but demonstrates only two phenotypic classes: individuals who develop the disease and those who do not. Theorize how a threshold trait such as Type 2 diabetes may be under the control of many polygenes, but express a limited number of phenotypes.

22 Population and Evolutionary Genetics

CHAPTER CONCEPTS

- Most populations and species harbor considerable genetic variation.
- This variation is reflected in the alleles distributed among populations of a species.
- The relationship between allele frequencies and genotype frequencies in an ideal population is described by the Hardy–Weinberg law.
- Selection, migration, and genetic drift can cause changes in allele frequency.
- Mutation creates new alleles in a population gene pool.
- Nonrandom mating changes population genotype frequency but not allele frequency.
- A reduction in gene flow between populations, accompanied by selection or genetic drift, can lead to reproductive isolation and speciation.
- Genetic differences between populations or species are used to reconstruct evolutionary history.

These ladybird beetles, from the Chiricahua Mountains in Arizona, show considerable phenotypic variation.

In the mid-nineteenth century, Alfred Russel Wallace and Charles Darwin identified natural selection as the mechanism of evolution. In his book, *On the Origin of Species,* published in 1859, Darwin provided evidence that populations and species are not fixed, but change, or evolve, over time as a result of natural selection. However, Wallace and Darwin could not explain either the origin of the variations that provide the raw material for evolution or the mechanisms by which such variations are passed from parents to offspring. Gregor Mendel published his work on the inheritance of traits in 1866, but it received little notice at the time. The rediscovery of Mendel's work in 1900 began a 30-year effort to reconcile Mendel's concept of genes and alleles with the theory of evolution by natural selection. As twentieth-century biologists applied the principles of Mendelian genetics to populations, both the source of variation (mutation and recombination) and the mechanism of inheritance (segregation of alleles) were explained. We now view evolution as a consequence of changes in genetic material through mutation and changes in allele frequencies in populations over time. This union of population genetics with the theory of natural selection generated a new view of the evolutionary process, called *neo-Darwinism.*

In addition to natural selection, other forces including mutation, migration, and drift, individually and collectively, alter allele frequencies and bring about evolutionary divergence that eventually may result in **speciation,** the formation of new species. Speciation is facilitated by environmental diversity. If a population is spread over a geographic range encompassing a

number of ecologically distinct subenvironments with different selection pressures, the populations occupying these areas may gradually adapt and become genetically distinct from one another. Genetically differentiated populations may remain in existence, become extinct, reunite with each other, or continue to diverge until they become reproductively isolated. Populations that are reproductively isolated are regarded as separate species. Genetic changes within populations can modify a species over time, transform it into another species, or cause it to split into two or more species.

Population geneticists investigate patterns of genetic variation within and among groups of interbreeding individuals. As changes in genetic structure form the basis for evolution of a population, population genetics has become an important subdiscipline of evolutionary biology. In this chapter, we examine the population genetics processes of **microevolution**—defined as evolutionary change within populations of a species—and then consider how molecular aspects of these processes can be extended to **macroevolution**—defined as evolutionary events leading to the emergence of new species and other taxonomic groups.

22.1 Genetic Variation Is Present in Most Populations and Species

A **population** is a group of individuals belonging to the same species that live in a defined geographic area and actually or potentially interbreed. In thinking about the human population, we can define it as everyone who lives in the United States, or in Sri Lanka, or we can specify a population as all the residents of a particular small town or village.

The genetic information carried by members of a population constitutes that population's **gene pool.** At first glance, it might seem that a population that is well-adapted to its environment must have a gene pool that is highly homozygous because it would seem likely that the most favorable allele at each locus is present at a high frequency. In addition, a look at most populations of plants and animals reveals many phenotypic similarities among individuals. However, a large body of evidence indicates that, in reality, most populations contain a high degree of heterozygosity. This built-in genetic variation is not necessarily apparent in the phenotype; hence, detecting it is not a simple task. Nevertheless, the amount of variation within a population can be revealed by several methods.

Detecting Genetic Variation

The detection and use of genetic variation in individuals and populations began long before genetics emerged as a science. Millennia ago, plant and animal breeders began using artificial selection to domesticate plants and animals. However, as genetic technology developed in the last century, the ability to detect and quantify genetic variation in genes, in individual genomes, and in the genomes of populations has grown exponentially.

FIGURE 22–1 The size difference between a chihuahua and a Great Dane illustrates the high degree of genetic variation present in the dog genome.

One of the more spectacular examples of how much variation exists in the gene pool of a species was the use of selective breeding to create hundreds of dog breeds in nineteenth-century England over a period of less than 75 years. Many people, seeing a chihuahua (about 10 inches high) and a Great Dane (about 42 inches high) for the first time, might find it difficult to believe they are both members of the same species (**Figure 22–1**).

Recombinant DNA Technology and Genetic Variation

After the discovery that DNA carries genetic information and the development of recombinant DNA technology, efforts centered on detecting genetic variation in the sequence of individual genes carried by individuals in a population.

In one such study, Martin Kreitman isolated, cloned, and sequenced copies of the alcohol dehydrogenase (*Adh*) gene from individuals representing five different populations of *Drosophila melanogaster*. The 11 cloned genes from these five populations contained a total of 43 nucleotide differences in the *Adh* sequence of 2721 base pairs (**Figure 22–2**). These variations are distributed throughout the gene: 14 in exon coding regions, 18 within introns, and 11 in untranslated flanking regions. Of the 14 variations in exons, only one leads to an amino acid replacement—the one in codon 192, resulting in the two known alleles of this gene. The other 13 nucleotide changes do not lead to amino acid replacements and are silent variations of this gene.

	Exon 3	Intron 3	Exon 4
Consensus *Adh* sequence:	C C C C	G G A A T	C T C C A*C T A G
Strain			
Wa-S	T T • A	C A • T A	A C • • • • • • •
Fl1-S	T T • A	C A • T A	A C • • • • • • •
Ja-S	• • • •	• • • • •	• • • T • T • C A
Fl-F	• • • •	• • • • •	• • G T C T C C •
Ja-F	• • A •	• • G • •	• • G T C T C C •

FIGURE 22–2 DNA sequence variation in parts of the *Drosophila Adh* gene in a sample of the 11 laboratory strains derived from the five natural populations. The dots represent nucleotides that are the same as the consensus sequence; letters represent nucleotide polymorphisms. An A/C polymorphism (A*) in codon 192 creates the two *Adh* alleles (F and S). All other polymorphisms are silent or noncoding.

Genetic Variation in Genomes

The development of **next-generation sequencing technology** has extended the detection of genomic variation from individuals to populations. The 1000 Genomes Project, started in 2008, is a global effort with the goal of identifying and cataloging 95 percent of the common genetic variations carried by the 7 billion people now inhabiting the planet. The Project's three pilot studies combined low-coverage whole-genome sequencing, exome sequencing of selected protein-coding regions, and deep sequencing of selected parents and one of their children. Studying 1000 genomes is just a start; eventually, the Project intends to sequence the genomes of 2500 individuals from 27 population groups worldwide.

To date, the pilot studies have identified 15 million single-nucleotide polymorphisms (SNPs), 1 million short insertions and/or deletions (indels), and 20,000 large structural variants in the human genome. In addition, many individuals were heterozygous for 250 to 300 genes with loss-of-function mutations, 20 to 40 percent of which are known to be associated with genetic disorders. The Project's overall goal is to explore and understand the relationship between genotype and phenotype. In humans, this translates into identifying variants associated with disease. For example, in studies to date, no single variant has been associated with diabetes; this implies that a combination of heritable multiple rare variants is related to this common disorder. Eventually, researchers hope to associate specific genetic variants with cellular pathways and networks associated with complex disorders such as hypertension, cardiovascular disease, and neurological disorders associated with protein accumulation such as Alzheimer disease and Huntington disease.

ESSENTIAL POINT

Genetic variation is widespread in most populations and provides a reservoir of alleles that serve as the basis for evolutionary changes within the population. ■

22.2 The Hardy–Weinberg Law Describes Allele Frequencies and Genotype Frequencies in Population Gene Pools

Often when we examine a single gene in a population, we find that combinations of the alleles of this gene result in individuals with different genotypes. For example, two alleles (*A* and *a*) of the *A* gene can be combined to produce three genotypes: *AA*, *Aa*, and *aa*. Key elements of population genetics depend on the calculation of allele frequencies and genotype frequencies in a gene pool, and the determination of how these frequencies change from one generation to the next. Population geneticists use these calculations to answer questions such as: How much genetic variation is present in a population? Are genotypes randomly distributed in time and space, or do discernible patterns exist? What processes affect the composition of a population's gene pool? Do these processes produce genetic divergence among populations that may lead to the formation of new species? Changes in allele frequencies in a population that do not directly result in species formation are examples of microevolution. In the following sections, we will discuss microevolutionary changes in population gene pools, and in later sections, we will consider macroevolution and the process of speciation.

The relationship between the relative proportions of alleles in the gene pool and the frequencies of different genotypes in the population was elegantly described in a mathematical model. This model, called the **Hardy—Weinberg law,** describes what happens to allele and genotype frequencies in an "ideal" population that is infinitely large and randomly mating, and that is not subject to any evolutionary forces such as mutation, migration, or selection.

Calculating Genotype Frequencies

The Hardy—Weinberg model uses the principle of Mendelian segregation and simple probability to explain the relationship between allele and genotype frequencies in a population. We can demonstrate how this works by considering a single autosomal gene with two alleles, *A* and *a,* in a population where the frequency of *A* is 0.7 and the frequency of *a* is 0.3. Note that $0.7 + 0.3 = 1$, indicating that all the alleles of gene *A* present in the population are accounted for.

These allele frequencies mean that the probability that any female gamete will contain *A* is 0.7, and the probability

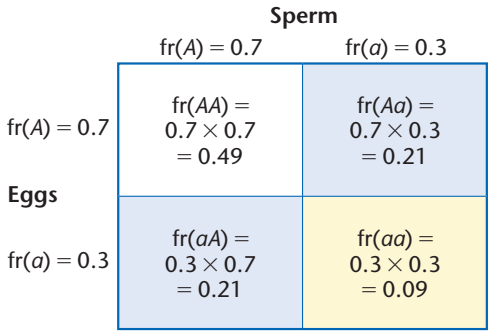


FIGURE 22–3 Calculating genotype frequencies from allele frequencies. Gametes represent samples drawn from the gene pool to form the genotypes of the next generation. In this population, the frequency of the *A* allele is 0.7, and the frequency of the *a* allele is 0.3. The frequencies of the genotypes in the next generation are calculated as 0.49 for *AA*, 0.42 for *Aa*, and 0.09 for *aa*. Under the Hardy–Weinberg law, the frequencies of *A* and *a* remain constant from generation to generation.

that a male gamete will contain *A* is also 0.7. The probability that *both* gametes will contain *A* is $0.7 \times 0.7 = 0.49$. Thus we predict that in the offspring, the genotype *AA* will occur 49 percent of the time. The probability that a zygote will be formed from a female gamete carrying *A* and a male gamete carrying *a* is $0.7 \times 0.3 = 0.21$, and the probability of a female gamete carrying *a* being fertilized by a male gamete carrying *A* is $0.3 \times 0.7 = 0.21$, so the frequency of genotype *Aa* in the offspring is $0.21 + 0.21 = 0.42 = 42$ percent. Finally, the probability that a zygote will be formed from two gametes carrying *a* is $0.3 \times 0.3 = 0.09$, so the frequency of genotype *aa* is 9 percent. As a check on our calculations, note that $0.49 + 0.42 + 0.09 = 1.0$, confirming that we have accounted for all possible genotypic combinations in the zygotes. These calculations are summarized in **Figure 22–3**.

Calculating Allele Frequencies

Now that we know the frequencies of genotypes in the next generation, what will be the allele frequencies in this new generation? Under the Hardy–Weinberg law, we assume that all genotypes have equal rates of survival and reproduction. This means that in the next generation, all genotypes contribute equally to the new gene pool. The *AA* individuals constitute 49 percent of the population, and we can predict that the gametes they produce will constitute 49 percent of the gene pool. These gametes all carry allele *A*. Similarly, *Aa* individuals constitute 42 percent of the population, so we predict that their gametes will constitute 42 percent of the new gene pool. Half (0.5) of these gametes will carry allele *A*. Thus, the frequency of allele *A* in the gene pool is $0.49 + (0.5)\,0.42 = 0.7$. The other half of the gametes produced by *Aa* individuals will carry allele *a*. The *aa* individuals constitute 9 percent of the population, so their gametes will constitute 9 percent of the new gene pool. All these gametes carry allele *a*. Thus, we can predict that the allele *a* in the new gene pool is $(0.5)\ 0.42 + 0.09 = 0.3$. As a check on our calculation, note that $0.7 + 0.3 = 1.0$, accounting for all of the gametes in the gene pool of the new generation.

The Hardy–Weinberg Law and Its Assumptions

Because the Hardy–Weinberg law is a mathematical model, we use variables instead of numerical values for the allele frequencies in the general case. Imagine a gene pool in which the frequency of allele *A* is represented by *p* and the frequency of allele *a* is represented by *q*, such that $p + q = 1$. If we randomly draw male and female gametes from the gene pool and pair them to make a zygote, the probability that both will carry allele *A* is $p \times p$. Thus, the frequency of genotype *AA* among the zygotes is p^2. The probability that the female gamete carries *A* and the male gamete carries *a* is $p \times q$, and the probability that the female gamete carries *a* and the male gamete carries *A* is $q \times p$. Thus, the frequency of genotype *Aa* among the zygotes is $2pq$. Finally, the probability that both gametes carry *a* is $q \times q$, making the frequency of genotype *aa* among the zygotes q^2. Therefore, the distribution of genotypes among the zygotes is

$$p^2 + 2pq + q^2 = 1$$

These calculations are summarized in **Figure 22–4**. They demonstrate the two main predictions of the Hardy–Weinberg model:

1. Allele frequencies in our population do not change from one generation to the next.
2. After one generation of random mating, genotype frequencies can be predicted from the allele frequencies.

In other words, there is no change in allele frequency, and for this locus, the population does not undergo any

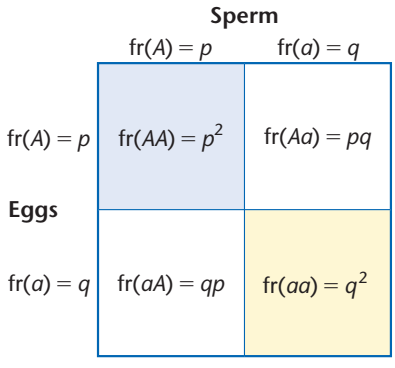


FIGURE 22–4 The general description of allele and genotype frequencies under Hardy–Weinberg assumptions. The frequency of allele *A* is *p*, and the frequency of allele *a* is *q*. After mating, the three genotypes *AA*, *Aa*, and *aa* have the frequencies p^2, $2pq$, and q^2, respectively.

microevolutionary change. The theoretical population described by the Hardy–Weinberg model is based on the following assumptions:

1. Individuals of all genotypes have equal rates of survival and equal reproductive success—that is, there is no selection.
2. No new alleles are created or converted from one allele into another by mutation.
3. Individuals do not migrate into or out of the population.
4. The population is infinitely large, which in practical terms means that the population is large enough that sampling errors and other random effects are negligible.
4. Individuals in the population mate randomly.

These assumptions are what make the Hardy–Weinberg model so useful in population genetics research. By specifying the conditions under which the population does not evolve, the Hardy–Weinberg model can be used to identify the real-world forces that cause allele frequencies to change. Application of this model can also reveal "neutral genes" in a population gene pool—those not being operated on by the forces of evolution.

The Hardy–Weinberg model has three additional important consequences:

1. Dominant traits do not necessarily increase from one generation to the next.
2. Genetic variability can be maintained in a population, since, once established in an ideal population, allele frequencies remain unchanged.
3. Under Hardy–Weinberg assumptions, knowing the frequency of just one genotype enables us to calculate the frequencies of all other genotypes at that locus.

This is particularly useful in human genetics because we can calculate the frequency of heterozygous carriers for recessive genetic disorders even when all we know is the frequency of affected individuals.

NOW SOLVE THIS

22–1 The ability to taste the compound PTC is controlled by a dominant allele *T*, while individuals homozygous for the recessive allele *t* are unable to taste PTC. In a genetics class of 125 students, 88 can taste PTC and 37 cannot. Calculate the frequency of the *T* and *t* alleles in this population and the frequency of the genotypes.

■ **HINT:** *This problem involves an understanding of how to use the Hardy–Weinberg law. The key to its solution is to determine which allele frequency (p or q) you must estimate first when homozygous dominant and heterozygous genotypes have the same phenotype.*

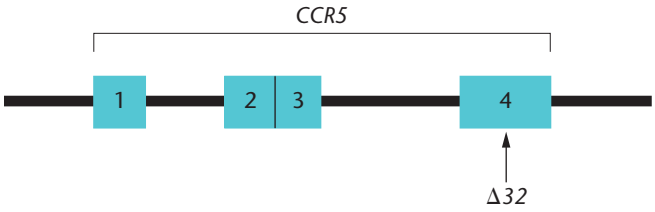


FIGURE 22–5 Organization of the *CCR5* gene in region 3p21.3 of human chromosome 3. The gene contains 4 exons and 2 introns (there is no intron between exons 2 and 3). The arrow shows the location of the 32-bp deletion in exon 4 that confers resistance to HIV-1 infection.

22.3 The Hardy–Weinberg Law Can Be Applied to Human Populations

To show how allele frequencies are measured in a real population, let's consider a gene that influences an individual's susceptibility to infection by HIV-1, the virus responsible for AIDS (acquired immunodeficiency syndrome). A small number of individuals who make high-risk choices (such as unprotected sex with HIV-positive partners) never become infected. Some of these individuals are homozygous for a mutant allele of a gene called *CCR5*.

The *CCR5* gene (**Figure 22–5**) encodes a protein called the C-C chemokine receptor-5, often abbreviated CCR5. Chemokines are signaling molecules associated with the immune system. The CCR5 protein is also used by strains of HIV-1 to gain entry into cells. The mutant allele of the *CCR5* gene contains a 32-bp deletion, making the encoded protein shorter and nonfunctional, blocking the entry of HIV-1 into cells. The normal allele is called *CCR51* (also called *1*), and the mutant allele is called *CCR5-Δ32* (also called *Δ32*).

Individuals homozygous for the mutant allele (*Δ32/Δ32*) are resistant to HIV-1 infection. Heterozygous (*1/Δ32*) individuals are susceptible to HIV-1 infection but progress more slowly to AIDS. **Table 22.1** summarizes the genotypes possible at the *CCR5* locus and the phenotypes associated with each.

The discovery of the *CCR5-Δ32* allele generates two important questions: Which human populations carry this

TABLE 22.1 ***CCR5* Genotypes and Phenotypes**

Genotype	Phenotype
1/1	Susceptible to sexually transmitted strains of HIV-1
1/Δ32	Susceptible but may progress to AIDS slowly
Δ32/Δ32	Resistant to most sexually transmitted strains of HIV-1

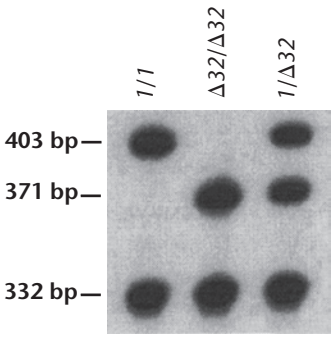


FIGURE 22–6 Allelic variation in the *CCR5* gene. Michel Samson and colleagues used PCR to amplify a part of the *CCR5* gene containing the site of the 32-bp deletion, cut the resulting DNA fragments with a restriction enzyme, and ran the fragments on an electrophoresis gel. Each lane reveals the genotype of a single individual. The *1* allele produces a 332-bp fragment and a 403-bp fragment; the *Δ32* allele produces a 332-bp fragment and a 371-bp fragment. Heterozygotes produce three bands.

allele, and how common is it? To address these questions, teams of researchers surveyed members of several populations. Genotypes were determined by direct analysis of DNA (Figure 22–6). In one population, 79 individuals had genotype *1/1*, 20 were *1/Δ32*, and 1 individual was *Δ32/Δ32*. We can see that this population has 158 *1* alleles carried by the *1/1* individuals plus 20 *1* alleles carried by *1/Δ32* individuals, for a total of 178. The frequency of the *CCR51* allele in the sample population is thus 178/200 = 0.89 = 89 percent. Copies of the *CCR5-Δ32* allele were carried by 20 *1/Δ32* individuals, plus 2 carried by the *Δ32/Δ32* individual, for a total of 22. The frequency of the *CCR5-Δ32* allele is thus 22/200 = 0.11 = 11 percent. Notice that $p + q = 1$, confirming that we have accounted for all the alleles of the *CCR51* gene in the population. Table 22.2 shows two methods for computing the frequencies of the alleles in the population surveyed.

Can we expect the *CCR5-Δ32* allele to increase in human populations because it offers resistance to infection by HIV? This specific question is difficult to answer directly, but as we will see later in this chapter, when factors such as natural selection, mutation, migration, or genetic drift are present, the allele frequencies in a population may change from one generation to the next.

By determining allele frequencies over more than one generation, it is possible to determine whether the frequencies remain in equilibrium because the Hardy–Weinberg assumptions are operating. Populations that meet the Hardy–Weinberg assumptions are not evolving because allele frequencies (for the generations tested) are not changing. However, a population may be in Hardy–Weinberg equilibrium for the alleles being tested, but other genes may not be in equilibrium.

Testing for Hardy–Weinberg Equilibrium in a Population

One way to see whether any of the Hardy–Weinberg assumptions do not hold in a given population is to determine whether the population's genotypes are in equilibrium. To do this, we first determine the genotype frequencies. This can be done directly from the phenotypes (if heterozygotes are recognizable), by analyzing proteins or DNA sequences, or indirectly using the frequency of the HIV-1 resistant phenotype in the population to calculate genotype frequencies using the Hardy–Weinberg law. We can then calculate the allele frequencies from the genotype frequencies. Finally, the allele frequencies in the parental generation are used to predict the genotype frequencies in the next generation. According to the Hardy–Weinberg law, genotype frequencies are predicted to fit the $p^2 + 2pq + q^2 = 1$ relationship. If they do not, then one or more of the assumptions are invalid for the population in question.

TABLE 22.2 Methods of Determining Allele Frequencies from Data on Genotypes

	Genotype			
(a) Counting Alleles	***1/1***	***1/Δ32***	***Δ32/Δ32***	**Total**
Number of individuals	79	20	1	100
Number of *1* alleles	158	20	0	178
Number of *Δ32* alleles	0	20	2	22
Total number of alleles	158	40	2	200
Frequency of *CCR51* in sample: 178/200 = 0.89 = 89%				
Frequency of *CCR5-Δ32* in sample: 22/200 = 0.11 = 11%				

	Genotype			
(b) From Genotype Frequencies	***1/1***	***1/Δ32***	***Δ32/Δ32***	**Total**
Number of individuals	79	20	1	100
Genotype frequency	79/100 = 0.79	20/100 = 0.20	1/100 = 0.01	1.00
Frequency of *CCR51* in sample: 0.79 + (0.5)0.20 = 0.89 = 89%				
Frequency of *CCR5-Δ32* in sample: (0.5)0.20 + 0.01 = 0.11 = 11%				

To demonstrate, let's examine *CCR5* genotypes in a hypothetical population. Our population is composed of 283 individuals; of these, 223 have genotype *1/1*; 57 have genotype *1/Δ32*; and 3 have genotype *Δ32/Δ32*. These numbers represent the following genotype frequencies: *1/1* = 223/283 = 0.788, *1/Δ32* = 57/283 = 0.201, and *Δ32/Δ32* = 3/283 = 0.011, respectively. From the genotype frequencies, we can compute the *CCR51* allele frequency as 0.89 and the frequency of the *CCR5-Δ32* allele as 0.11. Once we know the allele frequencies, we can use the Hardy–Weinberg law to determine whether this population is in equilibrium. The allele frequencies predict the genotype frequencies in the next generation as follows

Expected frequency of genotype

$$1/1 = p^2 = (0.89)^2 = 0.792$$

Expected frequency of genotype

$$1/\Delta 32 = 2pq = 2(0.89)(0.11) = 0.196$$

Expected frequency of genotype

$$\Delta 32/\Delta 32 = q^2 = (0.11)^2 = 0.012$$

These expected frequencies are nearly identical to the frequencies observed in the parental generation. Our test of this population has failed to provide evidence that Hardy–Weinberg assumptions are being violated. The conclusion can be confirmed by using the whole numbers utilized in calculating the genotype frequencies to perform a χ^2 analysis (see Chapter 3). In this case, neither the genotype frequencies nor the allele frequencies are changing in this population, meaning that the population is in equilibrium. As we will see in later sections of this chapter, forces such as natural selection, mutation, migration, and chance operate to bring about changes in allele frequency. These forces drive both microevolution and the formation of new species (macroevolution).

ESSENTIAL POINT

Populations that are not in Hardy–Weinberg equilibrium may be undergoing changes in allele frequency owing to forces such as selection, drift, migration, or nonrandom mating. ■

NOW SOLVE THIS

22–2 Determine whether the following two sets of data represent populations that are in Hardy–Weinberg equilibrium.

(a) *CCR5* genotypes: *1/1*, 60 percent; *1/Δ32*, 35.1 percent; *Δ32/Δ32*, 4.9 percent

(b) Sickle-cell hemoglobin: *SS*, 75.6 percent; *Ss*, 24.2 percent; *ss*, 0.2 percent (*S* = normal hemoglobin allele; *s* = mutant hemoglobin allele)

■ **HINT:** *This problem involves an understanding of how to use the Hardy–Weinberg law to determine whether populations are in genetic equilibrium. The key to its solution is to first determine the allele frequencies based on the genotype frequencies provided.*

Calculating Frequencies for Multiple Alleles in Populations

Although we have used one-gene two-allele systems as examples, many genes have several alleles, all of which can be found in a single population. The ABO blood group in humans (discussed in Chapter 4) is such an example. The locus *I* (isoagglutinin) has three alleles I^A, I^B, and *i*, yielding six possible genotypic combinations (I^AI^A , I^BI^B, *ii*, I^AI^B, I^Ai, I^Bi). Remember that in this case I^A and I^B are codominant alleles and that both of these are dominant to *i*. The result is that homozygous I^AI^A and heterozygous I^Ai individuals are phenotypically identical, as are I^BI^B and I^Bi individuals, so we can distinguish only four phenotypic blood-type combinations: Type A, Type B, Type AB, and Type O.

By adding another variable to the Hardy–Weinberg equation, we can calculate both the genotype and allele frequencies for the situation involving three alleles. Let *p, q,* and *r* represent the frequencies of alleles I^A, I^B, and *i*, respectively. Note that because there are three alleles

$$p + q + r = 1$$

Under Hardy–Weinberg assumptions, the frequencies of the genotypes are given by

$$(p + q + r)^2 = p^2 + q^2 + r^2 + 2pq + 2pr + 2qr = 1$$

If we know the frequencies of blood types for a population, we can then estimate the frequencies for the three alleles of the ABO system. For example, in one population sampled, the following blood-type frequencies are observed: A = 0.53, B = 0.133, O = 0.26. Because the *i* allele is recessive, the population's frequency of type O blood equals the proportion of the recessive genotype r^2. Thus,

$$r^2 = 0.26$$
$$r = \sqrt{0.26}$$
$$r = 0.51$$

Using *r,* we can calculate the allele frequencies for the I^A and I^B alleles. The I^A allele is present in two genotypes, I^AI^A and I^Ai. The frequency of the I^AI^A genotype is represented by p^2 and the I^Ai genotype by $2pr$. Therefore, the combined frequency of type A blood and type O blood is given by

$$p^2 + 2pr + r^2 = 0.53 + 0.26$$

If we factor the left side of the equation and take the sum of the terms on the right,

$$(p + r)^2 = 0.79$$
$$p + r = \sqrt{0.79}$$
$$p = 0.89 - r$$
$$p = 0.89 - 0.51 = 0.38$$

TABLE 22.3 Calculating Genotype Frequencies for Multiple Alleles in a Hardy–Weinberg Population Where the Frequency of Allele $I^A = 0.38$, Allele $I^B = 0.11$, and Allele $i = 0.51$

Genotype	Genotype Frequency	Phenotype	Phenotype Frequency
I^AI^A	$p^2 = (0.38)^2 = 0.14$	A	0.53
I^Ai	$2pr = 2(0.38)(0.51) = 0.39$		
I^BI^B	$q^2 = (0.11)^2 = 0.01$	B	0.12
I^Bi	$2qr = 2(0.11)(0.51) = 0.11$		
I^AI^B	$2pr = 2(0.38)(0.11) = 0.084$	AB	0.08
ii	$r^2 = (0.51)^2 = 0.26$	O	0.26

Having calculated p and r, the frequencies of allele I^A and allele i, we can now calculate the frequency for the I^B allele:

$$\begin{aligned} p + q + r &= 1 \\ q &= 1 - p - r \\ &= 1 - 0.38 - 0.51 \\ &= 0.11 \end{aligned}$$

The phenotypic and genotypic frequencies for this population are summarized in Table 22.3.

Calculating Heterozygote Frequency

A useful application of the Hardy–Weinberg law, especially in human genetics, allows us to estimate the frequency of heterozygotes in a population. The frequency of a recessive trait can usually be determined by identifying and counting individuals with the homozygous recessive phenotype in a sample of the population. With this information and the Hardy–Weinberg law, we can then calculate the allele and genotype frequencies for this gene.

Cystic fibrosis, an autosomal recessive trait, has an incidence of about 1/2500 (0.0004) in people of northern European ancestry. Individuals with cystic fibrosis are easily distinguished from the population at large by such symptoms as extra-salty sweat, excess amounts of thick mucus in the lungs, and susceptibility to bacterial infections. Because this is a recessive trait, individuals with cystic fibrosis must be homozygous. Their frequency in a population is represented by q^2 (provided that mating has been random in the previous generation). The frequency of the recessive allele is therefore

$$q = \sqrt{q^2} = \sqrt{0.0004} = 0.02$$

Knowing that the frequency of the recessive allele is about 2%, we can calculate the frequency of the normal (dominant) allele because $p + q = 1$. Using this equation, the frequency of p is

$$p = 1 - q = 1 - 0.02 = 0.98$$

Now that the allele frequencies are known, we can calculate the frequency of the heterozygous genotype. In the Hardy–Weinberg equation, the frequency of heterozygotes is $2pq$. Thus,

$$\begin{aligned} 2pq &= 2(0.98)(0.02) \\ &= 0.04 \text{ or 4 percent, or } 1/25 \end{aligned}$$

The results show that heterozygotes for cystic fibrosis are rather common (about 1/25 individuals, or 4 percent of the population), even though the frequency of homozygous recessives is only 1/2500, or 0.04 percent. However, keep in mind that these calculations are estimates because the population may not meet all Hardy–Weinberg assumptions.

NOW SOLVE THIS

22–3 If the albino phenotype occurs in 1/10,000 individuals in a population at equilibrium and albinism is caused by an autosomal recessive allele *a*, calculate the frequency of: (a) the recessive mutant allele; (b) the normal dominant allele; (c) heterozygotes in the population; (d) mating between heterozygotes.

HINT: *This problem involves an understanding of the method of calculating allele and genotype frequencies. The key to its solution is to first determine the frequency of the albinism allele in this population.*

22.4 Natural Selection Is a Major Force Driving Allele Frequency Change

To understand evolution, we must understand the forces that transform the gene pools of populations and can lead to the formation of new species. Chief among the mechanisms transforming populations is **natural selection,** discovered independently by Charles Darwin and Alfred Russel Wallace. The Wallace–Darwin concept of natural selection can be summarized as follows:

1. Individuals of a species exhibit variations in phenotype—for example, differences in size, agility, coloration, defenses against enemies, ability to obtain food, courtship behaviors, and flowering times.
2. Many of these variations, even small and seemingly insignificant ones, are heritable and are passed on to offspring.
3. Organisms tend to reproduce in an exponential fashion. More offspring are produced than can survive. This causes members of a species to engage in a struggle

for survival, competing with other members of the community for scarce resources. Offspring also must avoid predators, and in sexually reproducing species, adults must compete for mates.

4. In the struggle for survival, individuals with particular phenotypes will be more successful than others, allowing the former to survive and reproduce at higher rates.

As a consequence of natural selection, populations and species change. Traits that promote differential survival and reproduction will become more common, and traits that confer a lowered ability for survival and reproduction will become less common. This means that over many generations, traits that confer a reproductive advantage will increase in frequency, which in turn causes the population to become better adapted to its current environment. Over time, if selection continues, it may result in the appearance of new species.

Detecting Natural Selection in Populations

Recall that measuring allele frequencies and genotype frequencies using the Hardy–Weinberg law is based on several assumptions about an ideal population: large population size, lack of migration, presence of random mating, absence of selection and mutation, and equal survival rates of offspring.

However, if all genotypes do not have equal rates of survival or do not leave equal numbers of offspring, then allele frequencies may change from one generation to the next. To see why, let's imagine a population of 100 individuals in which the frequency of allele *A* is 0.5 and that of allele *a* is 0.5. Assuming the previous generation mated randomly, we find that the genotype frequencies in the present generation are $(0.5)^2 = 0.25$ for *AA,* $2(0.5)(0.5) = 0.5$ for *Aa,* and $(0.5)^2 = 0.25$ for *aa.* Because our population contains 100 individuals, we have 25 *AA* individuals, 50 *Aa* individuals, and 25 *aa* individuals.

Now let's suppose that individuals with different genotypes have different rates of survival: All 25 *AA* individuals survive to reproduce, 90 percent or 45/50 of the *Aa* individuals survive to reproduce, and 80 percent or 20/25 of the *aa* individuals survive to reproduce. When the survivors reproduce, each contributes two gametes to the new gene pool, giving us $2(25) + 2(45) + 2(20) = 180$ gametes. What are the frequencies of the two alleles in the surviving population? We have 50 *A* gametes from *AA* individuals, plus 45 *A* gametes from *Aa* individuals, so the frequency of allele *A* is $(50 + 45)/180 = 0.53$. We have 45 *a* gametes from *Aa* individuals, plus 40 *a* gametes from *aa* individuals, so the frequency of allele *a* is $(45 + 40)/180 = 0.47$.

These differ from the frequencies we started with. The frequency of allele *A* has increased, whereas the frequency of allele *a* has declined. A difference among individuals in survival or reproduction rate (or both) is an example of **natural selection.** Natural selection is the principal force that shifts allele frequencies within large populations by promoting differential survival and reproduction. It is one of the most important factors in evolutionary change.

Fitness and Selection

Selection occurs whenever individuals with a particular genotype enjoy an advantage in survival or reproduction over other genotypes. However, selection may vary over a wide range, from much less than 1 percent to 100 percent. In the previous hypothetical example, selection was strong. Weak selection might involve just a fraction of a percent difference in the survival rates of different genotypes. Advantages in survival and reproduction ultimately translate into increased genetic contribution to future generations. An individual organism's genetic contribution to future generations is called its **fitness.** Genotypes associated with high rates of reproductive success are said to have high fitness, whereas genotypes associated with low reproductive success are said to have low fitness.

Hardy–Weinberg analysis also allows us to examine fitness as a measure of the degree of natural selection. By convention, population geneticists use the letter *w* to represent fitness. Thus, w_{AA} represents the relative fitness of genotype *AA,* w_{Aa} the relative fitness of genotype *Aa,* and w_{aa} the relative fitness of genotype *aa.* Assigning the values $w_{AA} = 1$, $w_{Aa} = 0.9$, and $w_{aa} = 0.8$ would mean, for example, that all *AA* individuals survive, 90 percent of the *Aa* individuals survive, and 80 percent of the *aa* individuals survive, as in the previous hypothetical case.

Let's consider selection against deleterious alleles. Fitness values $w_{AA} = 1$, $w_{Aa} = 1$, and $w_{aa} = 0$ describe a situation in which *a* is a homozygous lethal allele. As homozygous recessive individuals die without leaving offspring, the frequency of allele *a* will decline. The decline in the frequency of allele *a* is described by the equation

$$q_g = \frac{q_0}{1 + gq_0}$$

where q_g is the frequency of allele *a* in generation *g,* q_o is the starting frequency of *a* (i.e., the frequency of *a* in generation zero), and *g* is the number of generations that have passed.

Figure 22–7 shows what happens to a lethal recessive allele with an initial frequency of 0.5. At first, because of the high percentage of *aa* genotypes, the frequency of allele *a* declines rapidly. The frequency of *a* is halved in only two generations. By the sixth generation, the frequency is

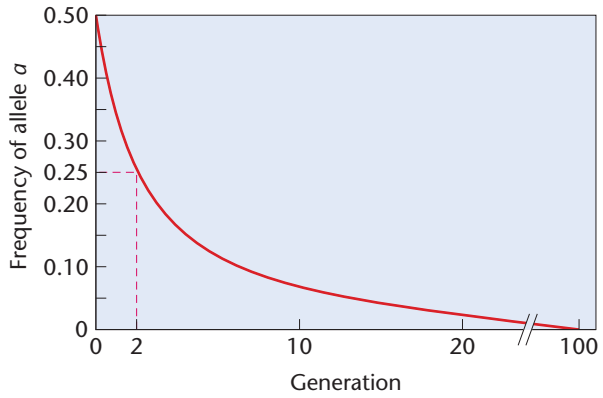


FIGURE 22–7 Change in the frequency of a lethal recessive allele, *a*. The frequency of *a* is halved in two generations and halved again by the sixth generation. Subsequent reductions occur slowly because the majority of *a* alleles are carried by heterozygotes.

halved again. By now, however, the majority of *a* alleles are carried by heterozygotes. Because *a* is recessive, these heterozygotes are not selected against. Consequently, as more time passes, the frequency of allele *a* declines ever more slowly. As long as heterozygotes continue to mate, it is difficult for selection to completely eliminate a recessive allele from a population.

Figure 22–8 shows the outcome of different degrees of selection against a nonlethal recessive allele, *a*. In this case, the intensity of selection varies from strong (red curve) to weak (blue curve), as well as intermediate values (yellow, purple, and green curves). In each example, the frequency of the deleterious allele, *a,* starts at 0.99 and declines over time. However, the rate of decline depends heavily on the strength of selection. When selection is strong and only 90 percent of the heterozygotes and 80 percent of the *aa* homozygotes survive (red curve), the frequency of allele *a* drops from 0.99 to less than 0.01 in about 85 generations. However, when selection is weak, and 99.8 percent of the heterozygotes and 99.6 percent of the *aa* homozygotes survive (blue curve), it takes 1000 generations for the frequency of allele *a* to drop from 0.99 to 0.93. Two important conclusions can be drawn from this example. First, over thousands of generations, even weak selection can cause substantial changes in allele frequencies; because evolution generally occurs over a large number of generations, selection is a powerful force in evolutionary change. Second, for selection to produce rapid changes in allele frequencies, the differences in fitness among genotypes must be large.

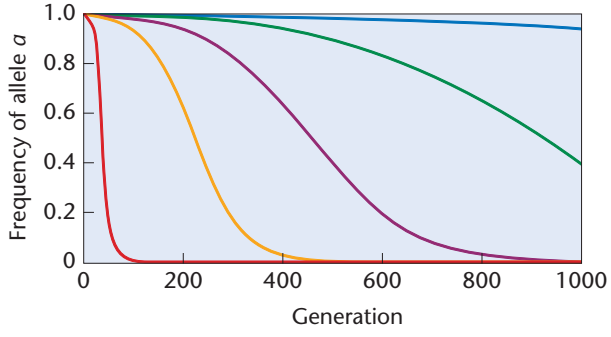


FIGURE 22–8 The effect of selection on allele frequency. The rate at which a deleterious allele is removed from a population depends heavily on the strength of selection.

There Are Several Types of Selection

The phenotype is the result of the combined influence of the individual's genotype at many different loci and the effects of the environment. Selection can be classified as (1) directional, (2) stabilizing, or (3) disruptive.

In **directional selection** traits at one end of a spectrum of phenotypes present in the population become selected for or against, usually as a result of changes in the environment. A carefully documented example comes from research by Peter and Rosemary Grant and their colleagues, who study the medium ground finches (*Geospiza fortis*) of Daphne Major Island in the Galapagos Islands. These researchers discovered that the beak size of these birds varies over time (**Figure 22–9**). In 1977, a severe drought killed some 80 percent of the finches on the island. Big-beaked birds survived at higher rates than small-beaked birds because when food became scarce, the big-beaked birds were able to eat a greater variety of seeds, especially larger ones with hard shells. After the drought ended, more plants were available, and beak size declined. Droughts in 1980 and 1982 again saw differential survival and reproduction, shifting the average beak size toward one phenotypic extreme.

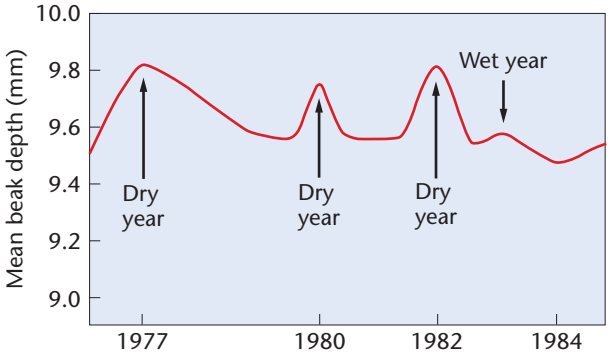


FIGURE 22–9 Beak size in finches during dry years increases because of strong selection. Between droughts, selection for large beak size is not as strong, and birds with smaller beak sizes survive and reproduce, increasing the number of birds with smaller beaks.

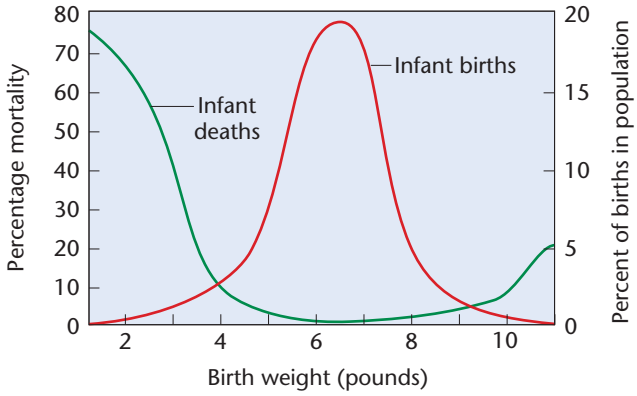


FIGURE 22–10 Relationship between birth weight and mortality in humans.

Stabilizing selection, in contrast, selects for intermediate phenotypes, with those at both extremes being selected against. Over time, this will reduce the phenotypic variance in the population but without a significant shift in the mean. One of the clearest demonstrations of stabilizing selection is from a study of human birth weight and survival for 13,730 children born over an 11-year period. **Figure 22–10** shows the distribution of birth weight, the percentage of mortality at five weeks, and the percent of births in the population (at right). Infant mortality increases on either side of the optimal birth weight of 7.5 pounds. Stabilizing selection acts to keep a population well adapted to its current environment.

Disruptive selection is selection against intermediate phenotypes and selection for phenotypes at both extremes. It can be viewed as the opposite of stabilizing selection because the intermediate types are selected against. This will result in a population with an increasingly bimodal distribution for a trait, as we can see in **Figure 22–11**. In experiments using *Drosophila,* after several generations of disruptive artificial selection for bristle number, in which only flies with high- or low-bristle numbers were allowed to breed, most flies could be easily placed in a low- or high-bristle category. In natural populations, such a situation might exist for a population in a heterogeneous environment.

ESSENTIAL POINT

The rate of change under natural selection depends on initial allele frequencies, selection intensity, and the relative fitness of different genotypes. ■

22.5 Mutation Creates New Alleles in a Gene Pool

Within a population, the gene pool is reshuffled each generation to produce new in the offspring. The enormous genetic variation present in the gene pool allows

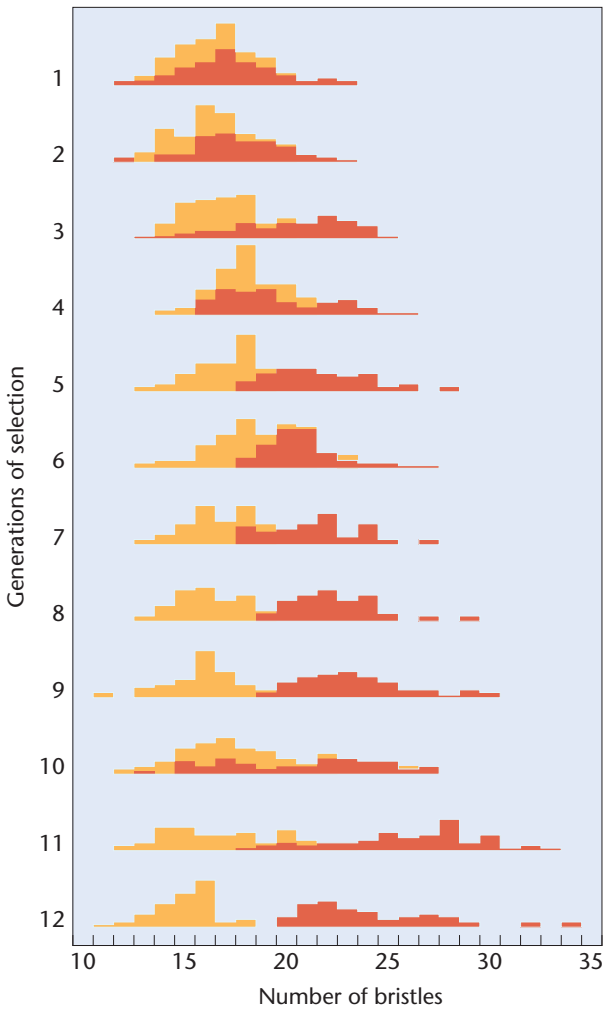


FIGURE 22–11 The effect of disruptive selection on bristle number in *Drosophila.* When individuals with the highest and lowest bristle number were selected, the population showed a nonoverlapping divergence in only 12 generations.

assortment and recombination to produce new combinations of genes already present in the gene pool. But assortment and recombination do not produce new alleles. **Mutation** alone acts to create new alleles. It is important to keep in mind that mutational events occur at random—that is, without regard for any possible benefit or disadvantage to the organism. Mutations not only create new alleles, but in very small populations can change allele frequencies. Let's consider a population of 20 individuals, and a gene with two alleles, *A* and *a*. If the frequency of *A* is 0.90, the frequency of *a* is 0.10. A mutational event changes one *A* allele into an *a* allele. This event reduces the frequency of the *A* allele from 0.90 to 0.85 and increases the frequency of the *a* allele from 0.10 to 0.15. In this section, we consider whether mutation, by itself, in the larger case, is a significant factor in changing allele frequencies.

To determine whether mutation is a significant force in changing allele frequencies, we measure the rate at which they are produced. As in our example, most mutations are recessive, so it is difficult to observe mutation rates directly in diploid organisms. Indirect methods use probability and statistics or large-scale screening programs to estimate mutation rates. For certain dominant mutations, however, a direct method of measurement can be used. To ensure accuracy, several conditions must be met:

1. The allele must produce a distinctive phenotype that can be distinguished from similar phenotypes produced by recessive alleles.
2. The trait must be fully expressed or completely penetrant so that mutant individuals can be identified.
3. An identical phenotype must never be produced by nongenetic agents such as drugs or chemicals.

Suppose that for a given gene that undergoes mutation to a dominant allele, 2 out of 100,000 births exhibit a mutant phenotype, but the parents are phenotypically normal. Because the zygotes that produced these births each carry two copies of the gene, we have actually surveyed 200,000 copies of the gene (or 200,000 gametes). If we assume that the affected births are each heterozygous, we have uncovered 2 mutant alleles out of 200,000. Thus, the mutation rate is 2/200,000 or 1/100,000, which in scientific notation is written as 1×10^{-5}. In humans, a dominant form of dwarfism known as **achondroplasia** fulfills the requirements for measuring mutation rates. Individuals with this skeletal disorder have an enlarged skull, short arms and legs, and can be diagnosed by X-ray examination at birth. In a survey of almost 250,000 births, the mutation rate (μ) for achondroplasia has been calculated as

$$\mu = 1.4 \times 10^{-5} \pm 0.5 \times 10^{-5}$$

Knowing the rate of mutation, we can estimate the extent to which mutation can change allele frequencies from one generation to the next. We represent the normal allele as *d* and the allele for achondroplasia as *D*.

Instead of a population of 20 individuals, imagine a population of 500,000 individuals in which everyone has genotype *dd*. The initial frequency of *d* is 1.0, and the initial frequency of *D* is 0. If each individual contributes two gametes to the gene pool, the gene pool will contain 1,000,000 gametes, all carrying allele *d*. Although the gametes are in the gene pool, 1.4 of every 100,000 *d* alleles mutate into a *D* allele. The frequency of allele *d* is now (1,000,000 − 14)/1,000,000 = 0.999986, and the frequency of allele *D* is 14/1,000,000 = 0.000014. From these numbers, it will clearly be a long time before mutation, by itself, causes any appreciable change in the allele frequencies in this population. In other words, mutation generates new alleles but, unless the population is very small, by itself does not alter allele frequencies at an appreciable rate.

22.6 Migration and Gene Flow Can Alter Allele Frequencies

The Hardy–Weinberg law assumes that migration does not take place. However, occasionally, **migration,** or gene flow, occurs when individuals move between the populations. Migration reduces the genetic differences between populations of a species and can increase the level of genetic variation in some populations.

Imagine a species in which a given locus has two alleles, *A* and *a*. There are two populations of this species, one on a mainland and one on an island. The frequency of *A* on the mainland is represented by p_m, and the frequency of *A* on the island is p_i. If there is migration from the mainland to the island, the frequency of *A* in the next generation on the island (p_i') is given by

$$p_i' = (1 - m)p_i + mp_m$$

where *m* represents migrants from the mainland to the island and that migration is random with respect to genotype.

As an example of how migration might affect the frequency of *A* in the next generation on the island (p_i'), assume that $p_i = 0.4$ and $p_m = 0.6$ and that 10 percent of the parents of the next generation are migrants from the mainland ($m = 0.1$). In the next generation, the frequency of allele *A* on the island will therefore be

$$\begin{aligned} p_i' &= [(1 - 0.1) \times 0.4] + (0.1 \times 0.6) \\ &= 0.36 + 0.06 \\ &= 0.42 \end{aligned}$$

In this case, the flow of genes from the mainland has changed the frequency of *A* on the island from 0.40 to 0.42 in a single generation.

These calculations reveal that the change in allele frequency attributable to migration is proportional to the differences in allele frequency between the donor and recipient populations *and* to the rate of migration. If either *m* is large or p_m is very different from p_i, then a rather large change in the frequency of *A* can occur in a single generation. If migration is the only force acting to change the allele frequency on the island, then equilibrium will be attained when $p_i = p_m$. These guidelines can often be used to estimate migration in cases where it is difficult to quantify. Even in large populations, over time, the effect of migration can substantially alter allele frequencies in populations, as shown for the I^B allele of the ABO blood group in **Figure 22–12**.

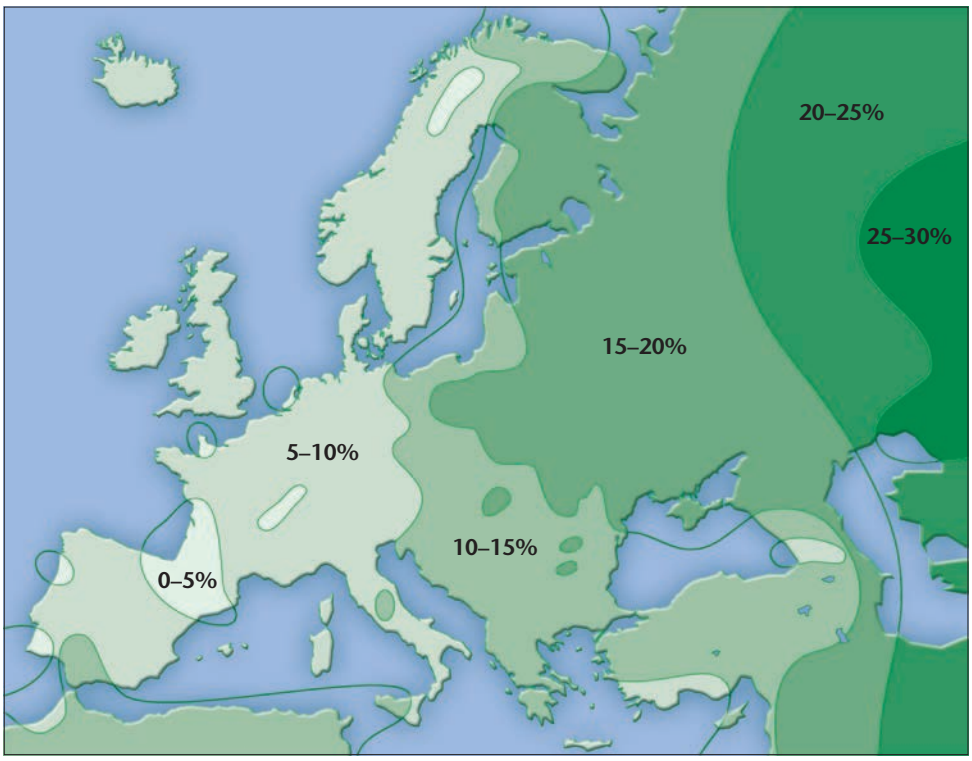


FIGURE 22–12 Migration as a force in evolution. The I^B allele of the *ABO* locus is present in a gradient from east to west. This allele shows the highest frequency in central Asia and the lowest in northeast Spain. The gradient parallels the waves of Mongol migration into Europe following the fall of the Roman Empire and is a genetic relic of human history.

22.7 Genetic Drift Causes Random Changes in Allele Frequency in Small Populations

In small populations, significant random fluctuations in allele frequencies are possible by chance alone, a situation known as **genetic drift.** In addition to small population size, drift can arise through the **founder effect,** which occurs when a population originates from a small number of individuals. Although the population may later increase to a large size, the genes carried by all members are derived from those of the founders (assuming no mutation, migration, or selection, and the presence of random mating). Drift can also arise via a **genetic bottleneck.** Bottlenecks develop when a large population undergoes a drastic but temporary reduction in numbers. Even though the population recovers, its genetic diversity has been greatly reduced. In summary, drift is a product of chance and can arise through small population size, founder effects, and bottlenecks. In the following section, we will examine how founder effects can affect allele frequencies.

Founder Effects in Human Populations

Allele frequencies in certain human populations demonstrate the role of genetic drift in natural populations. Native Americans living in the southwestern United States have a high frequency of oculocutaneous albinism (OCA). In the Navajo, who live primarily in northeast Arizona, albinism occurs with a frequency of 1 in 1500–2000, compared with whites (1 in 36,000) and African-Americans (1 in 10,000). There are four different forms of OCA (*OCA*1–4), all with varying degrees of melanin deficiency in the skin, eyes, and hair. OCA2 is caused by mutations in the *P* gene, which encodes a plasma membrane protein. To investigate the genetic basis of albinism in the Navajo, researchers screened for mutations in the *P* gene. In their study, all Navajo with albinism were homozygous for a 122.5-kb deletion in the *P* gene, spanning exons 10–20.

Using a set of PCR primers spanning the deletion, researchers were able to identify homozygous affected individuals, heterozygous carriers, and homozygous normal individuals (**Figure 22–13**). They surveyed 134 normally pigmented Navajo and 42 members of the Apache, a tribe closely related to the Navajo. Based on this sample, the heterozygote frequency in the Navajo is estimated to be 4.5 percent. No carriers were found in the Apache population that was studied.

The 122.5-kb deletion allele causing OCA2 albinism was found only in the Navajo population and not in members of other Native American tribes in the southwestern United States, suggesting that the mutant allele is specific to the Navajo and may have arisen in a single individual who was one of the small number of founders of the Navajo population. Workers originally estimated the age of the

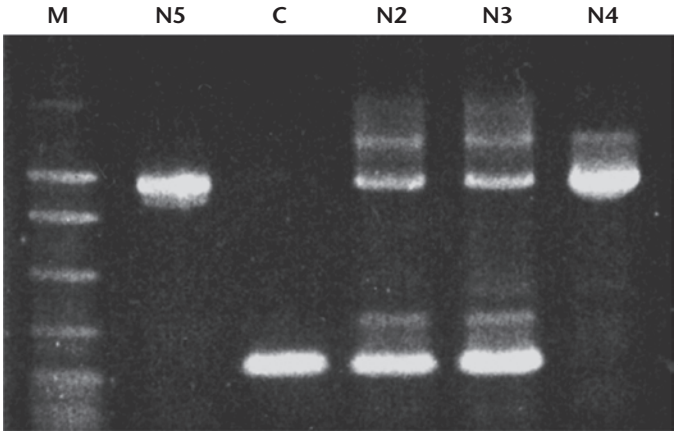


FIGURE 22–13 PCR screens of Navajo affected with albinism (N4 and N5) and the parents of N4 (N2 and N3). Affected individuals (N4 and N5) have a single, dense band at 606 bp; heterozygous carriers (N2 and N3) have two bands, one at 606 bp and one at 257 bp. The homozygous normal individual (C) has a single dense band at 257 bp. Each genotype produces a distinctive band pattern, allowing detection of heterozygous carriers in the population. Molecular size markers (M) are in the first lane.
Courtesy of Murray Brilliant, "A 122.5 kilobase deletion of P gene underlies the high prevalence of oculocutaneous albinism type 2 in the Navajo population." From: American Journal Human Genetics *72: 62–72, Figure 3, p. 67. Published by University of Chicago Press.*

mutation to be between 400 and 11,000 years, but tribal history and Navajo oral tradition indicated that the Navajo and Apache became separate populations between 600 and 1000 years ago. Because the deletion is not found in the Apaches, it probably arose in the Navajo population after the tribes split. On this basis, the deletion is estimated to be 400 to 1000 years old and probably arose as a founder mutation.

22.8 Nonrandom Mating Changes Genotype Frequency but Not Allele Frequency

We have explored how populations that do not meet the first four assumptions of the Hardy–Weinberg law, in the form of selection, mutation, migration, and genetic drift, can have changes in allele frequencies. The fifth assumption is that members of a population mate at random; in other words, any one genotype has an equal probability of mating with any other genotype in the population. Nonrandom mating can change the frequencies of genotypes in a population. Subsequent selection for or against certain genotypes has the potential to affect the overall frequencies of the alleles they contain, but it is important to note that nonrandom mating *does not itself directly change allele frequencies.*

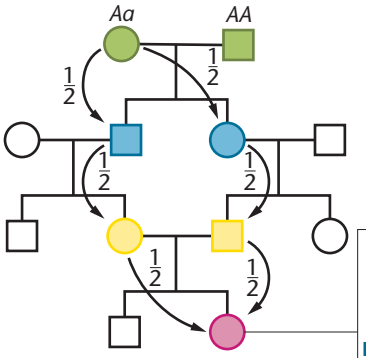
Nonrandom mating can take one of several forms. In **positive assortive mating,** similar genotypes are more likely to mate than dissimilar ones. This often occurs in humans: A number of studies have indicated that many people are more attracted to individuals who physically resemble them (and are therefore more likely to be genetically similar as well). **Negative assortive mating** occurs when dissimilar genotypes are more likely to mate; some plant species have inbuilt recognition systems that prevent fertilization between individuals with the same alleles at key loci. However, the form of nonrandom mating most commonly found to affect genotype frequencies in population genetics is **inbreeding.**

Inbreeding

Inbreeding occurs when mating individuals are more closely related than any two individuals drawn from the population at random; loosely defined, inbreeding is mating among relatives. For a given allele, inbreeding increases the proportion of homozygotes and decreases the proportion of heterozygotes in the population. A completely inbred population will theoretically consist only of homozygous genotypes. High levels of inbreeding can be harmful because it increases the probability that the number of individuals homozygous for deleterious and/or lethal alleles will increase in the population.

To describe the intensity of inbreeding in a population, Sewall Wright devised the **coefficient of inbreeding** (***F***). This coefficient quantifies the probability that the two alleles of a given gene present in an individual are identical *because they are descended from the same single copy of the allele in an ancestor.* If $F = 1$, all individuals in the population are homozygous, and both alleles in every individual are derived from the same ancestral copy. If $F = 0$, no individual has two alleles derived from a common ancestral copy.

One method of estimating F for an individual is shown in **Figure 22–14**. The fourth-generation female (shaded pink) is the daughter of first cousins (yellow). Suppose her great-grandmother (green) was a carrier of a recessive lethal allele, *a.* What is the probability that the fourth-generation female will inherit two copies of her great-grandmother's lethal allele? For this to happen, (1) the great-grandmother had to pass a copy of the allele to her son, (2) her son had to pass it to his daughter, and (3) his



The chance that this female will inherit two copies of her great-grandmother's *a* allele is

$$F = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{64}$$

Because the female's two alleles could be identical by descent from any of four different alleles,

$$F = 4 \times \frac{1}{64} = \frac{1}{16}$$

FIGURE 22–14 Calculating the coefficient of inbreeding *(F)* for the offspring of a first-cousin marriage.

daughter had to pass it to her daughter (the pink female). Also, (4) the great-grandmother had to pass a copy of the allele to her daughter, (5) her daughter had to pass it to her son, and (6) her son had to pass it to his daughter (the pink female). Each of the six necessary events has an individual probability of 1/2, and they *all* have to happen, so the probability that the pink female will inherit two copies of her great-grandmother's lethal allele is $(1/2)^6 = 1/64$. However, to calculate an overall value of F for the pink female as a child of a first-cousin marriage, remember that she could also inherit two copies of any of the other three dominant alleles present in her great-grandparents. Because any of four possibilities would give the pink female two alleles identical by descent from an ancestral copy,

$$F = 4 \times (1/64) = 1/16$$

ESSENTIAL POINT

Nonrandom mating in the form of inbreeding increases the frequency of homozygotes in the population and decreases the frequency of heterozygotes. ■

NOW SOLVE THIS

22–4 A prospective groom, who is normal, has a sister with cystic fibrosis (CF), an autosomal recessive disease. Their parents are normal. The brother plans to marry a woman who has no history of CF in her family. What is the probability that they will produce a CF child? They are both Caucasian, and the overall frequency of CF in the Caucasian population is 1/2500—that is, 1 affected child per 2500. (Assume the population meets the Hardy–Weinberg assumptions.)

■ **HINT:** *This problem involves an understanding of how recessive alleles are transmitted (see Chapter 3) and the probability of receiving a recessive allele from a heterozygous parent. The key to its solution is to first work out the probability that each parent carries the mutant allele.*

22.9 Speciation Occurs Via Reproductive Isolation

A **species** can be defined as a group of actually or potentially interbreeding organisms that is reproductively isolated in nature from all other such groups. In sexually reproducing organisms, speciation transforms the parental species into another species, or divides a single species into two or more separate species (**Figure 22–15**).

Populations within a species may carry considerable genetic variation, present as differences in alleles or allele frequencies at a variety of loci. Genetic divergence of these populations that result in different allele frequencies and/or different alleles in their gene pools can reflect the action of forces such as natural selection, mutation, and genetic drift.

When gene flow between populations is reduced or absent, the populations may diverge to the point that members of one population are no longer able to interbreed successfully with members of the other. When populations reach the point where they are reproductively isolated

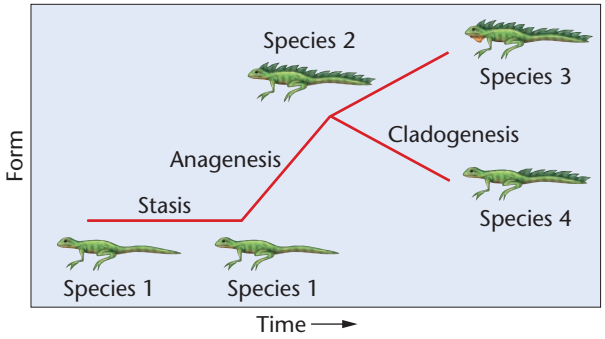


FIGURE 22–15 After a period with no change (stasis), species 1 is transformed into species 2, a process called anagenesis. Later, species 2 splits into two new species (species 3 and 4), a process called cladogenesis.

from one another, they have become different species, according to the biological species concept.

The biological barriers that prevent or reduce interbreeding between populations are called **reproductive isolating mechanisms.** These mechanisms may be ecological, behavioral, seasonal, mechanical, or physiological.

Prezygotic isolating mechanisms prevent individuals from mating in the first place. Individuals from different populations may not find each other at the right time, may not recognize each other as suitable mates, or may try to mate but find that they are unable to do so because of differences in mating behavior.

Postzygotic isolating mechanisms create reproductive isolation even when the members of two populations are willing and able to mate with each other. For example, mating may take place, and hybrid zygotes may be formed, but all or most of them may be inviable. Alternatively, the hybrids may be viable, but be sterile or suffer from reduced fertility. Yet again, the hybrids themselves may be fertile, but their progeny may have lowered viability or fertility. In all these situations, hybrids are genetic dead-ends. These postzygotic mechanisms act at or beyond the level of the zygote and are generated by genetic divergence.

Postzygotic isolating mechanisms waste gametes and zygotes and lower the reproductive fitness of hybrid survivors. Selection will therefore favor the spread of alleles that lead to the development of prezygotic isolating mechanisms, which in turn prevent interbreeding and the formation of hybrid zygotes and offspring. In animal evolution, one of the most effective prezygotic mechanisms is behavioral isolation, involving courtship behavior.

Changes Leading to Speciation

One form of speciation depends on the formation of geographic barriers between populations, which prevents gene flow between the isolated populations. Isolation allows the gene pools of these populations to diverge.

If the isolated populations later come into contact, several outcomes are possible. If reproductive isolating mechanisms are not in place, the populations will mate and will be regarded as one species. However, if reproductive isolating mechanisms have developed, the two populations will be regarded as separate species.

Formation of the Isthmus of Panama about 3 million years ago created a land bridge connecting North and South America and separated the Caribbean Sea from the Pacific Ocean. After identifying seven Caribbean species of snapping shrimp (**Figure 22–16**) and seven similar Pacific species, researchers matched them in pairs. Analysis of allele frequencies and mitochondrial DNA sequences confirmed that the ancestors of each pair were members of a single species. When the isthmus closed, each of the seven

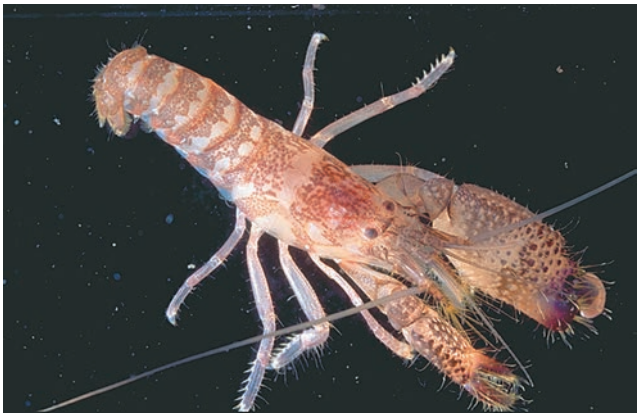


FIGURE 22–16 A snapping shrimp (genus *Alpheus*).

ancestral species was divided into two separate populations, one in the Caribbean and the other in the Pacific. But after 3 million years of separation, were members of these populations different species?

Males and females were paired together, and successful matings between Caribbean–Pacific couples versus those of Caribbean–Caribbean or Pacific–Pacific pairs were calculated. In three of the seven species pairs, transoceanic couples refused to mate altogether. Of the transoceanic pairs that mated, only 1 percent produced viable offspring, while 60 percent of same-ocean pairs produced viable offspring. We can conclude that 3 million years of separation has resulted in complete or nearly complete speciation, involving strong pre- and postzygotic isolating mechanisms for all seven species pairs.

The Rate of Macroevolution and Speciation

How much time is required for speciation? As we saw in the example above, the time needed for genetic divergence and formation of new species can occur over a span of several million years. In fact, the average time for speciation ranges from 100,000 to 10,000,000 years. However, rapid speciation over much shorter time spans has been reported in a number of cases, including fishes in East African lakes, marine salmon, palm trees on isolated islands, polyploid plants, and brown algae in the Baltic Sea.

In Nicaragua, Lake Apoyo was formed within the last 23,000 years in the crater of a volcano (**Figure 22–17**). This small lake is home to two species of cichlid fish: the Midas cichlid, *Amphilophus citrinellus*, and the Arrow cichlid, *A. zaliosus*. The Midas is the most common cichlid in the region and is found in nearby lakes; the Arrow cichlid is found only in Lake Apoyo.

To establish the evolutionary origin of the Arrow cichlid, researchers used a variety of approaches, including phylogenetic, morphological, and ecological analyses. Sequence analysis of mitochondrial DNA established that the two species form a group with a common ancestor (a

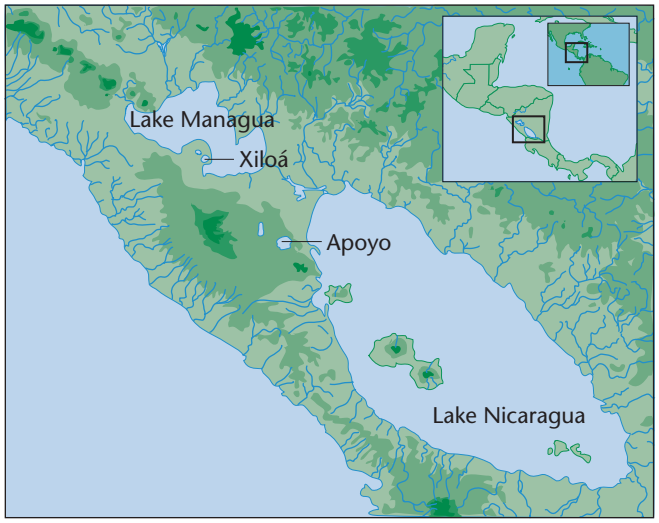


FIGURE 22–17 Lake Apoyo in Nicaragua occupies the crater of an inactive volcano. The lake formed about 23,000 years ago. Two species of cichlid fish in the lake share a close evolutionary relationship.

monophyletic group). Further genomic analysis of both species using a PCR-based method strengthened the conclusion that these two species are monophyletic and that *A. zaliosus* evolved from *A. citrinellus*. Members of the two species have distinctive morphologies (**Figure 22–18**), including jaw specializations that reflect different food preferences, which were confirmed by analysis of stomach contents. In addition, the two species are reproductively isolated, a conclusion substantiated by laboratory experiments. Using a molecular clock calibrated for cichlid mtDNA, researchers have estimated that *A. zaliosus* evolved from *A. citrinellus* sometime within the last 10,000 years. This estimate, and examples from other species, provide unambiguous evidence that, depending on the strength of selection and that of other parameters of the Hardy–Weinberg law, species formation can occur over a much shorter time scale than the usual range of 100,000–10,000,000 years.

22.10 Phylogeny Can Be Used to Analyze Evolutionary History

Speciation is associated with genetic divergence of populations. Therefore, we should be able to use genetic differences and similarities among present-day species to reconstruct their evolutionary histories. These relationships are most often presented in the form of phylogenetic trees (**Figure 22–19**), which show the ancestral relationships among a group of organisms. These groups can be species, or larger groups such as phyla. In a phylogenetic tree, branches represent the relationships among lineages over time. The length of a branch can be derived from a time scale, showing the length of time between speciation

(a)

(b)

FIGURE 22–18 The two species of cichilds in Lake Apoyo exhibit distinctive morphologies: (a) *Amphilophus citrinellus,* (b) *Amphilophus zaliosus.*

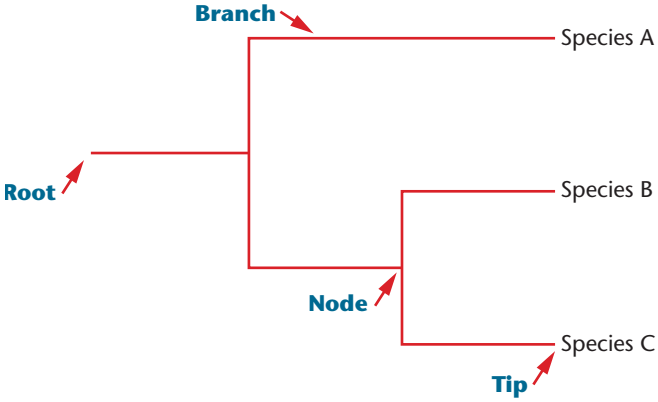


FIGURE 22–19 Elements of a phylogenetic tree showing the relationships among several species. The root represents a common ancestor to all species on the tree. Branches represent lineages through time. The points at which the branches separate are called nodes, and at the tips of the branches are the living or extinct species.

events. Branch points, or nodes, show when a species split into two or more species. Each node represents a common ancestor of the species diverging at that node. The tips of the branches represent species (or a larger group) alive today (or those that ended in extinction). Groups that consist of an ancestral species and all its descendants are called monophyletic groups. The root of a phylogenetic tree represents the oldest common ancestor to all the groups shown in the tree. Trees can be constructed from differences in morphology of living organisms, from fossils, and the molecular sequences of proteins, RNA, and DNA.

Constructing Phylogenetic Trees from DNA Sequences

Advances in DNA sequencing technology have made genetic and genomic information from many species available, and today, most phylogenetic trees are constructed using DNA sequences.

Constructing a species-level phylogenetic tree using DNA sequences requires three steps:

1. DNA sequences representing a gene or genome of interest from a number of different species must be acquired. With the proliferation of DNA sequencing projects, these are usually available from public databases.
2. The sequences must be aligned with each other so that the related parts of each sequence can be compared to see if they are the same or different. The sequences to be compared can be imported into software programs that maximize the number of aligned base pairs by inserting gaps as needed. As discussed earlier, more distantly related species have acquired more DNA differences because of the longer time that has elapsed since they last shared a common ancestor. More closely related species have fewer DNA differences because there has been less time for accumulation of DNA differences since they last shared a common ancestor.
3. These DNA differences are used to construct a phylogenetic tree, often beginning with the most closely related sequences and working backwards through sequences that are less closely related.

Reconstructing Vertebrate Evolution by Phylogenetic Analysis

One of the most important steps in the evolutionary history of our species was the ancient transition of vertebrates from the ocean to the land. For more than a century, biologists have debated and argued about which group of lobe-finned fish crawled ashore as the ancestor of all terrestrial vertebrates (amphibians, reptiles, birds, and mammals). In past years, phylogenetic trees constructed from the fossil record, from living species, and from mitochondrial DNA sequences pointed to the lungfish (**Figure 22–20**) as the closest living relative to terrestrial vertebrates, but could not rule out the possibility that vertebrates may have two common ancestors, the lungfish and another organism, the coelacanth (Figure 22–20).

(a)

(b)

FIGURE 22-20 Phylogenetic evidence indicates that the lungfish (a) and not the coelacanth (b) is a common ancestor of amphibians, reptiles, birds. and mammals.

Recently, the coelacanth genome has been sequenced, and the data from this study have reopened the question of which group shares a common ancestor with our species and all other land vertebrates. Using sequence data from the coelacanth, the lungfish, and selected vertebrate species, researchers aligned and analyzed information from 251 protein-coding genes to construct a phylogenetic tree (**Figure 22–21**). The results strongly support earlier work indicating that terrestrial vertebrates are more closely related to the lungfish than to the coelacanth. Thus, the door has been closed on this important evolutionary question.

Molecular Clocks Measure the Rate of Evolutionary Change

In many cases, we would like to estimate not only which members of a set of species are most closely related, but

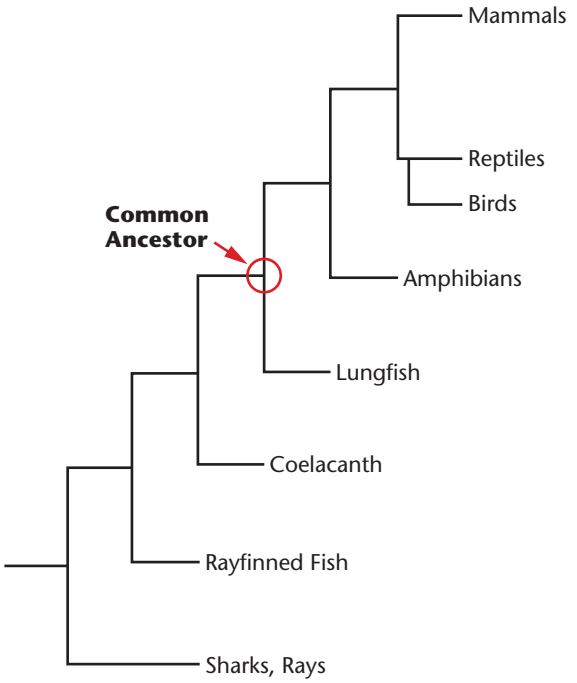


FIGURE 22-21 A phylogenetic tree of selected jawed vertebrates, including the lungfish and the coelacanth, shows that the lungfish shares the most recent common ancestor with these vertebrates.

also when their common ancestors lived. The ability to construct phylogenetic trees from protein and nucleic acid sequences led to the development of molecular clocks, which use the rate of change in amino acid or nucleotide sequences as a way to estimate the time of divergence from a common ancestor.

To be useful, molecular clocks must be carefully calibrated. Molecular clocks can only measure changes in amino acids or nucleotides; they are linear over certain time scales, and times and dates must be added to the clock using independent evidence such as the fossil record. **Figure 22–22** shows a molecular clock showing divergence times for humans and other vertebrates based on the fossil record [Figure 22–22(a)] and molecular data [Figure 22–22(b)]. In both cases, changes in amino acid sequence and nucleotide sequence increase linearly with time.

The Complex Origins of Our Genome

Current fossil, molecular, and genomic evidence indicates that our species, *Homo sapiens*, arose in Africa about 200,000 years ago from earlier species of *Homo*. When populations of *H. sapiens* first expanded out of Africa sometime between 50,000 and 70,000 years ago, parts of Europe and Asia were already occupied by members of other human species. Advances in DNA sequencing technology and new methods of DNA extraction that allow the recovery of genomic DNA from fossil remains have created a new field, called **paleogenomics,** which in turn, has revolutionized the study of human evolution. The genomes of two extinct groups who lived in the Middle East, Asia, and Europe, the Neanderthals and the Denisovans, have been sequenced and compared with the genomes of present-day humans. The results show that modern human populations outside Africa, including those of the Middle East, Europe, Asia, Australia/Oceania, and the Americas, carry sequences from these two groups. Here is what we know about the genome of the Neanderthals and the contributions they made to our genome.

The first Neanderthal genome was assembled in 2010 from three skeletons discovered in a Croatian cave. Since then, genomes from several other Neanderthals have been

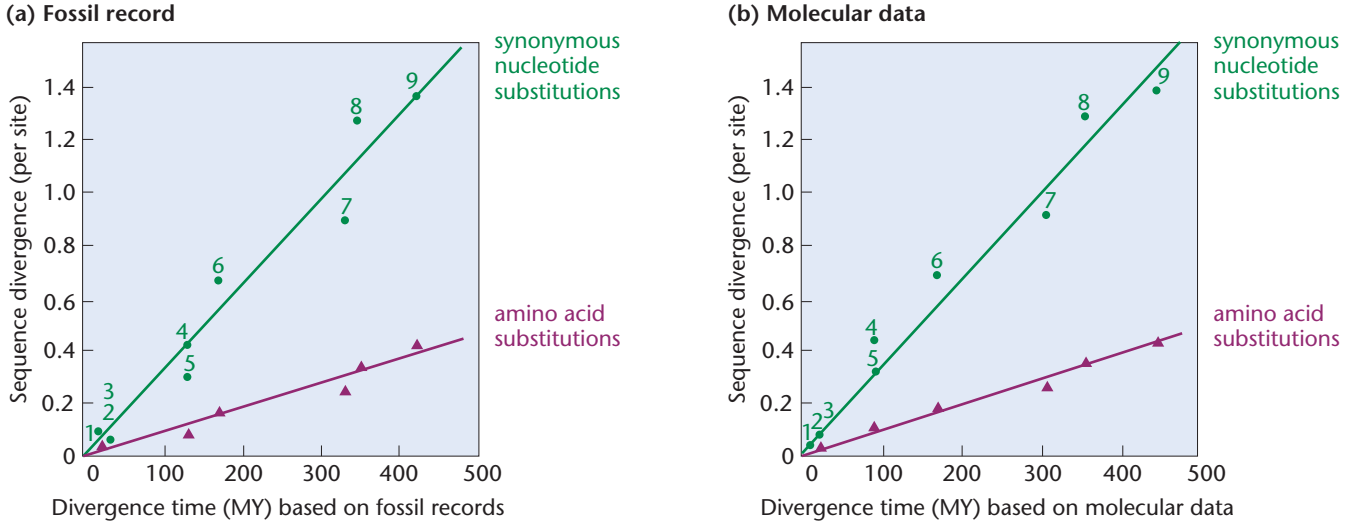


FIGURE 22–22 Relationship between the number of amino acid substitutions and the number of nucleotide substitutions for 4198 nuclear genes from 10 vertebrate species. Humans versus (1) chimpanzee, (2) orangutan, (3) macaque, (4) mouse, (5) cow, (6) opossum, (7) chicken, (8) western clawed frog, and (9) zebrafish. In (a) the data are calculated by divergence times based on the fossil record, and in (b), based on synonymous nucleotide substitutions, which are mutations that do not result in any changes in the amino acid sequence of a protein.

sequenced. Comparative genome analysis shows that the genomes of our species and the Neanderthals are the same size (about 3.2 billion base pairs) and are 99.7 percent identical.

Populations of *H. neanderthalensis* lived in Europe and western Asia from some 300,000 years ago until they disappeared about 40,000 years ago. For at least 30,000 years, Neanderthals coexisted with anatomically modern humans (*H. sapiens*) in regions of the Middle East and Europe, providing an opportunity for interbreeding between these species. In fact, gene flow from extinct Neanderthals to modern humans through interbreeding is estimated to represent 2 percent of the genome of non-African populations. Thus, the 99.7 percent sequence identity between the two species includes the 2 percent contributed by Neanderthals that has become fixed in the genome of our species. However, different individuals carry different portions of the Neanderthal genome; taken together, upward of 20 percent of the Neanderthal genome may be present in the genomes of modern non-African populations.

From these studies, two conclusions can be drawn. First, Neanderthals are not direct ancestors of our species. Second, Neanderthals and members of our species did interbreed, and Neanderthals contributed to our genome. Thus, although Neanderthals are extinct, some of their DNA has survived and is a fixed part of our genome.

In 2008, human fossils were discovered in a cave near Denisova, Siberia (**Figure 22–23**). A complete mtDNA genome sequence showed that these fossils belonged to a group separate from both Neanderthals and our species. They were named the Denisovans. A nuclear Denisovan genome sequence shows that they are more closely related to Neanderthals than to our species. In addition, the Denisovan genome contains sequences from another, as yet unknown, archaic group that made no contribution to the Neanderthal genome.

Analysis of modern human populations shows that 4 to 6 percent of the DNA in the genomes of residents of the Melanesian Islands in the South Pacific is derived from the Denisovans, and smaller amounts of Denisovan DNA are found in the genomes of Australian aborigines, as well as Polynesians, Fujians, east Indonesians, and some populations of East Asia. As things stand now, we know that as a result of gene flow, some members of our species outside of Africa carry DNA from one or two other human groups (Figure 22–23).

FIGURE 22–23 The cave in Denisova, Siberia, where the Denisovan fossils were discovered.

The Neanderthal and Denisovan genomes were assembled from fossil remains that are 40,000 to 80,000 years old. The recent sequencing of a genome from a 700,000-year-old horse fossil opens the possibility that genome sequences can be recovered from fossils of much older human species. For now, using the paleogenomic techniques currently available, we can expect exciting answers to questions about the similarities and differences between our genome and those of other human species, providing revolutionary insights into the evolution of our species and other human species that preceded us on this planet.

GENETICS, TECHNOLOGY, AND SOCIETY

Tracking Our Genetic Footprints out of Africa

Based on the physical traits and distribution of hominid fossils, most paleoanthropologists agree that a large-brained, tool-using hominid they call *Homo erectus* appeared in east Africa about 2 million years ago. This species used simple stone tools and hunted, but did not fish, build houses, or follow ritual burial practices. About 1.7 million years ago, *H. erectus* spread into Eurasia and south Asia. Most scientists also agree that *H. erectus* likely developed into *H. heidelbergensis*—a species that became the ancestor to our species (in Africa), Neanderthals (in Europe), and Denisovans (in Asia). These hominids were anatomically robust, with large, heavy skeletons and skulls. These groups disappeared 50,000 to 30,000 years ago—around the same time that anatomically modern humans (*H. sapiens*) appeared all over the world. The events that led to the appearance of *H. sapiens* throughout Europe and Asia are a source of controversy.

At present, two main hypotheses explain the origins of modern humans: the multiregional hypothesis and the out-of-Africa hypothesis. The multiregional hypothesis is based primarily on archaeological and fossil evidence. It proposes that *H. sapiens* developed gradually and simultaneously all over the world from existing *H. heidelbergensis* groups, including

Neanderthals. Interbreeding between these groups eventually made *H. sapiens* a genetically homogeneous species. Natural selection then created the regional variants that we see today. In the multiregional view, our genetic makeup should include significant contributions from many *H. heidelbergensis* groups, including Neanderthals. In contrast, the out-of-Africa hypothesis, based primarily on genetic analyses of modern human populations, contends that *H. sapiens* evolved from the descendants of *H. heidelbergensis* in sub-Saharan Africa about 200,000 years ago. A small band of *H. sapiens* (probably fewer than 1000) then left Africa around 50,000 years ago. By 40,000 years ago, they had reached Europe, Asia, and Australia. In the out-of-Africa model, *H. sapiens* replaced all existing hominins. In this way, *H. sapiens* became the only species in the genus by about 30,000 years ago.

Although the out-of-Africa hypothesis is still debated, most genetic evidence appears to support it. Humans all over the globe are remarkably similar genetically. DNA sequences from any two people chosen at random are 99.9 percent identical. More genetic identity exists between two persons chosen at random from a human population than between two chimpanzees chosen at random from a chimpanzee population. Interestingly, about 90 percent of the genetic differences that do exist occur between individuals rather than between populations. This unusually high degree of genetic relatedness in all humans around the world supports the idea that our species arose recently from a small founding group of humans.

Studies of mitochondrial DNA sequences from current human populations reveal that the highest levels of genetic variation occur within African populations. Africans show twice the mitochondrial DNA sequence diversity of non-Africans. This implies that the earliest branches of *H. sapiens* diverged in Africa and had a longer time to accumulate mitochondrial DNA mutations, which are thought to accumulate at a constant rate over time.

DNA sequences from mitochondrial, Y-chromosome, and chromosome-21 markers support the idea that human roots are in east Africa and that the migration out of Africa occurred through Ethiopia, along the coast of the Arabian Peninsula, and outward to Eurasia and Southeast Asia. Recent data based on nuclear microsatellite variants and whole genome single-nucleotide polymorphism (SNP) analysis further support the notion that humans migrated out of Africa and dispersed throughout the world from a small founding population.

As with any explanation of human origins, the out-of-Africa hypothesis is actively debated. As methods to sequence DNA from ancient fossils improve, it may be possible to fill the gaps in the genetic pathway leading out of Africa and to resolve those age-old questions about our origins.

Your Turn

1. Recent sequencing data suggest that there was some interbreeding between *H. sapiens,* Neanderthals and Denisovans. Discuss the evidence that supports the idea that these groups interbred. How might interbreeding have affected the survival of *H. sapiens* out of Africa?

Start your investigations by reading Kelso, J. and Prufer, K. 2014. Ancient humans and the origin of modern humans. *Curr. Opin. Genet. Dev*. 29: 133–138.

2. If all people on Earth are very similar genetically, how did we come to have the range of physical differences, which some describe as racial differences? How has modern genomics contributed to the debate about the validity and definition of the term *race*?

For an interesting discussion of race, human variation, and genomic studies, see Lewontin, R.C. 2006. Confusion about human races, on the *Social Sciences Research Center website*—raceandgenomics.ssrc.org/Lewontin. *A study of the genetic differences between human population groups can be found in* Witherspoon, D.J. et al. 2007. Genetic similarities within and between human populations. *Genetics 176(1): 351–359.*

3. Geneticists study mitochondrial and Y-chromosome DNA to determine the ancestry of modern humans. Why are these two types of DNA used in lineage studies? What is meant by the terms *mitochondrial Eve* and *Y-chromosome Adam*?

To read the original paper hypothesizing a mitochondrial Eve, see Cann, R.L. et al. 1987. Mitochondrial DNA and human evolution. *Nature* 325: 31–36. *For a discussion of Y-chromosome Adam, see* Gibbons, A. 1997. Y Chromosome shows that Adam was an African. *Science* 278: 804–805.

CASE STUDY | An unexpected outcome

A newborn screening program identified a baby with a rare autosomal recessive disorder called arginosuccinate aciduria (AGA), which causes high levels of ammonia to accumulate in the blood. Symptoms usually appear in the first week after birth and can progress to include severe liver damage, developmental delay, and mental retardation. AGA occurs with a frequency of about 1 in 70,000 births. There is no history of this disorder in either the father's or mother's family. This case raises several questions:

1. Since it appears that the unaffected parents are heterozygotes, would it be considered unusual that there would be no family history of the disorder? How would they be counseled about risks to future children?
2. If the disorder is so rare, what is the frequency of heterozygous carriers in the population?
3. What are the chances that two heterozygotes will meet and have an affected child?

INSIGHTS AND SOLUTIONS

1 Tay–Sachs disease is caused by loss-of-function mutations in a gene on chromosome 15 that encodes a lysosomal enzyme. Tay-Sachs is inherited as an autosomal recessive condition. Among Ashkenazi Jews of Central European ancestry, about 1 in 3600 children is born with the disease. What fraction of the individuals in this population are carriers?

Solution: If we let p represent the frequency of the wild-type enzyme allele and q the total frequency of recessive loss-of-function alleles, and if we assume that the population is in Hardy–Weinberg equilibrium, then the frequencies of the genotypes are given by p^2 for homozygous normal, $2pq$

(continued)

Insights and Solutions—continued

for carriers, and q^2 for individuals with Tay–Sachs. The frequency of Tay–Sachs alleles is thus

$$q = \sqrt{q^2} = \sqrt{\frac{1}{3600}} = 0.017$$

Since $p + q = 1$, we have

$$p = 1 - q = 1 - 0.017 = 0.983$$

Therefore, we can estimate that the frequency of carriers is

$$2pq = 2(0.983)(0.017) = 0.033 \text{ or about 1 in 30}$$

2 A single plant twice the size of others in the same population suddenly appears. Normally, plants of that species reproduce by self-fertilization and by cross-fertilization. Is this new giant plant simply a variant, or could it be a new species? How would you determine which it is?

Solution: One of the most widespread mechanisms of speciation in higher plants is polyploidy, the multiplication of entire sets of chromosomes. The result of polyploidy is usually a larger plant with larger flowers and seeds. There are two ways of testing the new variant to determine whether it is a new species. First, the giant plant should be crossed with a normal-sized plant to see whether the giant plant produces viable, fertile offspring. If it does not, then the two different types of plants would appear to be reproductively isolated. Second, the giant plant should be cytogenetically screened to examine its chromosome complement. If it has twice the number of its normal-sized neighbors, it is a tetraploid that may have arisen spontaneously. If the chromosome number differs by a factor of two and the new plant is reproductively isolated from its normal-sized neighbors, it is a new species.

Problems and Discussion Questions

MasteringGenetics™ Visit for instructor-assigned tutorials and problems.

HOW DO WE KNOW?

1. Population geneticists study changes in the nature and amount of genetic variation in populations, the distribution of different genotypes, and how forces such as selection and drift act on genetic variation to bring about evolutionary change in populations and the formation of new species. From the explanation given in the chapter, what answers would you propose to the following fundamental questions?
   (a) How do we know how much genetic variation is in a population?
   (b) How do geneticists detect the presence of genetic variation as different alleles in a population?
   (c) How do we know whether the genetic structure of a population is static or dynamic?
   (d) How do we know when populations have diverged to the point that they form two different species?
   (e) How do we know the age of the last common ancestor shared by two species?

CONCEPT QUESTION

2. Review the Chapter Concepts on page 457. All these pertain to the principles of population genetics and the evolution of species. Write a short essay describing the roles of mutation, migration, and selection in bringing about speciation. ■

3. Price et al. (1999. *J. Bacteriol.* 181: 2358–2362) conducted a genetic study of the toxin transport protein (PA) of *Bacillus anthracis,* the bacterium that causes anthrax in humans. Within the 2294-nucleotide gene in 26 strains they identified five point mutations—two missense and three synonyms—among different isolates. Necropsy samples from an anthrax outbreak in 1979 revealed a novel missense mutation and five unique nucleotide changes among ten victims. The authors concluded that these data indicate little or no horizontal transfer between different *B. anthracis* strains.
   (a) Which types of nucleotide changes (missense or synonyms) cause amino acid changes?
   (b) What is meant by horizontal transfer?
   (c) On what basis did the authors conclude that evidence of horizontal transfer is absent from their data?
4. The genetic difference between two *Drosophila* species, *D. heteroneura* and *D. sylvestris,* as measured by nucleotide diversity, is about 1.8 percent. The difference between chimpanzees (*P. troglodytes*) and humans (*H. sapiens*) is about the same, yet the latter species are classified in different genera. In your opinion, is this valid? Explain why.
5. The use of nucleotide sequence data to measure genetic variability is complicated by the fact that the genes of higher eukaryotes are complex in organization and contain 5′ and 3′ flanking regions as well as introns. Researchers have compared the nucleotide sequence of two cloned alleles of the *γ-globin* gene from a single individual and found a variation of 1 percent. Those differences include 13 substitutions of one nucleotide for another and 3 short DNA segments that have been inserted in one allele or deleted in the other. None of the changes takes place in the gene's exons (coding regions). Why do you think this is so, and should it change our concept of genetic variation?
6. Calculate the frequencies of the *AA, Aa,* and *aa* genotypes after one generation if the initial population consists of 0.2 *AA,* 0.6 *Aa,* and 0.2 *aa* genotypes and meets the requirements of the Hardy–Weinberg relationship. What genotype frequencies will occur after a second generation?
7. Consider rare disorders in a population caused by an autosomal recessive mutation. From the frequencies of the disorder in the population given, calculate the percentage of heterozygous carriers.
   (a) 0.0064
   (b) 0.000081
   (c) 0.09
   (d) 0.01
   (e) 0.10
8. What must be assumed in order to validate the answers in Problem 7?
9. In a population that meets the Hardy–Weinberg equilibrium assumptions, 81% of the individuals are homozygous for a recessive allele. What percentage of the individuals would be expected to be heterozygous for this locus in the next generation?
10. In a population of cattle, the following color distribution was noted: 36% red (*RR*), 48% roan (*Rr*), and 16% white (*rr*). Is this population in a Hardy–Weinberg equilibrium? What will be the distribution of genotypes in the next generation if the Hardy–Weinberg assumptions are met?
11. Consider a population in which the frequency of allele *A* is $p = 0.7$ and the frequency of allele *a* is $q = 0.3$, and where the

alleles are codominant. What will be the allele frequencies after one generation if the following occurs?

(a) $w_{AA} = 1, w_{Aa} = 0.9$, and $w_{aa} = 0.8$

(b) $w_{AA} = 1, w_{Aa} = 0.95$, and $w_{aa} = 0.9$

(c) $w_{AA} = 1, w_{Aa} = 0.99, w_{aa} = 0.98$

(d) $w_{AA} = 0.8, w_{Aa} = 1, w_{aa} = 0.8$

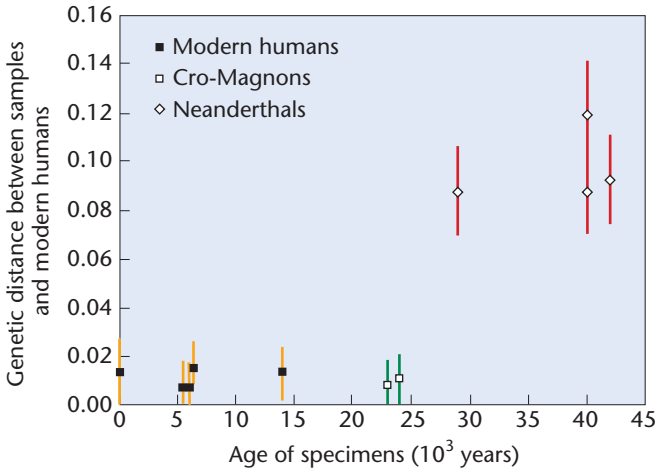
12. In a population of 10,000 individuals, where 3600 are *MM*, 1600 are *NN*, and 4800 are *MN*, what are the frequencies of the *M* alleles and the *N* alleles?
13. Under what circumstances might a lethal dominant allele persist in a population?
14. A certain form of albinism in humans is recessive and autosomal. Assume that 1% of the individuals in a given population are albino. Assuming that the population is in Hardy–Weinberg equilibrium, what percentage of the individuals in this population is expected to be heterozygous?
15. One of the first Mendelian traits identified in humans was a dominant condition known as *brachydactyly*. This gene causes an abnormal shortening of the fingers or toes (or both). At the time, some researchers thought that the dominant trait would spread until 75 percent of the population would be affected (because the phenotypic ratio of dominant to recessive is 3:1). Show that the reasoning was incorrect.
16. What is the original source of genetic variation in a population? Which natural factors affect changes in this original variation?
17. Achondroplasia is a dominant trait that causes a characteristic form of dwarfism. In a survey of 50,000 births, five infants with achondroplasia were identified. Three of the affected infants had affected parents, while two had normal parents. Calculate the mutation rate for achondroplasia and express the rate as the number of mutant genes per given number of gametes.
18. A recent study examining the mutation rates of 5669 mammalian genes (17,208 sequences) indicates that, contrary to popular belief, mutation rates among lineages with vastly different generation lengths and physiological attributes are remarkably constant (Kumar, S., and Subramanian, S. 2002. *Proc. Natl. Acad. Sci. [USA]* 99: 803–808). The average rate is estimated at 12.2×10^{-9} per bp per year. What is the significance of this finding in terms of mammalian evolution?
19. A form of dwarfism known as Ellis–van Creveld syndrome was first discovered in the late 1930s, when Richard Ellis and Simon van Creveld shared a train compartment on the way to a pediatrics meeting. In the course of conversation, they discovered that they each had a patient with this syndrome. They published a description of the syndrome in 1940. Affected individuals have a short-limbed form of dwarfism and often have defects of the lips and teeth, and polydactyly (extra fingers). The largest pedigree for the condition was reported in an Old Order Amish population in eastern Pennsylvania by Victor McKusick and his colleagues (1964). In that community, about 5 per 1000 births are affected, and in the population of 8000, the observed frequency is 2 per 1000. All affected individuals have unaffected parents, and all affected cases can trace their ancestry to Samuel King and his wife, who arrived in the area in 1774. It is known that neither King nor his wife was affected with the disorder. There are no cases of the disorder in other Amish communities, such as those in Ohio or Indiana.

    (a) From the information provided, derive the most likely mode of inheritance of this disorder. Using the Hardy–Weinberg law, calculate the frequency of the mutant allele in the population and the frequency of heterozygotes, assuming Hardy–Weinberg conditions.

    (b) What is the most likely explanation for the high frequency of the disorder in the Pennsylvania Amish community and its absence in other Amish communities?
20. List the barriers that prevent interbreeding and give an example of each.
21. Present a rationale for using DNA sequence polymorphisms as an index of genetic diversity. Is genetic diversity directly proportional to evolutionary (phylogenetic) diversity?
22. Are there nucleotide substitutions that will not be detected by electrophoretic studies of a gene's protein product?
23. In a recent study of cichlid fish inhabiting Lake Victoria in Africa, Nagl et al. (1998. *Proc. Natl. Acad. Sci. [USA]* 95: 14,238–14,243) examined suspected neutral sequence polymorphisms in noncoding genomic loci in 12 species and their putative river-living ancestors. At all loci, the same polymorphism was found in nearly all of the tested species from Lake Victoria, both lacustrine and riverine. Different polymorphisms at these loci were found in cichlids at other African lakes.

    (a) Why would you suspect neutral sequences to be located in noncoding genomic regions?

    (b) What conclusions can be drawn from these polymorphism data in terms of cichlid ancestry in these lakes?
24. What genetic changes take place during speciation?
25. Some critics have warned that the use of gene therapy to correct genetic disorders will affect the course of human evolution. Evaluate this criticism in light of what you know about population genetics and evolution, distinguishing between somatic gene therapy and germ-line gene therapy.
26. Comparisons of Neanderthal mitochondrial DNA with that of modern humans indicate that they are not related to modern humans and did not contribute to our mitochondrial heritage. However, because Neanderthals and modern humans are separated by at least 25,000 years, this does not rule out some forms of interbreeding causing the modern European gene pool to be derived from both Neanderthals and early humans (called Cro-Magnons). To resolve this question, Caramelli et al. (2003. *Proc. Natl. Acad. Sci. [USA]* 100: 6593–6597) analyzed mitochondrial DNA sequences from 25,000-year-old Cro-Magnon remains and compared them to four Neanderthal specimens and a large dataset derived from modern humans. The results are shown in the graph.



The *x*-axis represents the age of the specimens in thousands of years; the *y*-axis represents the average genetic distance. Modern humans are indicated by filled squares; Cro-Magnons, open squares; and Neanderthals, diamonds.

(a) What can you conclude about the relationship between Cro-Magnons and modern Europeans? What about the relationship between Cro-Magnons and Neanderthals?

(b) From these data, does it seem likely that Neanderthals made any mitochondrial DNA contributions to the Cro-Magnon gene pool or the modern European gene pool?

Epigenetics

The somatic cells of the human body contain 20,000 to 25,000 genes. In the more than 200 cell types present in the body, different cell-specific gene sets are transcribed, while the rest of the genome is transcriptionally inactive. During development, as embryonic cells gradually become specialized cells and exhibit adult phenotypes, programs of gene expression become increasingly restricted. Until recently, it was thought that most regulation of gene expression is coordinated by *cis*-regulatory elements as well as DNA-binding proteins and transcription factors, and that this regulation can occur at any of the steps in gene expression (see Chapter 14). However, as we have learned more about genome organization and the regulation of gene expression, it is clear that classical regulatory mechanisms cannot fully explain how some phenotypes arise or why phenotypes change during the life cycle. For example, monozygotic twins have identical genotypes but often develop different phenotypes. In addition, although one allele of each gene is inherited maternally and one is inherited paternally, in some cases, only the maternal or paternal allele is expressed, while the other is transcriptionally silent.

The newly emerging field of epigenetics is providing us with a basis for understanding how heritable changes other than those in DNA sequence can influence phenotypic variation (**ST Figure 1–1**). These advances greatly extend our understanding of the molecular basis of gene regulation and have application in wide-ranging areas including genetic disorders, cancer, and behavior.

An **epigenetic trait** is a stable, mitotically and meiotically heritable phenotype that results from changes in gene expression without alterations in the DNA sequence. **Epigenetics** is the study of the ways in which these changes alter cell- and tissue-specific patterns of gene expression (see Box 1). Epigenetic regulation of gene expression uses reversible modifications of DNA and chromatin structure to mediate the interaction of the genome with a variety of environmental factors and generates changes in the patterns of gene expression in response to these factors. The **epigenome** refers to the epigenetic state of a cell. During its life span, an organism has one genome, but this genome can be modified in diverse cell types at different times to produce many epigenomes.

The newly emerging field of epigenetics is providing us with a basis for understanding how heritable changes other than those in DNA sequence can influence phenotypic variation.

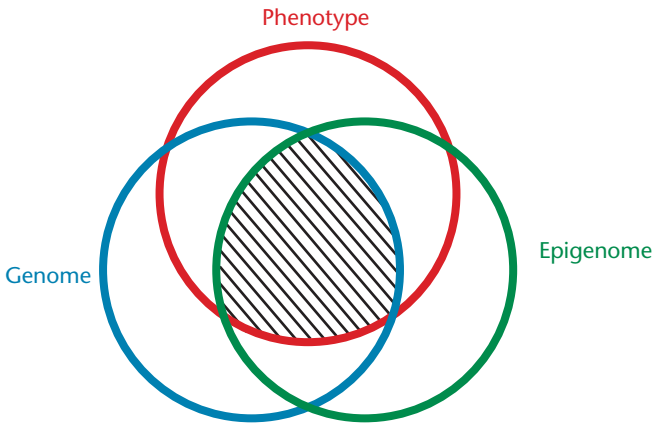
Current research efforts are focused on several aspects of epigenetics: how an epigenome arises in developing and differentiated cells and how these epigenomes are transmitted via mitosis and meiosis, making them heritable traits. In addition, because epigenetically controlled alterations to the genome are associated with common diseases such as cancer, diabetes, and asthma, efforts are also being directed toward developing drugs that can modify or reverse disease-associated epigenetic changes in cells.

Here we will focus on the association of epigenetics with some heritable genetic disorders, cancer, and environment–genome interactions. Because epigenetic changes are potentially reversible, we will also examine how knowledge of molecular mechanisms of epigenetics is being used to develop drugs and treatments for human diseases.

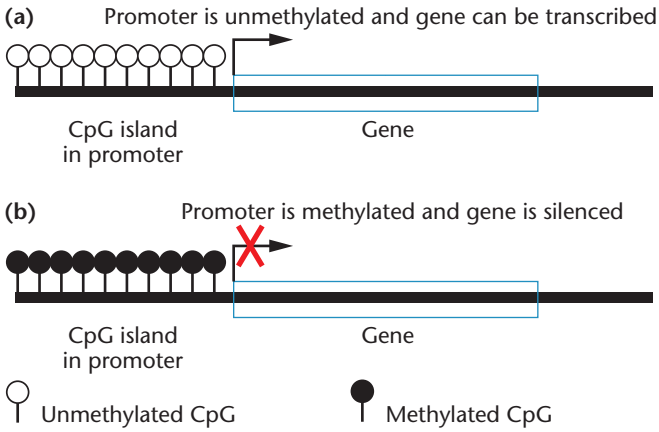
Epigenetic Alterations to the Genome

Unlike the genome, which is identical in all cell types of an organism, the epigenome is cell-type specific and changes throughout the life cycle in response to environmental cues. Like the genome, the epigenome can be transmitted to daughter cells by mitosis and to future generations by meiosis. In the following sections, we will examine mechanisms of epigenetic changes and their role in development, aging, cancer, and environment–genome interactions, providing a snapshot of the many roles played by this recently discovered mechanism of gene regulation.

There are three major epigenetic mechanisms: (1) reversible modification of DNA by the addition or removal of methyl groups; (2) remodeling of chromatin by the addition or removal of chemical groups to histone proteins; and (3) regulation of gene expression by small, noncoding RNA molecules.



ST FIGURE 1–1 The phenotype of an organism is the product of interactions between the genome and the epigenome. The genome is a constant from fertilization throughout life, but cells, tissues, and the organism develop different epigenomes as a result of epigenetic reprogramming of gene activity in response to environmental stimuli. These reprogramming events lead to phenotypic changes through the life cycle.



ST FIGURE 1–2 Methylation patterns of CpG dinucleotides in promoters control activity of the adjacent genes. CpG islands outside and within genes also have characteristic methylation patterns, contributing to the overall level of genome methylation.

Methylation

In mammals, DNA methylation takes place after replication and during cell differentiation. This process involves the addition of a methyl group ($-CH_3$) to cytosine, a reaction catalyzed by a family of enzymes called methyltransferases. Methylation takes place almost exclusively on cytosine bases located adjacent to a guanine base, a combination called a CpG dinucleotide. Many of these dinucleotides are clustered in regions called CpG islands, which are located in and near promoter sequences adjacent to genes (**ST Figure 1–2**). In euchromatin, CpG islands and promoters adjacent to essential genes (housekeeping genes) and cell-specific genes are unmethylated, making these genes available for transcription. In heterochromatic regions of the genome, genes with adjacent methylated CpG islands and methylated promoters are transcriptionally silenced. The methyl groups in CpG dinucleotides occupy the major groove of DNA and block the binding of transcription factors necessary to form transcription complexes.

During development, methylation of CpG islands is a normal process and plays a role in the inactivation of X chromosomes in females (see Chapter 5 for a detailed discussion of X inactivation). In addition, DNA methylation plays a role in parent-specific allele expression, a process called imprinting.

Histone Modification and Chromatin Configuration

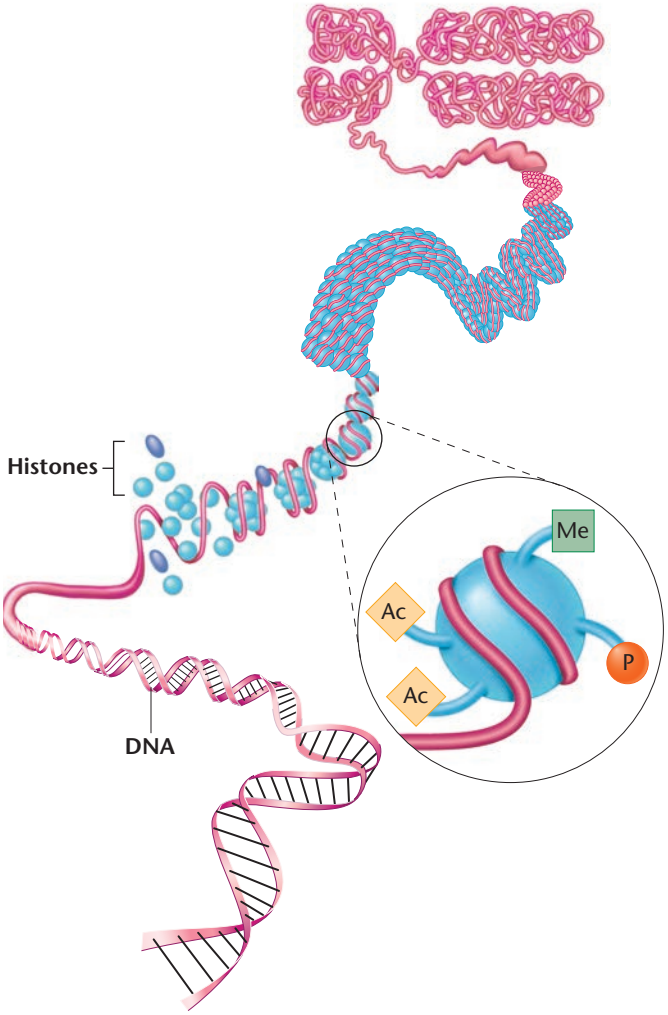
In addition to DNA methylation, chromatin remodeling is an important epigenetic mechanism of gene regulation. Recall that chromatin is a dynamic structure composed of

BOX 1

The Beginning of Epigenetics

C.H. Waddington coined the term *epigenetics* in the 1940s to describe how environmental influences on developmental events can affect the phenotype of the adult. He showed that environmental alterations during development induced alternative phenotypes in organisms with identical genotypes. Using *Drosophila melanogaster,* Waddington found that wing vein patterns could be altered by administering heat shocks during pupal development. Offspring of flies with these environmentally induced changes showed the alternative phenotype without the need for continued environmental stimulus. He called this phenomenon "genetic assimilation." In other words, interactions between the environment and the genome during certain stages of development produced heritable phenotypic changes.

In the 1970s, Holliday and Pugh proposed that changes in the program of gene expression during development depends on the methylation of specific bases in DNA, and that altering methylation patterns affects the resulting phenotype. Waddington's pioneering work, the methylation model of Holliday and Pugh, and the discovery that expression of genes from both the maternal and paternal genomes is required for normal development, all helped set the stage for the birth of epigenetics and epigenomics as fields of scientific research.



ST FIGURE 1–3 Clusters of histones in nucleosomes have their N-terminal tails covalently modified in epigenetic modifications that alter patterns of gene expression. Ac = acetyl groups, Me = methyl groups, P = phosphate groups.

DNA wound around a core of 8 histone proteins to form nucleosomes. Several sets of proteins are involved in modifying histones: "writers" and "erasers" that add or remove chemical groups, and "readers" that interpret these modifications. The N-terminal region of each histone molecule extends beyond the nucleosome, and the amino acids in these tails can be chemically modified by the addition of acetyl, methyl, and phosphate groups (**ST Figure 1–3**).

These modifications remodel chromatin structure and loosen the binding between histones and DNA, and shift the spacing of nucleosomes, making genes accessible for transcription [**ST Figure 1–4 (a)**]. Chromatin remodeling is a reversible process, and the removal of chemical groups changes chromatin from an "open" to a "closed" configuration and silences genes by making them unavailable for transcription [**ST Figure 1–4(b)**].

We are learning that specific combinations of histone modifications control the transcriptional status of a chromatin region. For example, histone methylation can either increase or decrease gene activity, depending on which amino acids are methylated and how many methyl groups are added. A large combination of histone modifications are possible, and the sum of the complex patterns and interactions of histone modifications that alter chromatin structure and gene expression is called the **histone code.** These changes allow differentiated cells to carry out cell-specific patterns of gene transcription and to respond to external signals that modify these patterns without any changes in DNA sequence.

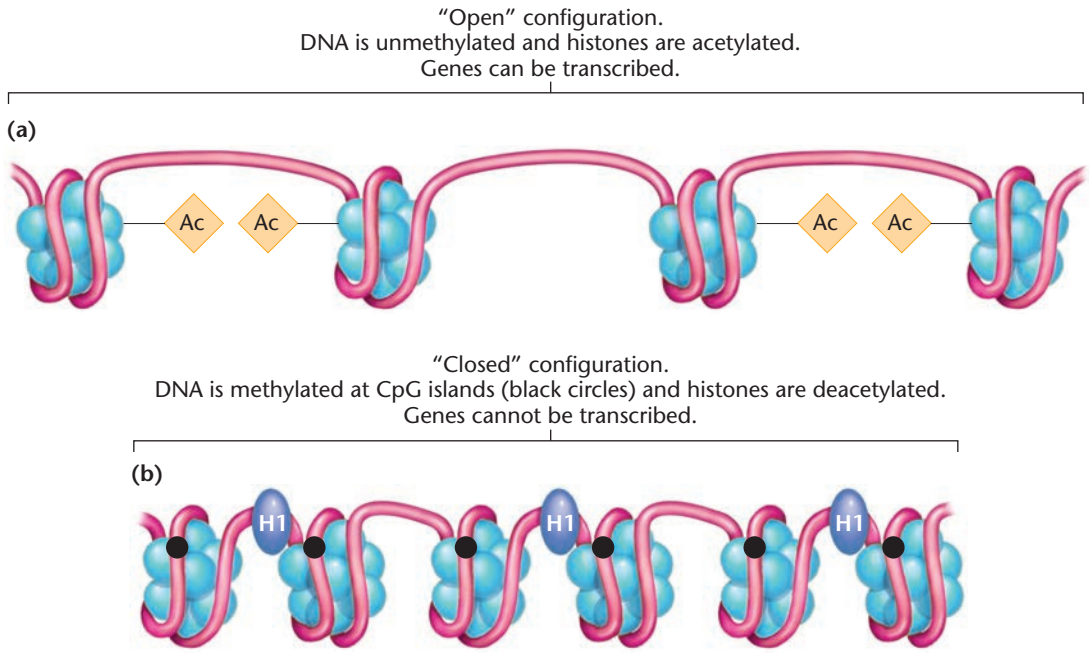
MicroRNAs and Long Noncoding RNAs

In addition to messenger RNA (mRNA), genome transcription produces several classes of noncoding RNAs. Two of these, microRNAs (miRNAs) and long, noncoding RNAs (lncRNAs), play important roles in epigenetic regulation of gene expression. The structure and function of these RNAs are discussed in Special Topic Chapter 2—Emerging Roles of RNA. miRNAs are involved in controlling pattern formation in developing embryos, the timing of developmental events, and physiological processes such as cell signaling. Recent evidence shows that miRNAs also play roles in the development of cardiovascular disease and cancer.

Primary miRNA transcripts are processed into precursor molecules about 70–100 nucleotides long, containing a double-stranded stem loop and single-stranded regions. Processing removes the single-stranded regions, and the loops move to the cytoplasm where they are altered further. The resulting double-stranded RNA is incorporated into a protein complex where one RNA strand is removed and degraded, forming a mature RNA-Induced Silencing Complex (RISC) containing the remaining single miRNA strand. RISCs act as posttranscriptional repressors of gene expression by binding to and destroying target mRNA molecules carrying sequences complementary to the RISC miRNA. mRNAs that are partially complementary to the RISC miRNA are modified, making these target mRNAs less likely to be translated by ribosomes, resulting in downregulation of gene expression.

In addition to forming RISC complexes, miRNA can also associate with a different set of proteins to form RNA-Induced Transcriptional Silencing (RITS) complexes. RITS complexes reversibly convert euchromatic chromosome regions into facultative heterochromatin, silencing the genes located within these newly created heterochromatic regions. Unlike the constitutive heterochromatin at telomeres and centromeres, facultative heterochromatin can be reversibly converted to euchromatin, making genes in this region once again accessible for transcription.

Long noncoding RNAs (lncRNAs) share properties in common with mRNAs; they often have 5′ caps, 3′ poly-A tails, and are spliced. What distinguishes lncRNAs from coding



ST FIGURE 1–4 Epigenetic modifications to the genome alter the spacing of nucleosomes and alter the availability of genes for transcription.

(mRNA) transcripts is the lack of an extended open reading frame that codes for this insertion of amino acids into a polypeptide. As epigenetic modulators, lncRNAs bind to chromatin-modifying enzymes and direct their activity to specific regions of the genome. At these sites, lncRNAs direct chromatin modification, altering the pattern of gene expression.

As information is accumulated about the molecular mechanisms associated with epigenetic regulation of gene expression, it has become clear that mutations in genes associated with these mechanisms are the basis for human genetic disorders. Dozens of such diseases have been identified, including Rett syndrome, a disorder of the nervous system, caused by mutations associated with DNA methylation. Weaver syndrome (a growth disorder) is associated with mutation of histone modification genes, and fragile-X syndrome (see Chapter 6) is associated with defects in miRNA processing.

In summary, epigenetic modifications alter chromatin structure by several mechanisms including DNA methylation, histone acetylation, histone deacetylation, and action of miRNAs, without changing the sequence of DNA. These epigenetic changes create an epigenome that, in turn, can regulate normal development or generate physiological responses to environmental signals.

Epigenetics and Development: Imprinting

Some epigenetic modifications are heritable and are transmitted to daughter cells at cell division. This ensures that during development, cell and tissue-specific patterns of gene transcription are initiated and maintained. In this way, developing nerve cells maintain their identity as they divide and form components of the central and peripheral nervous system.

The DNA carried by sperm and eggs are highly methylated. However, shortly after fertilization, most of the methylation marks associated with differentiated cells are erased. This resets embryonic cells to a pluripotent state, allowing them to undergo new epigenetic modifications to form the more than 200 cell types found in the adult body. About the time the embryo is implanting in the wall of the uterus, cells take on tissue-specific epigenetic identities, and methylation patterns and histone modifications change rapidly to reflect those seen in differentiated cells.

Some genomic regions, however, escape these rounds of global demethylation and remethylation. The genes contained in these regions remain imprinted with the methylation marks of the maternal and paternal chromosomes. This inherited pattern of methylation in CpG-rich regions and in promoter sequences produces allele-specific imprinting. The maternally and paternally inherited copies of imprinted genes remain transcriptionally silent during embryogenesis and later stages of development.

In humans, imprinted genes are usually found in clusters that can occupy more than 1,000 kb of DNA. Because these genes are located near each other at a limited number of sites in the genome, mutation in one imprinted gene can affect the function of adjacent imprinted genes, amplifying the mutation's phenotypic impact. Mutations in imprinted genes can arise by producing changes in the DNA sequence

or by causing dysfunctional epigenetic changes, called **epimutations**, both of which can cause heritable changes in gene activity.

Later in development, a second wave of epigenetic reprogramming occurs in the germ cells, which are located in the fetal gonads. In this wave, all epigenetic modifications, even those on imprinted genes, are completely removed from the germ-cell chromosomes. Once this process is accomplished, sex-appropriate epigenetic imprinting modifications are added to the germ-cell chromosomes. Germ cells in the developing ovary are programmed with female imprints, and in male germ cells, the chromosomes receive male imprints (ST Figure 1–5).

Most human disorders associated with imprinting originate during fetal growth and development. Imprinting defects cause Prader–Willi syndrome, Angelman syndrome, Beckwith–Wiedemann syndrome, and several other diseases (ST Table 1.1). However, given the number of candidate genes and the possibility that additional imprinted genes remain to be discovered, the overall number of imprinting-related genetic disorders may be much higher.

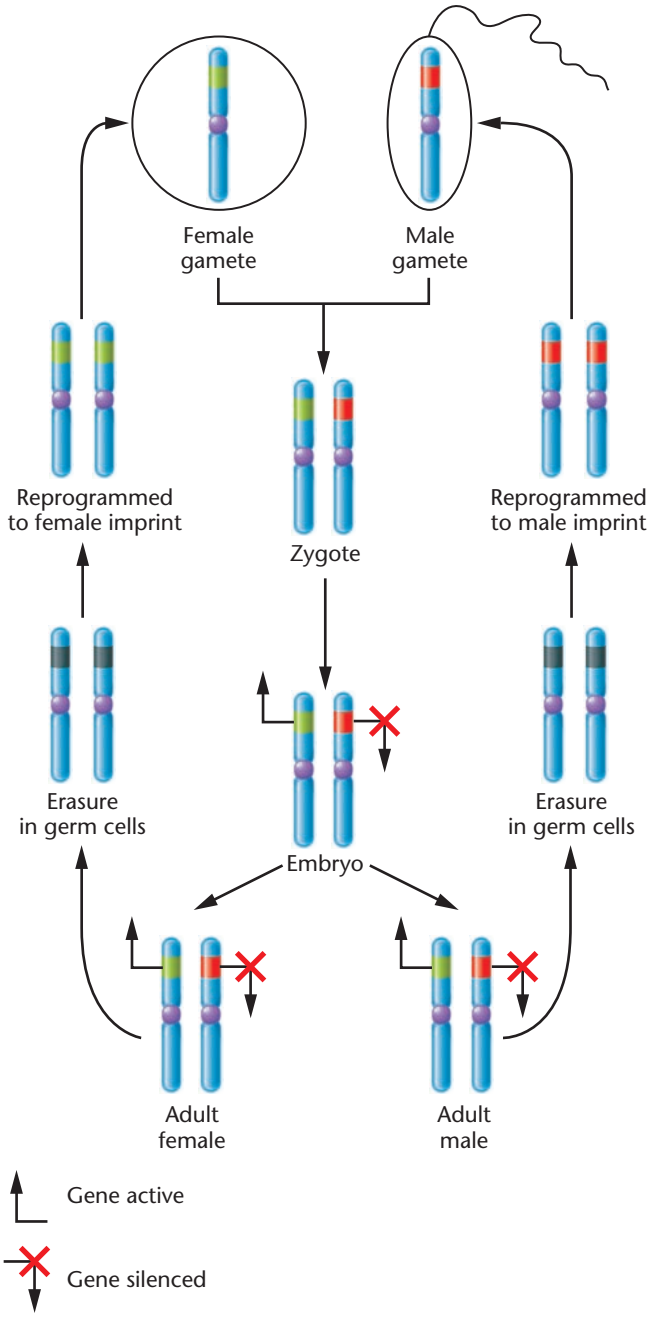
In humans, most known imprinted genes encode growth factors or other growth-regulating genes. An autosomal dominant disorder of imprinting, Beckwith–Wiedemann syndrome (BWS), offers insight into how disruptions of epigenetically imprinted genes lead to an abnormal phenotype. BWS is a prenatal overgrowth disorder with abdominal wall defects, enlarged organs, large birth weight, and predisposition to cancer. BWS is not caused by mutation, nor is it associated with any chromosomal aberration. Instead it is a disorder of imprinting and is caused by abnormal methylation patterns and resulting altered patterns of gene expression.

Genes associated with BWS are located in a cluster of imprinted genes on the short arm of chromosome 11. All genes in this cluster regulate growth during prenatal development. One of the genes in this cluster is *IGF2* (insulin growth factor 2). Normally, the paternal allele of *IGF2* is expressed, and the maternal allele is silenced.

In many individuals with BWS, the maternal *IGF2* allele is not silenced. As a result, both the maternal and paternal alleles are transcribed, resulting in the overgrowth of tissues that are characteristic of this disease.

The known number of imprinted genes represents only a small fraction (less than 1 percent) of the mammalian genome, but they play major roles in regulating growth during prenatal development. Because they act so early in life, external or internal factors that disturb the epigenetic pattern of imprinting or the expression of imprinted genes can have serious phenotypic consequences.

In the United States, assisted reproductive technologies (ART), including *in vitro* fertilization (IVF), are now



ST FIGURE 1–5 Imprinting patterns in germ cells are reprogrammed each generation to form sex-specific patterns of epigenetic modifications.

ST TABLE 1.1 Some Imprinting Disorders in Humans

Disorder	Locus
Albright hereditary osteodystrophy	20q13
Angelman syndrome	15q11-q15
Beckwith–Wiedemann syndrome	11p15
Prader–Willi syndrome	15q11-q15
Silver–Russell syndrome	Chromosome 7
Uniparental disomy 14	Chromosome 14

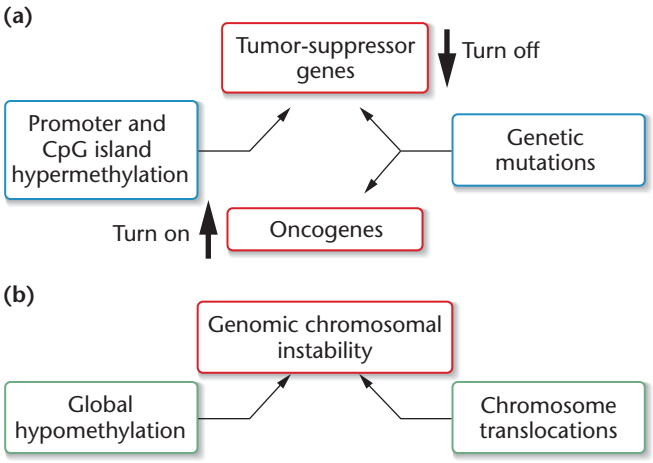
used in over 1 percent of all births. Over the past decade, several studies have suggested that children born through the use of ART have an increased risk for imprinting errors (epimutations) created by the manipulation of gametes or embryos.

For example, the use of ART results in a four- to nine-fold increased risk of BWS; in addition, there are increased risks for Prader–Willi syndrome and Angelman syndrome. Studies of children with Angelman syndrome or BWS conceived by IVF have shown that they have reduced levels or loss of maternal-specific methylation at known imprinting sites in the genome, confirming the role of epigenetics in these cases. Although imprinting errors are uncommon in the general population (BWS occurs in about 1 in 15,000 births), epimutations may be a significant risk factor for those conceived by ART. However, large-scale and longitudinal studies will be needed to assess the relationship among imprinting abnormalities, growth disorders, and ART.

Epigenetics and Cancer

Until recently, the conventional view has been that cancer is clonal in origin and begins in a single cell that has accumulated a suite of dominant and recessive mutations in genes that promote (proto-oncogenes) or inhibit (tumor-suppressor genes) cell division, allowing it to escape control of the cell cycle. Subsequent mutations allow cells of the tumor to become metastatic, spreading the cancer to other locations in the body where new malignant tumors appear.

As far back as the 1980s, however, researchers observed that cancer cells had much lower levels of methylation than normal cells derived from the same tissue. Subsequent research by many investigators showed that complex changes in DNA methylation patterns are associated with cancer. These studies showed that genomic hypomethylation is a property of all cancers examined to date. DNA hypomethylation reverses the transcriptional inactivation, leading to unrestricted transcription of many gene sets, including those associated with the development of cancer. It also relaxes control over imprinted genes, causing cells to acquire new growth properties. In addition, selective hypermethylation of CpG islands and subsequent gene silencing are also properties of cancer cells, indicating that aberrant patterns of DNA methylation are a universal property of cancer. As a result, cancer is now viewed as a disease that results from the accumulation of both epigenetic *and* genetic changes that lead to alterations in gene expression and the development of malignancy (**ST Figure 1–6**).



ST FIGURE 1–6 The development and maintenance of malignant growth in cancer involves gene mutations, hypomethylation, hypermethylation, overexpression of oncogenes, and the silencing of tumor-suppressor genes.

Hypomethylation of repetitive DNA sequences in heterochromatic regions is associated with an increase in chromosome rearrangements and changes in chromosome number, both of which are characteristic of cancer cells. In addition, hypomethylation of repetitive sequences leads to transcriptional activation of transposable DNA sequences such as LINEs and SINEs, further increasing genomic instability.

While widespread hypomethylation is a hallmark of cancer cells, hypermethylation at CpG islands and promoters silences certain genes, including tumor-suppressor genes (**ST Table 1.2**), often in a tumor-specific pattern. For example, *BRCA1* is hypermethylated and inactivated in breast and ovarian cancer, and *MLH1* is hypermethylated in some forms of colon cancer. Inactivation of tumor-suppressor genes by hypermethylation is thought to play

ST TABLE 1.2 Some Cancer-Related Genes Inactivated by Hypermethylation in Human Cancers

Gene	Locus	Function	Related Cancers
BRCA1	17q21	DNA repair	Breast, ovarian
APC	5q21	Nucleocytoplasmic signaling	Colorectal, duodenal
MLH1	3p21	DNA repair	Colon, stomach
RB1	13q14	Cell-cycle control point	Retinoblastoma, osteosarcoma
AR	Xq11-12	Nuclear receptor for androgen; transcriptional activator	Prostate
ESR1	6q25	Nuclear receptor for estrogen; transcriptional activator	Breast, colorectal

an important complementary role to mutational changes that accompany the transformation of normal cells into malignant cells. For example, in a bladder cancer cell line, one allele of the cell-cycle control gene *CDKN2A* is mutated, and the other, normal allele is inactivated by hypermethylation. Because both alleles are inactivated (although by different mechanisms), cells are able to escape control of the cell cycle and divide continuously. In many clinical cases, a combination of mutation and epigenetic hypermethylation occurs in familial forms of cancer.

However, genes other than those controlling the cell cycle are also hypermethylated in some cancers; these include genes that control or participate in DNA repair, differentiation, apoptosis, and drug resistance.

In addition to dysregulation of DNA methylation, many cancers also have altered patterns of chromatin conformation. Chromatin remodeling is controlled by the reversible covalent modification of histone proteins in nucleosome cores. This process involves three classes of enzymes: "writers" that add chemical groups to histones; "erasers" that remove these groups; and "readers" that recognize and interpret the chemical marks. Abnormal regulation of each of these enzyme classes results in disrupted histone profiles and is associated with a variety of cancer subtypes. Acetylation of histones is strongly correlated with the activation of gene transcription by reducing the strength of the interaction with DNA, making promoters available to RNA polymerase. Mutations in genes of the histone acetyltransferase (HAT) family and genes of the histone deacetylase (HDAC) family, which encode enzymes that remove acetyl groups and induce transcriptional repression, are linked to the development of cancer. For example, individuals with Rubinstein–Taybi syndrome inherit a germ-line mutation that produces a dysfunctional HAT and have a greater than 300-fold increased risk of cancer.

Abnormalities in histone deacetylation have been identified as a early stage in the transformation of normal cells into cancer cells. HDAC complexes are selectively recruited to tumor-suppressor genes by mutated, oncogenic DNA binding proteins. Action of the HDAC complexes at these genes converts the chromatin to a closed configuration and inhibits transcription, causing the cell to lose control of the cell cycle and undergo uncontrolled division.

Many of the mechanisms that cause epigenetic changes in cancer cells are not well understood, partly because they take place very early in the conversion of a normal cell to a cancerous one and partly because by the time the cancer is detected, alterations in the methylation pattern have already occurred. The fact that such changes occur very early in the transformation process has led to the proposal that epigenetic changes leading to cancer may occur within adult stem cells in normal tissue. Three lines of evidence support this idea: (1) epigenetic mechanisms can replace mutations as a way of silencing individual tumor-suppressor genes or activating oncogenes; (2) global hypomethylation may cause genomic instability and the large-scale chromosomal changes that are a characteristic feature of cancer; and (3) epigenetic modifications can silence multiple genes, making them more effective in transforming normal cells into malignant cells than sequential mutations of single genes.

In addition to changing ideas about the origins of cancer, the role of epigenetic mechanisms in cancer has opened the way to develop new classes of drugs for chemotherapy. The focus of epigenetic therapy is the reactivation of genes silenced by methylation or histone modification, essentially reprogramming the pattern of gene expression in cancer cells. Several epigenetic drugs have been approved by the U.S. Food and Drug Administration, and another 12 to 15 drugs are in clinical trials. One new drug, Vidaza, is used in the treatment of myelodysplastic syndrome, a precursor to leukemia, and for treatment of acute myeloid leukemia. This drug is an analog of cytidine and is incorporated into DNA during replication during the S phase of the cell cycle. Methylation enzymes (methyltransferases) bind irreversibly to decitabine, preventing methylation of DNA at other sites, effectively reducing the amount of methylation in cancer cells. Other drugs that inhibit histone deacetylases (HDAC) have been approved by the FDA for use in epigenetic cancer therapy. The HDAC inhibitor drug Zolinza is used to treat some forms of lymphoma. The inhibition of HDAC activity reactivates tumor-suppressor gene activity, bringing the tumor cells under cell-cycle control and halting uncontrolled cell division. Epigenetic cancer therapy drugs discovered to date are only moderately effective and are best used in combination with conventional chemotherapy drugs. To develop more effective epigenetic therapy, several basic questions must be answered (Box 2). Research into the mechanisms and genomic locations of epigenetic modifications in cancer cells will help promote the design of more potent drugs for epigenetic chemotherapy.

Epigenetics and the Environment

The epigenome receives and integrates intracellular, extracellular, and environmental signals with the information encoded in the genome to generate programs of gene expression in response to these signals. This means that organisms are able to adapt to and respond to internal and external stimuli throughout their life span. One of the most important sources of such signals is the environment. Environmental agents including nutrition, chemicals, medical or recreational drugs, as well as social interactions, stress, and exercise, exert effects on the epigenome.

BOX 2

What More We Need to Know about Epigenetics and Cancer

The discovery that epigenetic changes may be as important as genetic changes in the origin, maintenance, and metastasis of cancers has opened new avenues of cancer research. Key discoveries about epigenetic mechanisms include the finding of tumor-specific deregulation of genes by altered DNA methylation profiles and histone modifications, the discovery that epigenetic changes in histones or DNA methylation are interconnected, and the recognition that epigenetic changes can affect hundreds of genes in a single cancer cell. These advances were made in the span of a few years, and while it is clear that epigenetics plays a key role in cancer, many questions remain to be answered before we can draw conclusions about the relative contributions of genetics and epigenetics to the development of cancer. Some of these questions are as follows:

- Is global hypomethylation in cancer cells a cause or an effect of the malignant condition?
- Do these changes arise primarily in stem cells or in differentiated cells?
- Once methylation alterations begin, what triggers hypermethylation in cancer cells?
- Is hypermethylation a process that targets certain gene classes, or is it a random event?
- Can we develop drugs that target cancer cells and reverse tumor-specific epigenetic changes?
- Can we target specific genes for reactivation, while leaving others inactive?

In addition, physical factors, including seasons of the year, storms, and temperature, can also generate changes in gene expression mediated through the epigenome. Many of these epigenomic changes are heritable, and influence gene expression in future generations. In humans it is difficult to directly determine the relative contributions of environmental or learned behavior as factors in changing the epigenome, but there is evidence that environmental factors such as maternal nutrition and exposure to agents that affect the developing fetus can have detrimental effects during adulthood.

Women who were pregnant during the 1944–1945 famine in the Netherlands had children with increased risks for obesity, diabetes, and coronary heart disease. In addition, as adults, these individuals had significantly increased risks for schizophrenia and other neuropsychiatric disorders. Members of the F_2 generation also had abnormal patterns of weight gain and growth.

The most direct evidence for the role of environmental factors in modifying the epigenome comes from studies in experimental animals. A low-protein diet fed to pregnant rats results in permanent changes in the expression of several genes in both the F_1 and F_2 offspring. Increased expression of liver genes is associated with hypomethylation of their respective promoter regions. Other evidence indicates that epigenetic changes triggered by this diet modification were gene-specific.

A dramatic example of how epigenome modifications affect the phenotype comes from the study of coat color in mice, where color is controlled by the dominant allele *Agouti* (*A*). In homozygous *AA* mice, the gene is active only during a specific time during hair development, resulting in a yellow band on an otherwise black hair shaft, producing the agouti phenotype. A mutant allele (A^{vy}) causes yellow pigment formation along the entire hair shaft, producing a yellow phenotype. This allele is the result of the insertion of a transposable element near the transcription start site of the *Agouti* gene. A promoter element within the transposon is responsible for this change in gene expression. Researchers found that the degree of methylation in the transposon's promoter is related to the amount of yellow pigment deposited in the hair shaft and that the amount of methylation varies from individual to individual. The result is variation in coat-color phenotypes even in genetically identical mice (**ST Figure 1–7**). In these mice, coat colors range from yellow (unmethylated promoter) to pseudoagouti (highly methylated promoter). In addition to a gradation in coat color, there is also a gradation in body weight. Yellow mice are more obese than the brown, pseudoagouti mice.

Yellow | Slightly mottled | Mottled | Heavily mottled | Pseudo-agouti

ST FIGURE 1–7 Variable expression of yellow phenotype in mice caused by diet-related epigenetic changes in the genome.

To evaluate the role of environmental factors in modifying the epigenome, the diet of pregnant A^{vy} mice was supplemented with methylation precursors, including folic acid, vitamin B_{12}, and choline. In the offspring, coat-color variation was reduced and shifted toward the pseudoagouti (highly methylated) phenotype. The shift in coat color was accompanied by increased methylation of the transposon's promoter. These findings have applications to epigenetic diseases in humans. For example, the risk of colorectal cancer is linked directly to folate dietary deficiency and activity differences in enzymes leading to the synthesis of methyl donors.

In addition to foods that mediate epigenetic changes in gene expression via methylation, it has recently been reported that miRNAs from some plant foods such as rice and potatoes enter the blood stream of humans and regulate expression of target genes. Follow-up studies have not confirmed these findings, but further work should carefully explore the possibility that genetic information can be transferred by dietary intake.

Epigenetics and Behavior

A growing body of evidence shows that epigenetic changes, including alterations in DNA methylation and histone modification, have important effects on behavioral phenotypes. In mice, two regions of the brain show preferential expression of maternal or paternal alleles. Upward of 1000 genes in the developing brain are imprinted, supporting the idea that epigenetic mechanisms operating in different regions of the brain may represent a major form of behavioral regulation.

In humans, epigenetic changes have been documented during the progression of neurodegenerative disorders and in neuropsychiatric diseases, both of which show altered behavioral phenotypes. Epigenetic changes to the nervous system occur in Alzheimer disease, Parkinson disease, Huntington disease, and in schizophrenia and bipolar disorder. However, because the phenotypes in these disorders are influenced by a number of factors including genetic predispositions, events in prenatal development, and prenatal and postnatal environmental effects, it is not yet possible to define a cause and effect relationship between epigenomic changes and the onset and intensity of neural disorders.

One of the most significant findings in the epigenetics of behavior is that stress-induced epigenetic changes that occur prenatally or early in life can influence behavior (and physical health) later in adult life, and potentially be transmitted to future generations. For example, an early study showed that newborn rats that experienced reduced levels of maternal nurturing (low-MN) did not adapt well to stress and to anxiety-inducing situations in adulthood. In rats and humans, the hypothalamic region of the brain mediates stress reactions by controlling levels of glucocorticoid hormones via the action of cell-surface glucocorticoid receptors (GRs). In rats exposed to normal levels of nurturing care early in life (high-MN), GR expression is increased and adults are stress-adaptive. However, low-MN rats had reduced levels of GR transcription and were less able to adapt to stress. The relevant observation was that the differences in GR expression were associated with differences in DNA methylation and histone acetylation levels in the GR gene promoter. Low-MN rats had significantly higher levels of methylation than high-MN rats. Subsequent research showed that differences in DNA methylation are present in hundreds of genes across the genome, all of which show differential expression in low-MN and high-MN adults. Significantly, in low-MN adults, administering drugs that lower methylation levels reversed the effect of poor early-life nurturing and improved their stress responses. Later studies showed that these phenotypes can be transmitted across generations. Female rats raised by more nurturing mothers are more attentive to their own newborns, whereas those raised by less nurturing mothers are much less attentive and less nurturing to their offspring.

Similar epigenetic changes triggered by prenatal or early childhood environmental factors may alter later behavior in humans. For example, it is known that a history of child abuse increases the risk of suicide later in life. One study examined epigenomic differences in brain tissue from two classes of suicide victims and in others who died suddenly of unrelated causes. One class of suicide victims had experienced childhood abuse, and the other had no history of child abuse. Those who died suddenly of unrelated causes also had no history of child abuse. High levels of GR gene promoter methylation were found in suicide victims with a history of child abuse, but not in the other two groups. These results are consistent with those found in experimental animals and suggest that parental care, epigenomic variation, GR expression, and adult behavior are linked in both rats and humans. Further research may lead to the development of drugs to treat depression and help prevent suicide in humans.

As the role of the **epigenome** (the epigenetic state of a cell) in disease has become increasingly clear, researchers across the globe have formed multidisciplinary projects to map all the epigenetic changes that occur in the normal genome and to study the role of the epigenome in specific diseases.

These include:

- NIH Roadmap Epigenomics Project
- Human Epigenome Project (HEP)
- International Human Epigenome Consortium (IHEC)
- International Cancer Genome Consortium (ICGC)

In our conclusion, we will discuss some of these projects and their goals. The NIH Roadmap Epigenomics Project focuses on how epigenetic mechanisms controlling stem cell differentiation and organ formation generate biological responses to external and internal stimuli that result in disease. Part of this Project, the Human Epigenome Atlas collects and catalogs detailed information about epigenomic modifications at specific loci in different cell types, different physiological states, and different genotypes. These data allow researchers to perform integrative and comparative analysis of epigenomic data across genomic regions or entire genomes.

The Human Epigenome Project is a multinational, public/private consortium organized to identify, map, and establish the functional significance of all DNA methylation patterns in the human genome across all major tissue types in the body. Analysis of these methylation patterns may show that genetic responses to environmental cues mediated by epigenetic changes are a pathway to disease.

The International Human Epigenome Consortium (IHEC) is a global program established to determine how the epigenome has altered human populations in response to environmental factors. The consortium is cataloging the epigenomes of 1000 people from different populations around the world and will also include the epigenomes of 250 different cell types.

Although these projects are in the early stages of development, the information already available strongly suggests that we are on the threshold of a new era in genetics, one in which we can study the development of disease at the genomic level and understand the impact of environmental factors on gene expression. The results of these projects may help explain how environmental settings in early life can affect predisposition to adulthood diseases.

Visit the Study Area in MasteringGenetics for a list of further readings on this topic, including journal references and selected web sites.

Review Questions

1. What are the major mechanisms of epigenetic genome modification?
2. What parts of the genome are reversibly methylated? How does this affect gene expression?
3. What are the roles of proteins in histone modification?
4. Describe how reversible chemical changes to histones are linked to chromatin modification.
5. What is the histone code?
6. What is the difference between silencing genes by imprinting and silencing by epigenetic modifications?
7. Why are changes in nucleosome spacing important in changing gene expression?
8. How do microRNAs regulate epigenetic mechanisms during development?
9. What is the role of imprinting in human genetic disorders?

Discussion Questions

1. Imprinting disorders do not involve changes in DNA sequence, but only the methylated state of the DNA, or the modification of histones. Does it seem likely that imprinting disorders could be treated prenatally or prevented by controlling the maternal environment in some way, perhaps by dietary changes?
2. Should fertility clinics be required by law to disclose that some assisted reproductive technologies (ART) can result in epigenetic diseases? How would you and your partner balance the risks of ART with the desire to have a child?
3. How can the role of epigenetics in cancer be reconciled with the idea that cancer is caused by the accumulation of mutations in tumor-suppressor genes and proto-oncogenes?
4. If true, would the knowledge that plant miRNAs can affect gene expression in your body affect your food choices?

Emerging Roles of RNA

In 1958, Francis Crick proposed his theory for the central dogma of molecular biology, which described how genetic information flows from DNA to RNA to protein. Proteins, as the end products of the gene, were viewed as the functional units of the cell, while DNA was considered a stable molecule that stores genetic information. RNA was considered as the temporary message between DNA and protein. This insightful model was the basis for the emerging field of molecular biology and is still considered accurate today. However, this model did not anticipate many other types of **noncoding RNAs (ncRNAs)** that today are known to play important roles in genetic processes. Advancing rapidly in the early 1990s and continuing to this day, studies have revealed that ncRNAs are more abundant, more diverse, and more important than we could ever have imagined.

"Noncoding RNAs have myriad functions in the cell, including catalyzing reactions, modifying protein activity, defending the cell against foreign nucleic acids, and regulating gene expression."

While it has been known for well over a decade that only ~2 percent of 3.2 billion base pairs making up the human genome encode proteins, we have recently learned that more than 75 percent of the human genome is in fact transcribed into RNA. Among these many transcripts are numerous examples of ncRNAs that have experimentally determined biological roles. Earlier in the text, we introduced several examples. You may recall from studying DNA replication that RNA serves as a primer for DNA synthesis and as a template molecule for reverse transcription of telomere sequences by telomerase (Chapter 10). And **small nuclear RNAs (snRNAs)** bound by several proteins **(snRNPs)** catalyze the splicing of pre-mRNAs as they are converted to mature transcripts (Chapter 12). Still another example is the relatively recent finding that the 23*S* prokaryotic and 28*S* eukaryotic rRNAs within the ribosome are responsible for catalyzing peptide bond formation during translation (Chapter 13).

The key to RNA's versatility lies in its structural diversity. While RNA is transcribed as a single strand that can serve as an informational molecule such as mRNA, it can also form complementary bonds with other nucleic acids, creating DNA/RNA duplexes and RNA/RNA duplexes (double-stranded RNA), or it can fold back on itself to form hairpin or stem-loop structures. As with proteins, such three-dimensional folding imparts enormous structural and thus functional diversity to RNA molecules. Additionally, RNAs can associate with proteins and modify their activity or recruit them to other nucleic acids through complementary base pairing.

As our knowledge of the diversity of RNA structure and function has increased, it has shifted several paradigms for how we think about life at the molecular level. Whereas we once thought only proteins could be enzymes, we are now searching for ways to use catalytic RNAs as therapeutics to fight disease. Whereas we once thought that most of the human genome was "junk DNA," we are now discovering that most noncoding sequences can be transcribed and even display important anticancer functions. Whereas we never imagined that RNA could regulate gene expression, we now know that it does so at the level of both transcription and translation. As you will soon learn, RNA discoveries have even changed our views on how life began. In this Special Topics chapter we will examine the versatility of RNA and highlight some modern applications of RNAs as research tools and therapeutic agents.

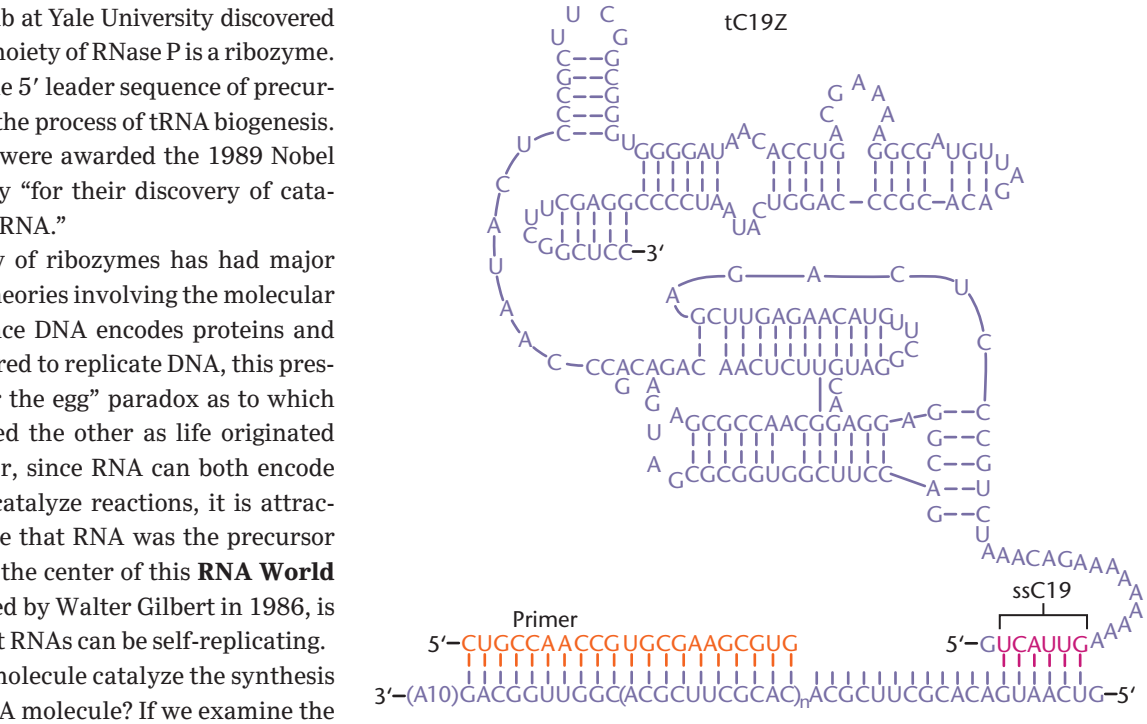
Catalytic Activity of RNAs: Ribozymes and the Origin of Life

In the 1960s, several scientists including Carl Woese, Francis Crick, and Leslie Orgel postulated that RNAs could be catalysts based on the fact that RNAs can form complex secondary structures. However, the first examples of RNAs acting as biological catalysts, or **ribozymes,** came many years later in the early 1980s from two different studies. Tom Cech's lab at the University of Colorado at Boulder discovered that introns (Group I introns) in some RNAs could be spliced out in the complete absence of any proteins. Such self-splicing introns demonstrated that RNAs could break and form phosphodiester bonds (see Chapter 12).

Sidney Altman's lab at Yale University discovered that the M1 RNA moiety of RNase P is a ribozyme. RNase P cleaves the 5′ leader sequence of precursor tRNAs during the process of tRNA biogenesis. Cech and Altman were awarded the 1989 Nobel Prize in Chemistry "for their discovery of catalytic properties of RNA."

The discovery of ribozymes has had major implications for theories involving the molecular origins of life. Since DNA encodes proteins and proteins are required to replicate DNA, this presents a "chicken or the egg" paradox as to which may have preceded the other as life originated on Earth. However, since RNA can both encode information and catalyze reactions, it is attractive to hypothesize that RNA was the precursor to cellular life. At the center of this **RNA World Hypothesis,** coined by Walter Gilbert in 1986, is the possibility that RNAs can be self-replicating.

Can an RNA molecule catalyze the synthesis of an identical RNA molecule? If we examine the naturally occurring ribozymes, the repertoire of ribozyme-catalyzed reactions is fairly limited. Most ribozymes catalyze the cleavage or formation of phosphodiester bonds. One notable exception is the ribosome, as discussed earlier; the 23*S* prokaryotic and 28*S* eukaryotic large subunit rRNAs are ribozymes that catalyze peptide bond formation. However, naturally occurring ribozymes are limited in their catalytic activity in that they generally are only able to catalyze a reaction once. For example, self-splicing introns cut themselves out and ligate the exons together but can neither repeat this reaction on other molecules nor reverse it. Thus far, there are only three examples of naturally occurring ribozymes that are capable of "multiple turnover": the ribosome, snRNPs, and RNase P. So the naturally occurring ribozymes fall far short of the self-replicating molecules postulated in the RNA World Hypothesis. Nonetheless, one may predict that self-replicating RNAs did once exist but that during evolution, they were usurped by the more stable informational storage of DNA and the more efficient catalysis of enzymes.



ST FIGURE 2–1 A ribozyme polymerase. The lab-synthesized tC19Z ribozyme is capable of copying RNA templates up to 95 nucleotides long, with a sequence complementary to the ssC19 5′-end of the ribozyme.

Genetic Engineering of Ribozymes

To better understand the potential of ribozymes, scientists have turned toward genetically engineering them. By starting with a large collection of lab-synthesized RNA molecules of varying sequences, scientists have been able to identify and isolate molecules with specific catalytic activity. Using this process, dubbed ***in vitro* evolution,** lab-synthesized ribozymes with novel functions have been created. These functions include RNA and DNA phosphorylation, carbon–carbon bond formation, RNA aminoacylation, and other reactions. Importantly, ribozymes with RNA polymerase activity have also been isolated. Recently, the Holliger lab at the University of Cambridge reported a ribozyme polymerase called tC19Z, which can reliably copy RNA molecules up to 95 nucleotides long (**ST Figure 2–1**). One template RNA that tC19Z copied successfully was that of another ribozyme, which was subsequently shown to be functional. Such a finding lends further credibility to the RNA World Hypothesis, but it remains an open question as to how complex molecules such as the ribonucleotide building blocks of RNAs could be synthesized in a prebiotic world.

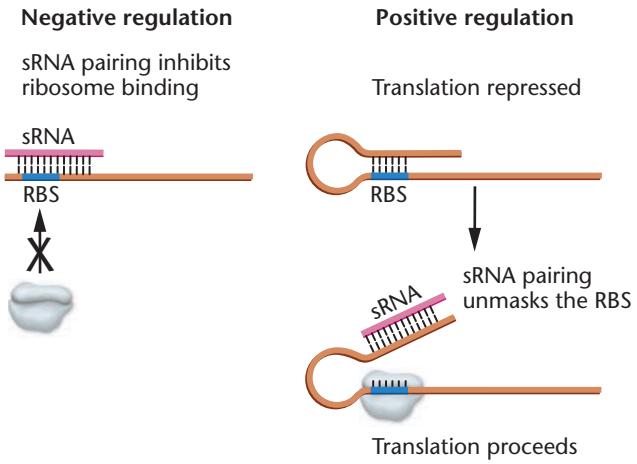
Genetically engineered ribozymes are also currently being investigated for therapeutic use. Angiozyme, designed to block angiogenesis, or the growth of new blood vessels, was the first ribozyme to be tested in clinical trials. Cancer cells often secrete a protein called vascular endothelial growth factor (VEGF), which signals through VEGF receptors on cells of blood vessels to induce vascularization of tumors. This provides the cancer cells a supply of oxygen and nutrients as well as a route for spreading to other parts of the body. To battle this problem, Angiozyme targets VEGF receptor mRNAs for destruction to prevent formation of the receptor proteins, thus stopping angiogenesis. Angiozyme showed promising results in animal tests and progressed to phase II clinical trials on patients with metastatic breast cancer. Although analysis of patient blood

serum showed a significant reduction in VEGF receptor levels, Angiozyme was not effective at fighting cancer and further development of the drug was halted. This study suggests that there is potential for ribozymes as drugs because there was specific target inhibition, but the study also demonstrates the complex nature of fighting cancer. There are ongoing clinical trials for ribozymes engineered for multiple targets associated with fighting HIV infection.

Small Noncoding RNAs Play Regulatory Roles in Prokaryotes

Prokaryotic small noncoding RNAs (**sRNAs**) were discovered decades ago, but their regulatory functions are still being elucidated and new sRNAs are still being discovered. It is thought that *E. coli* contains roughly 80–100 sRNAs, and other species are reported to have three times that number. sRNAs are generally between 50 and 500 nucleotides long and are involved in gene regulation and the modification of protein function. sRNAs involved in gene regulation are often transcribed from loci that partially overlap the coding genes that they regulate. However, they are transcribed from the opposite strand of DNA and in the opposite direction, making them complementary to mRNAs transcribed from that locus. In other cases, sRNAs are complementary to target mRNAs, but are transcribed from loci that do not overlap target genes. sRNAs regulate gene expression by binding to mRNAs (usually at the 5′ end) that are being transcribed. In some cases, sRNA binding to mRNAs blocks translation of the mRNA by blocking the ribosome binding site (RBS). In other cases, binding enhances translation by preventing secondary structures from forming in the mRNA that would block translation, often by masking the RBS (see **ST Figure 2–2**). Thus, sRNAs can be both negative and positive regulators of gene expression.

sRNAs have been shown to play important roles in gene regulation in response to changing environmental conditions or stress. For example, the sRNA *DsrA* of *E. coli* is upregulated in response to low temperature and promotes the expression of genes that enable the long-term survival of the cell under stressful conditions, or stationary phase. *DsrA* binds to *rpoS* mRNA to promote the translation of the RpoS stress response sigma factor (see Chapter 12), which is the primary transcriptional regulator of genes that promote stationary phase. In contrast, *RyhB* sRNA from *E. coli* is a negative regulator of gene expression. In response to low iron levels, *RyhB* is transcribed to inhibit the translation of several nonessential iron-containing enzymes so that the more critical iron-containing enzymes can utilize what little iron is present in the cytoplasm.

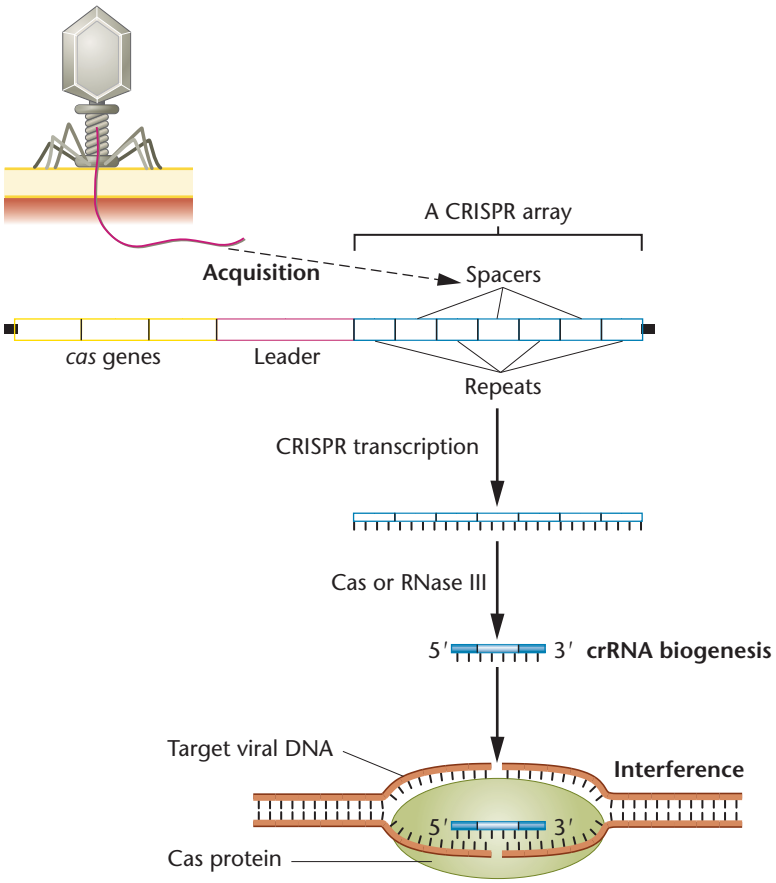


ST FIGURE 2–2 Bacterial small noncoding RNAs regulate gene expression. Bacterial sRNAs can be negative regulators of gene expression by binding to mRNAs and preventing translation by masking the ribosome binding site (RBS), or they can be positive regulators of gene expression by binding to mRNAs and preventing secondary structures (that would otherwise mask an RBS) and enable translation.

In addition to regulating mRNA translation, sRNAs can modulate protein activity. The *E. coli* 6*S* RNA is a 184-nucleotide sRNA that accumulates during stationary phase and functions to repress the transcription of genes that mediate vegetative growth. It accomplishes this by binding to RNA polymerases associated with the RpoD sigma factor that promotes the transcription of genes required for vegetative growth. 6*S* RNA has a hairpin structure that binds to the active site of the RNA polymerase and mimics double-stranded DNA, thus blocking the RNA polymerase from binding to promoter regions of target genes.

Prokaryotes Have an RNA-Guided Viral Defense Mechanism

Biological warfare between viruses (bacteriophages) and prokaryotes as a result of their coevolution has led to a diversity of defense mechanisms. For example, bacteria express endonucleases (restriction enzymes), which cleave specific DNA sequences. Such restriction enzymes destroy foreign bacteriophage DNA, while the host's genomic DNA is protected by DNA methylation. These same restriction enzymes have been adopted by molecular biologists for use in recombinant DNA technology (see Chapter 17). Bacteria can also defend against bacteriophage attack by blocking bacteriophage adsorption, blocking DNA insertion, and inducing suicide in infected cells. All of these defense mechanisms are **innate** because they are not tailored to a specific pathogen. In contrast, recent research has identified



ST FIGURE 2–3 The CRISPR/Cas system mediates an adaptive immune response in prokaryotes. CRISPR loci contain repeat sequences derived from foreign viral DNA separated by spacer sequences. *cas* genes, located nearby in the genome, are involved in CRISPR-mediated adaptive immunity. The CRISPR/Cas mechanism for adaptive immunity has three steps of acquisition (viral DNA is inserted into CRISPR loci), crRNA biogenesis (CRISPR loci are transcribed and processed into short crRNAs), and interference (crRNAs direct Cas proteins to viral DNA to destroy it).

an **adaptive** viral defense strategy whereby previous infection by a pathogen provides immunity to that cell and its descendants. In 2007, researchers at a Danish food science company called Danisco sought to create a strain of *Streptococcus thermophilus* that was more resistant to bacteriophage attack, thus making it more efficient for use in the production of yogurt and cheese. The Danisco researchers found that when they exposed *S. thermophilus* to a particular bacteriophage, it essentially became immune to it. The Danisco research group and others identified an adaptive defense mechanism that uses RNA as a guide for destroying viral DNA.

The adaptive viral defense mechanism is dependent on a genomic feature called a **CRISPR,** so named because it contains clustered regularly interspaced short palindromic repeats (**ST Figure 2–3**). CRISPR sequences were first identified in 1987 in the *E. coli* genome based on a simple description of repeated DNA sequences with unique spacer sequences between them. These spacers remained a mystery until 2005 when three independent studies demonstrated that the spacer sequences of CRISPRs were identical to bacteriophage sequences. We now know that insertion of fragments of phage DNA into CRISPR loci is required for adaptive immunity. CRISPR-associated genes (***cas***) are found adjacent to CRISPR loci and are required for various aspects of adaptive immunity (ST Figure 2–3). There are many different types of Cas proteins, and they include proteins with DNase, RNase, and helicase domains, as well as proteins of unknown function.

The **CRISPR/Cas** mechanism for adaptive immunity has three discrete steps (ST Figure 2–3). The first step is the integration of invading phage DNA into CRISPR loci and is often referred to as **acquisition.** It is not completely clear how the host cell fragments phage DNA and inserts it into these specific CRISPR loci. However, one of the Cas proteins, Cas1 from *E. coli*, is a double-stranded DNase and is thought to be involved in this process. Certain sequences in phage genomes also appear to be targeted for CRISPR integration.

After acquisition, the second step is the transcription of CRISPR loci and the processing of CRISPR-derived RNAs (crRNAs). This step is referred to as **crRNA biogenesis,** which also requires assistance from Cas proteins. CasE from *E. coli* is an endoribonuclease that cleaves long precursor crRNAs (pre-crRNAs) into mature crRNAs. Each mature crRNA contains a spacer sequence (determined by the integrated phage DNA fragment) flanked by repeat sequences. In some species, CRISPR repeat sequences form secondary structures such as stem-loops that are required for processing. In other species, a short RNA called a transactivating CRISPR RNA (tracrRNA) is also required for processing. tracrRNAs are complementary to CRISPR repeat sequences in the pre-crRNA and bind to them to form secondary structures.

The third and final step is the targeting and cleavage of phage DNA sequences complementary to crRNAs. This step is referred to as **interference.** Mature crRNAs associate with one or more Cas proteins to form CRISPR-associated ribonucleoprotein complexes, which bind to foreign DNA through crRNA-DNA complementary interactions. When a complementary crRNA-DNA match is made, the complex catalyzes cleavage of both strands of the foreign DNA within the region of complementarity. Cas3, found in many species, has both helicase and DNase domains and is important for target DNA cleavage. Amazingly, this system is able to distinguish "self" DNA from foreign DNA; otherwise

the CRISPR/Cas system would cleave CRISPR spacers in the genome since they contain bacteriophage sequences. "Self" identification is possible because the crRNAs contain CRISPR repeat sequences that are complementary to the "self" DNA and inhibit cleavage. Complementarity outside of the spacer region prevents DNA cleavage presumably because of conformational changes to the CRISPR-associated ribonucleoprotein complexes.

Although the details of the mechanism have been studied in just a select few species, it is already apparent that the CRISPR/Cas system is variable with respect to the number of CRISPR loci, the number of repeats, the length of the repeats, the length of the spacers, and the *cas* genes. It is also apparent, however, that some variation of the CRISPR/Cas adaptive defense system is present in roughly 48 percent of sequenced bacterial species and roughly 95 percent of sequenced Archaea, suggesting that it is an ancient and successful defense strategy. Furthermore, evidence suggests that this defense strategy is not just active against bacteriophages; this system is active against plasmid DNA as well.

CRISPR/Cas has been harnessed as an important tool for molecular biology. By expressing components of the CRISPR/Cas system in eukaryotic cells, scientists are able to specifically and efficiently target genes for mutation. The CRISPR/Cas system of *Streptococcus pyrogenes* has been used most extensively for this purpose because only three things are needed to target a specific DNA sequence: the mature crRNA, the Cas9 endonuclease, and a tracrRNA (see above) required for crRNA interaction with Cas9. However, the crRNA and tracrRNA can be genetically engineered and expressed as a hybrid small guide RNA (sgRNA), which provides the requisite secondary structure needed for activity. Thus, by inserting Cas9 and sgRNAs, one can target a specific gene for interruption. This can be done by inserting the genes encoding for Cas9 and the sgRNA or by injecting *in vitro* transcribed sgRNA and mRNA for Cas9. Since this technique was published in 2012, it has been successfully used to target genes for mutation in *Drosophila, C. elegans*, zebrafish, *Arabidopsis*, mice, and cultured human cells, demonstrating the broad applicability and the fast pace with which this is being adopted as a tool for science.

CRISPR/Cas technology also has potential for specific genome editing in human embryos. In fact, this has already been achieved. A team of researchers in China led by Dr. Junjiu Huang used CRISPR/Cas technology to create specific edits in the genomes of human embryos. However, these embryos were abnormal triploids formed *in vitro* when two sperm fertilized the same oocyte. Such triploid embryos are inviable early in development. While this study showed that targeted genome editing in human embryos is indeed possible with CRISPR/Cas, the study also demonstrated that the specific edits occurred at a low efficiency and that the genomes also suffered non-specific, off-target alterations. The fast pace of CRISPR/Cas genome editing technology has led many researchers to call for a moratorium on use of CRISPR/Cas genome editing of human DNA in any scenarios in which the targeted changes would be inherited by the next generation. This, however, does leave open the possibility of using CRISPR/Cas for gene therapy on somatic cells that will not be inherited (see Box 1).

Small Noncoding RNAs Mediate the Regulation of Eukaryotic Gene Expression

As discussed above, small noncoding RNAs (sRNAs) regulate gene expression and modulate protein activity in prokaryotes. Indeed, small noncoding RNAs are present in eukaryotes as well. Despite the fact that they are both called small noncoding RNAs, the prokaryotic and eukaryotic varieties differ in terms of their length, biosynthesis, mechanisms of action, and repertoire of regulatory activities. To help keep this distinction clear, scientists have applied the abbreviation sRNA to refer to the prokaryotic variety and refer to **small noncoding RNAs** in eukaryotes as **sncRNAs.** sncRNAs are short (20–31 nucleotides long) double-stranded RNAs with a 2-nucleotide 3′ overhang and are involved in silencing gene expression at the transcriptional or posttranscriptional levels through a process called **RNA-induced gene silencing.**

There are three different types of sncRNAs, which are classified according to their biogenesis and function. Small interfering RNAs (**siRNAs**) protect the cell from exogenous RNAs, such as from retroviral genomes. However, the exploitation of the siRNA mechanism for research and application has already proved indispensable and still has enormous potential. MicroRNAs (**miRNAs**) play important roles in regulating gene expression, and their malfunction is associated with human diseases (see Chapter 15 for additional coverage of siRNAs and miRNAs). The recently identified and poorly understood Piwi-interacting RNAs (**piRNAs**) protect the germ cells from the harmful effects of mobile DNA elements. We will consider siRNAs and miRNAs in more depth below.

siRNAs and RNA Interference

Nobel Prize-winning research from Andrew Fire and Craig Mello in 1998 demonstrated that when double-stranded RNA (dsRNA) is introduced into the cells of *C. elegans*, it leads to a "potent and specific" degradation of complementary mRNAs through a process we call **RNA interference** (**RNAi**). We now know a great deal about the molecular basis of RNAi, that it functions broadly in eukaryotes, and how to harness it as a tool for research and as a therapeutic

BOX 1

RNA-Guided Gene Therapy with CRISPR/Cas Technology

The CRISPR/Cas system has proved to be an efficient and specific genome-editing tool that has been exploited by scientists to induce mutations in target genes of interest. Based on this success, a company called Editas Medicine, based out of Cambridge, Massachusetts, aims "to translate its genome editing technology into a novel class of human therapeutics that enable precise and corrective molecular modification to treat the underlying cause of a broad range of diseases at the genetic level." Simply put, Editas Medicine hopes to develop CRISPR/Cas technology for gene therapy.

Founded by five academic scientists with expertise in CRISPR/Cas technology and supported by $43 million in venture capital investment, Editas Medicine thinks that CRISPR/Cas technology may offer a solution to some of the problems of traditional gene therapy strategies. Although we do not know their precise strategies or their intended target diseases, one can speculate on how CRISPR/Cas can be used for gene therapy. Since the CRISPR/Cas system uses RNA as a guide to direct the Cas9 endonuclease to complementary DNA sequences in the genome, small guide RNAs (sgRNAs) may be engineered to target disease-causing mutant alleles. For recessive genetic disorders, CRISPR/Cas can be used to induce a double-stranded break in the target gene, which activates the DNA repair machinery. If a normal copy of the gene is inserted into cells along with the CRISPR/Cas components, it can be used as a template for correcting the mutant allele by homologous recombination (see Chapter 14). This has a distinct advantage over traditional gene therapy methods because the gene is corrected in its native location in the genome rather than being randomly inserted or not integrated into a chromosome at all.

Although traditional gene therapy methods are not useful for treating dominant genetic diseases, CRISPR/Cas may be effective and thus extend the reach of gene therapy to a greater range of diseases. Most of the individuals afflicted with dominant disorders are heterozygous; they have a dominant mutant allele and a normal allele. In theory, CRISPR/Cas may be used to specifically inactivate the dominant mutant allele, rendering it a recessive, loss-of-function allele. In many cases, the undisrupted allele would then be able to restore normal gene function.

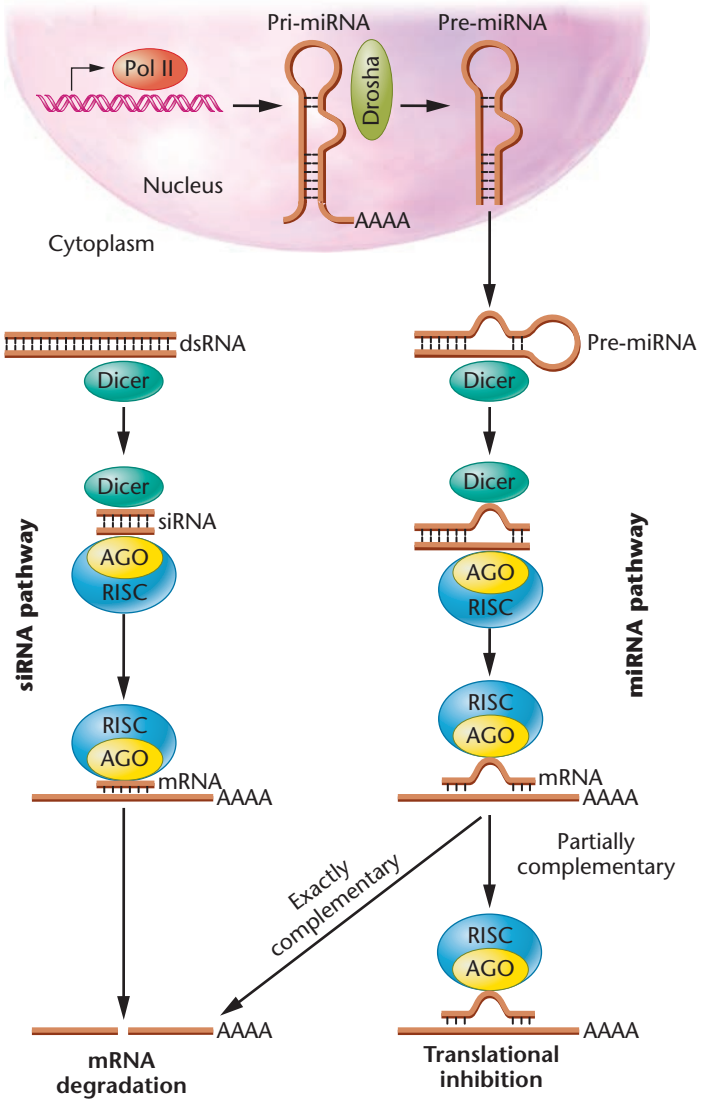
Despite encouraging foundational studies on CRISPR/Cas genome editing, bright minds working on the problem, and financial backing, there are some important hurdles to overcome before Editas Medicine can achieve its goals. For example, how will sgRNAs and Cas9 (or genes encoding these components) be delivered to the appropriate target cells in the patient effectively and without causing damaging side-effects? Delivery is a difficult problem for all gene therapy strategies. However, a concern specific to CRISPR/Cas is that some studies found that the Cas9 endonuclease cleaves DNA at off-target loci in the genome. It will be important to achieve absolute target specificity before this strategy will be safe for therapeutic applications.

agent. When double-stranded RNAs (dsRNAs) are present in the cytoplasm of a eukaryotic cell, they are recognized and cleaved into ~22-nucleotide-long RNAs with a 2-nucleotide 3′ overhang by an RNAse III protein called **Dicer.** These **siRNAs** then associate with the **RNA-induced silencing complex** (**RISC**), which contains an **Argonaute** family protein that binds RNA and has endonuclease or "slicer" activity. RISC cleaves and evicts one of the two strands of the siRNA and retains the other strand as a siRNA "guide" to recruit RISC to a complementary mRNA. RISC then cleaves the mRNA in the middle of the region of siRNA/mRNA complementarity (see **ST Figure 2–4**). Cleaved mRNA fragments lacking a methylated cap or a poly-A tail are then quickly degraded in the cell by nucleases.

The discovery of RNAi in 1998 led to an almost immediate revolution in the investigation of gene function. As long as a gene's sequence is known, one can quickly synthesize dsRNA corresponding to that sequence to inhibit or "knock down" that gene's function. As an example of how quickly RNAi became a tool for research, five years after the discovery of RNAi, the Ahringer lab from the University of Cambridge used RNAi on a massive scale to determine the loss-of-function phenotypes for 86 percent of the genes in the *C. elegans* genome. Subsequently, many genome-wide RNAi screens have been performed to test genes, one at a time, for roles in molecular, cellular, and developmental biology pathways.

siRNAs are also important tools for biomedical research and have great potential as therapeutic agents (see Chapter 15). Thus far, there have been over 30 clinical trials for siRNA drugs to treat diseases, such as asthma, cancer, and vision impairment. Despite substantial evidence that siRNAs are effective against a broad range of gene targets, siRNA delivery to target cells remains difficult. siRNAs are rapidly degraded when injected into the bloodstream, and they cannot passively cross the plasma membrane. Due to these siRNA delivery problems, several pharmaceutical companies, such as Novartis, Pfizer, and Roche, abandoned research and clinical trials involving siRNA drugs.

Since 2011, there has been a recent resurgence of siRNA drug development due to reports of successful delivery of siRNAs using either nanoparticles or cell-penetrating peptides. Clearly, an effective delivery solution will need to be



ST FIGURE 2–4 RNA interference pathways. Double-stranded RNA is processed into short interfering RNAs (siRNAs) by Dicer. siRNAs then associate with the RNA induced silencing complex (RISC) containing an Argonaute (AGO) family protein. RISC unwinds the siRNAs into single-stranded siRNAs and cleaves mRNAs complementary to the siRNA. MicroRNA (miRNA) genes are transcribed as primary-miRNAs (pri-miRNAs), which are trimmed at the 5′ and 3′ ends by Drosha to form pre-miRNAs. Pre-miRNAs are exported to the cytoplasm and processed by Dicer. These miRNAs then associate with RISC, and single-stranded miRNAs target RISC to mRNAs. If the miRNA and mRNA are perfectly complementary, the mRNA is destroyed; however, if there is a partial match, translation is inhibited.

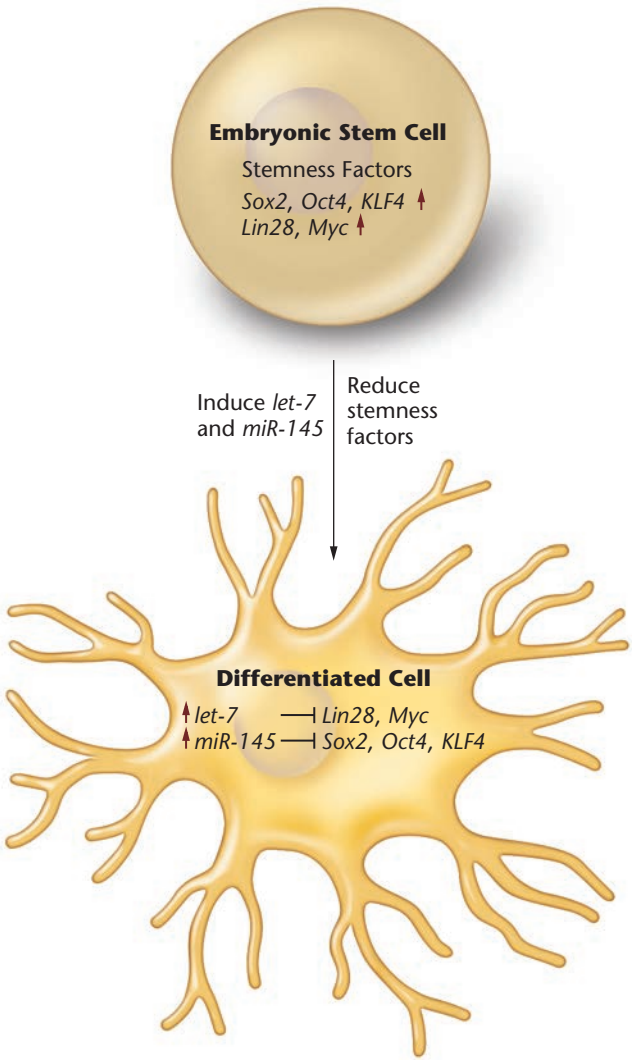
identified before siRNAs will gain widespread use as therapeutic agents. As a promising start, Alnylam Pharmaceuticals recently reported successful phase II clinical trials with Patisiran, an siRNA drug and nanoparticle delivery system that treats familial amyloidotic polyneuropathy (FAP). FAP is a disorder characterized by nervous system and cardiac problems due to aggregation of a mutant form of the transthyretin (TTR) protein. Importantly, Patisiran treatment reduced TTR protein levels in blood serum by 80% and halted progression of nervous system problems in FAP patients.

Although the initial studies of RNAi involved the introduction of lab-synthesized dsRNAs, naturally occurring dsRNAs have been described as well, such as bi-directional transcription of genomic loci and retroviral genomes. siRNAs produced without insertion of lab-synthesized RNAs are called **endogenous siRNAs** (**endo-siRNAs**). In fact, it is thought that the RNAi pathway evolved, in part, as a defense strategy against retroviruses, which often have dsRNA genomes. In that model, retroviral dsRNA would be chopped into siRNAs, which then guide RISC to retroviral transcripts to destroy them.

miRNAs Regulate Posttranscriptional Gene Expression

Another type of sncRNA present in eukaryotes is the miRNA (see Chapter 15). Unlike siRNAs, miRNAs are derived from the self-complementary transcripts of miRNA genes. These initial transcripts called **primary miRNAs** (**pri-miRNAs**) are "capped and tailed," and some contain introns like messenger RNAs. However, due to their self-complementary sequences, pri-miRNAs form hairpin structures. A nuclear enzyme called **Drosha** removes the non–self-complementary 5′ and 3′ ends to produce **pre-miRNAs.** These hairpins, which are single-stranded, are then exported to the cytoplasm and cleaved by **Dicer** to produce mature double-stranded miRNAs. miRNAs are very similar to siRNAs and associate with RISC to target complementary mRNAs. If the miRNA/mRNA match is perfect (common in plants), the target mRNA is cleaved by RISC, but if the miRNA/mRNA match is partial (common in animals) it blocks translation by the ribosome (ST Figure 2–4).

miRNAs are found in animals, plants, viruses, and possibly fungi, and have been shown to regulate genes involved in diverse cellular processes such as stress responses in plants, development in *C. elegans*, and cell-cycle control in mammalian cells. Why is miRNA-mediated posttranscriptional regulation so widespread? Wouldn't it be more efficient for the cell to repress a gene's transcription rather than transcribe a second gene to destroy/inhibit the mRNA of the first gene? The answer to these questions partly comes from the fact that mRNAs may be translated many times after transcription is stopped. To achieve an efficient and rapid change in gene expression, a cell can turn off transcription and employ an miRNA to target the existing mRNAs in the cytoplasm. For example, miRNAs are key regulators of mammalian embryonic stem cells (ES cells), the cells of the embryo that give rise to all differentiated cell types of the organism. ES cells express the "stemness genes" *Oct4*, *Sox2*, *KLF4*, *Lin28*,



ST FIGURE 2–5 miRNAs regulate embryonic stem cell differentiation. Several genes associated with embryonic stem cell identity (*Sox2, Oct4, KLF4, Lin28,* and *Myc*) are repressed in differentiating cells by *miR-145* and *let-7* to enable cells to acquire specialized functions and inhibit self-renewal.

and *Myc*, which suppress differentiation and promote stem cell maintenance or self-renewal. Loss of these genes results in premature differentiation of ES cells while persistent activity results in an inability to produce specialized cells—both scenarios are lethal. To enable daughter cells of ES cells to differentiate, they express *miR-145* and *let-7* miRNAs, which target and downregulate stemness mRNAs (**ST Figure 2–5**). A better understanding of miRNA regulation of cellular differentiation is likely to improve stem cell therapy by ensuring that patients receive cells that properly differentiate into the desired cell types rather than fail to differentiate and pose a risk for tumor formation.

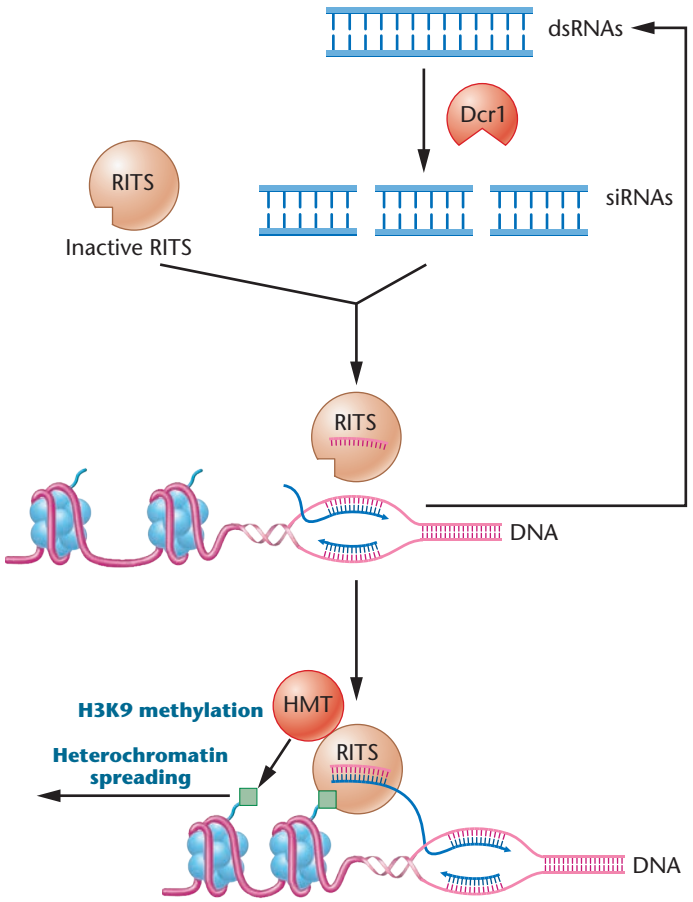
RNA-Induced Transcriptional Silencing

Heterochromatic regions of the eukaryotic chromosome, characterized by tightly packaged and transcriptionally repressed DNA, play critical structural roles (see Chapter 11). For example, centromeres are heterochromatic and are important for attaching chromosomes to spindle microtubules during cell division. Heterochromatic sequences such as centromeres are characterized by the modification of histone proteins within nucleosomes that are important for packaging and transcriptional repression. For example, nucleosomes associated with heterochromatin are generally methylated at lysine 9 on histone 3 (H3K9Me). Proteins associated with chromatin condensation and transcriptional repression such as the heterochromatin protein 1 (HP1) bind to H3K9Me and mediate heterochromatin formation. H3K9 methylation is dependent on the activity of a **histone methyltransferase** (HMT); several studies have identified DNA binding proteins that complex with HMTs and direct this complex to specific sequences. Importantly, studies have also found that sncRNAs can also direct heterochromatin formation to specific sites in the genome.

In contrast to RNAi, which results in posttranscriptional inhibition of gene expression, sncRNAs involved in heterochromatin formation use a mechanism known as **RNA-induced transcriptional silencing** (**RITS**). RITS was first described in the fission yeast *Schizosaccharomyces pombe* when the deletion of several genes involved in the RNAi pathway, such as Argonaute and Dicer, resulted in loss of H3K9 methylation at centromere nucleosomes and loss of centromere function. In 2004, the Moazed lab at Harvard University identified a complex including an Argonaute protein, a heterochromatin-associated protein, and siRNAs produced by Dicer. The siRNAs associated with this RITS complex are complementary to centromeric DNA and are required for recruiting the complex to the centromere and heterochromatin formation.

Paradoxically, transcription from heterochromatic regions such as centromeres is important for transcriptional silencing by RITS. The siRNAs that target the RITS complex to centromeres have siRNAs derived from centromeric transcription that occurs during S phase of the cell cycle. When centromeric transcription occurs, an **RNA-directed RNA polymerase** (**RdRP**) binds to nascent transcripts and catalyzes the formation of dsRNA from single-stranded transcripts. The dsRNA is then a substrate for Dicer and leads to the formation of siRNAs complementary to centromeric DNA. The RITS complex associates with these siRNAs and targets centromeric transcripts. Through an unknown mechanism, this complex at centromeric loci leads to recruitment of HMTs and H3K9 methylation, which triggers heterochromatin formation (**ST Figure 2–6**). In support of this model, tethering of the RNAi machinery to other chromosomal loci leads to heterochromatin formation and transcriptional silencing of these loci.

While RITS was characterized in *S. pombe*, RNAi machinery is known to be involved in transcriptional silencing and



ST FIGURE 2–6 RNA-induced transcriptional silencing in fission yeast (*S. pombe*). Bi-directional transcription of centromeric DNA (or RdRP activity on single-stranded transcripts) produces dsRNA that is processed by Dicer into siRNAs. These siRNAs associate with the RITS complex and recruit it to nascent transcripts. Histone methyltransferase (HMT) associated with RITS mediates H3K9 methylation of nearby nucleosomes, which leads to heterochromatin assembly.

heterochromatin formation in plants, *Drosophila*, *C. elegans*, and mammals. Further studies will help to elucidate how similar these mechanisms are in diverse species.

Long Noncoding RNAs Are Abundant and Have Diverse Functions

In addition to small noncoding RNAs discussed above, eukaryotic genomes also encode many **long noncoding RNAs** (**lncRNAs**). One obvious distinction is that lncRNAs are longer than the small noncoding RNAs and are often arbitrarily designated to be longer than 200 nucleotides. lncRNAs are produced in a similar fashion to mRNAs; following transcription they are modified with a methylated cap, a poly-A tail, and can be spliced. In contrast to mRNAs, they have **no** start and stop codons, indicating that they do not encode protein. The conservative estimate is that the human genome encodes ~17,000 lncRNAs; studies of just a small fraction of these have demonstrated that lncRNAs have diverse cellular functions.

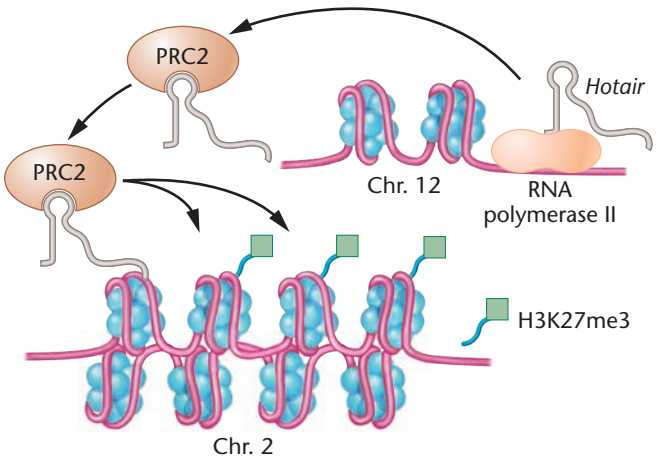
lncRNAs Mediate Transcriptional Repression by Interacting with Chromatin-Regulating Complexes

An immediate clue regarding the function of lncRNAs emerged from the observation that whereas mature mRNAs are most frequently localized to the cytoplasm, mature lncRNAs are most commonly found in the nucleus. Consistent with this finding, many lncRNAs are involved in gene regulation by recruiting chromatin-regulating complexes to specific loci in the genome. A classic example is X chromosome inactivation in mammals (see Chapter 5). The *X-inactive specific transcript* (*Xist*) gene on the X chromosome encodes a 17-kb lncRNA, which is critical to X chromosome inactivation. Of the two X chromosomes present in mammalian females, the one randomly chosen for inactivation expresses *Xist* lncRNAs that coat the chromosome. *Xist* recruits the polycomb repressor complex 2 (PRC2) to the inactivated X, which mediates chromosome-wide trimethylation of lysine 27 on histone 3 of nucleosomes, leading to chromatin condensation and transcriptional silencing. It is unclear how *Xist* spreads across the inactivated X, but recent studies demonstrate that the first loci on the X chromosome to acquire *Xist* lncRNAs are located near the *Xist* gene in terms of the three-dimensional architecture of the chromosome. Surprisingly, it appears that proximity to the site of transcription is more important than a specific sequence required for *Xist* lncRNA localization.

Similar mechanisms of lncRNA-mediated transcriptional repression are found in diverse eukaryotes and include gene-specific as well as chromosome-wide transcriptional silencing. For example, *Hotair* is a 2.2-kb lncRNA expressed from human chromosome 12 that mediates transcriptional repression of several target genes on chromosome 2. Similar to *Xist*, *Hotair* recruits PRC2 to target genes, which causes H3K27 methylation and transcriptional silencing (**ST Figure 2–7**). While the normal role of *Hotair* is not specifically known, overexpression of *Hotair* in cancer cells is linked with a poor prognosis due to increased chances for metastasis. Hopefully, future studies will help to elucidate how *Hotair* overexpression affects specific genes to promote cancer metastasis.

lncRNAs Regulate Transcription Factor Activity

lncRNAs can also associate with and regulate the activity of transcription factors. For example, the androgen receptor (AR) both binds testosterone and serves as transcription

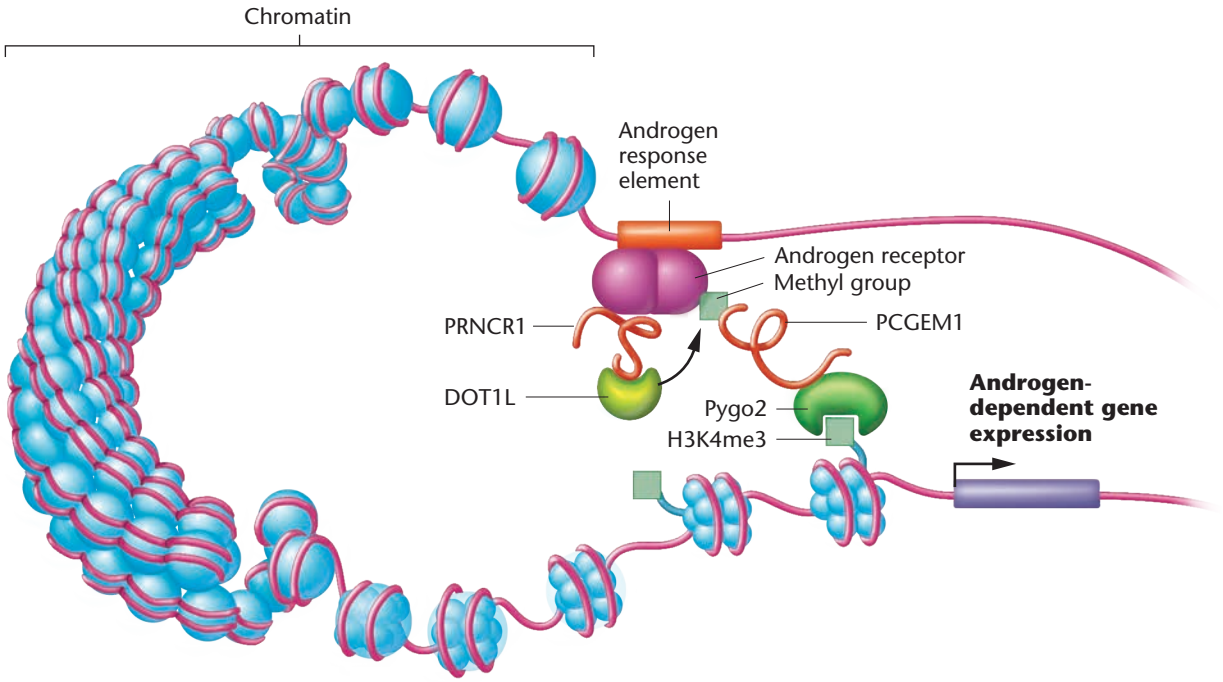


ST FIGURE 2–7 The *Hotair* long noncoding RNA regulates chromatin. The lncRNA *Hotair* is expressed from human chromosome 12 and recruits the polycomb repressor complex 2 (PRC2) to several target genes on chromosome 2. PRC2 catalyzes methylation of nearby nucleosomes (H3K27me3), which mediates chromatin condensation. Overexpression of *Hotair* in cancer indicates poor prognosis and likelihood of metastasis, presumably due to repression of the target genes on chromosome 2.

factor to activate target genes important for male development. Recent studies have demonstrated that AR's function as a transcription factor is regulated by two lncRNAs: PRNCR1 and PCGEM1. AR binds to enhancer sequences called androgen response elements (AREs) that are located at a distance from the target genes they enhance. In order to activate transcription of the target genes, the DNA is looped to bring the enhancer and AR closer to the target gene's promoter and the transcriptional machinery. PRNCR1 and PCGEM1 play important roles as "liaisons" in making this DNA loop by physically bridging the gap between AR and the promoter (**ST Figure 2–8**). Without these lncRNAs, AR cannot activate transcription of target genes. This work has added significance because androgen signaling is known to promote the growth of prostate cancer. Even eliminating testosterone by castration only delays cancer progression since the cancer cells often evolve to have mutant forms of AR that do not require testosterone for activity. Preliminary studies show that inhibiting PRNCR1 and PCGEM1 blocks the growth of both androgen-dependent and androgen-independent prostate cancer cells, suggesting that blocking these lncRNAs may be a new strategy for fighting prostate cancer.

mRNA Localization and Translational Regulation in Eukaryotes

In this final section we shift from noncoding RNAs to messenger RNAs (mRNAs). Although we have known since 1960 that mRNAs encode for proteins, we are still learning about how mRNAs are posttranscriptionally regulated in eukaryotes. It has become clear that mRNAs are not



ST FIGURE 2–8 Long noncoding RNAs mediate transcriptional activation. The lncRNAs PCGEM1 and PRNCR1 act as liaisons between the androgen receptor (AR) and the promoter to activate the transcription of target genes. PRNCR1 binds to AR and recruits DOT1L, an enzyme that methylates AR. PCGEM1 binds methylated AR and recruits Pygo2, a protein that binds methylated H3K4 at promoters.

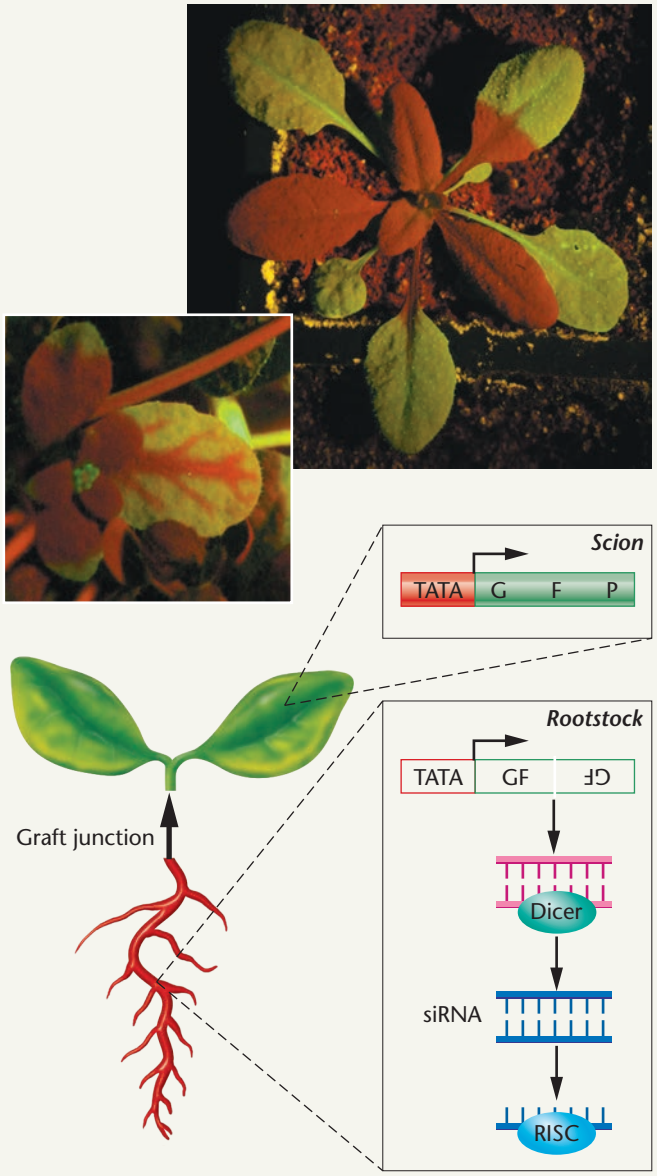
BOX 2

Do Extracellular RNAs Play Important Roles in Cellular Communication?

In addition to the abundance and diversity of RNAs within cells, RNAs are also found outside of cells. RNAs are prevalent in blood, sweat, tears, and other body fluids. These **extracellular RNAs** (**exRNAs**), which include mRNAs, miRNAs, siRNAs, lncRNAs, and tRNAs, are often enclosed in vesicles to protect them from degradation by ubiquitous RNases. Since RNAs are secreted from cells, it raises the possibility that they are internalized by other cells and can serve as a form of cell–cell communication. Although there is not yet convincing evidence that this is true in mammals, studies from plants (*A. thaliana*) and *C. elegans* clearly show that RNAs passed from cell to cell are indeed functional.

Experiments in *A. thaliana* demonstrated that exRNAs influence cells at a long distance from the cells that secrete them. When roots expressing a transgene carrying an inverted repeat RNA to target green fluorescent protein (GFP) were grafted to a plant expressing GFP in leaves, GFP expression was silenced in the leaves and siRNAs derived from the inverted repeats were found in the vascular tissue of the plant's stem (**ST Figure 2–9**). This strongly suggests that siRNAs can be transported within plants and affect gene expression in distant cells. An attractive implication of this mechanism is that a plant's ability to sense environmental conditions (such as stress) in one part of the plant can lead to an adaptive response in another part of the plant.

Studies in *C. elegans* have demonstrated that the effects of RNAi are both systemic and, in some cases, transgenerational. When worms are fed dsRNA, a membrane protein called SID-2 mediates the endocytosis of dsRNA into intestinal cells, which in turn likely secrete the dsRNAs into the body cavity. A second dsRNA channel (transport) protein, called SID-1, mediates the uptake of dsRNAs from the body cavity into most of the other cells of the worm where Dicer cleaves the dsRNAs into siRNAs that target mRNAs for destruction. Even cells of the germ line receive dsRNAs through this mechanism, thus passing the effect on to offspring. It is thought that that this mechanism evolved as part



ST FIGURE 2–9 RNAs are transported between cells in plants. *A. thaliana* scion (stem and leaves) expressing the green fluorescent protein (GFP) was grafted onto a rootstock expressing an inverted repeat of a partial GFP sequence. Transcription of the inverted repeats in roots created hairpin RNAs that were processed by Dicer to generate siRNAs. siRNAs spread through the plant and, with RISC, inhibited GFP expression in the leaves, demonstrating that siRNAs can signal between cells. GFP-positive regions are green, while non-GFP (silenced) regions are red due to autofluorescence of chlorophyll. Note that silencing is more prominent along vascular tissue, through which the siRNAs are transported.

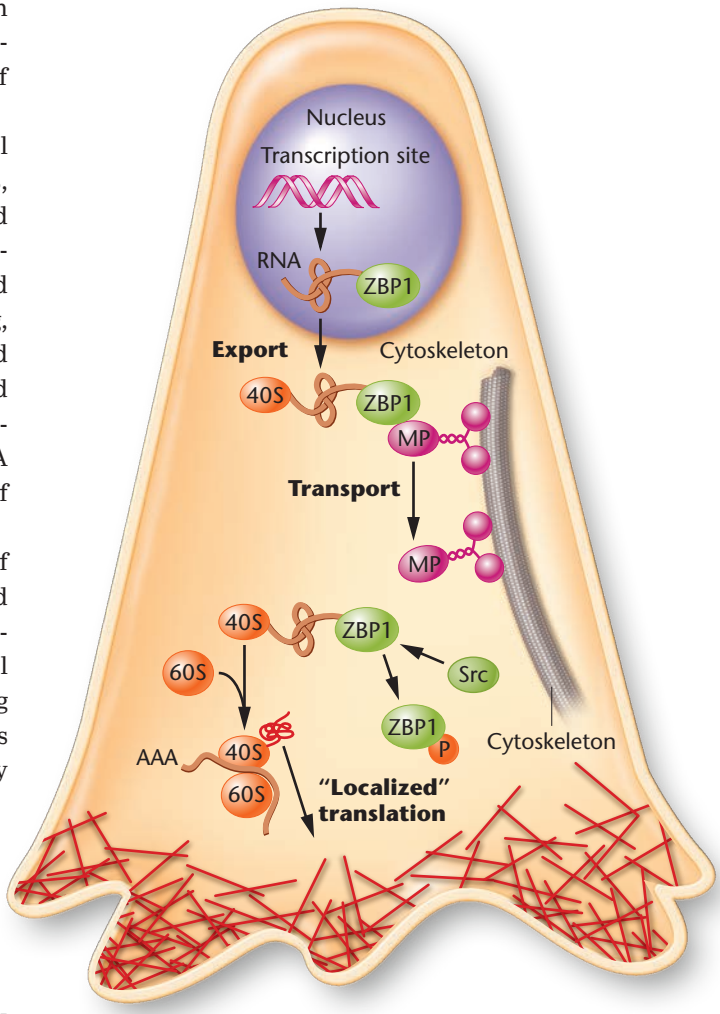
of an immune response whereby a damaged viral particle that spills its viral dsRNA triggers a systemic RNAi response against that virus preceding an infection.

In an effort to learn more about exRNAs and the roles they play in humans, in 2013 the National Institutes of Health (NIH) launched a $17 million grant program to specifically fund research in this emerging field. As Director of the NIH Francis Collins announced: "Expanding our understanding of this emerging scientific field could help us determine the role extracellular RNA plays in health and disease, and unlocking its mysteries may provide our nation's scientists with new tools to better diagnose and treat a wide range of diseases."

simply translated as soon as they reach the cytoplasm, but rather that mRNAs can be localized to discrete locations within the cell and then locally translated. mRNA localization and localized translational control create asymmetric protein distributions within the cell that define cellular regions with distinct functions. For example, different proteins localized to the highly branched dendrites of a neuron enable them to receive sensory information, while the proteins present in the axon of a neuron mediate the release of neurotransmitters that signal to other cells.

Since mRNA localization and localized translational control are common and important in diverse cell types, it is important to understand how mRNAs are localized and translationally regulated. As soon as mRNAs are transcribed, they begin to associate with a class of proteins called **RNA-binding proteins** (**RBPs**), which influence splicing, nuclear export, localization, stability, degradation, and translation. Much of the research on mRNA localization and translational regulation has focused on RBPs and their interactions with mRNAs. One of the best-described RBP/mRNA interactions is the localization and translational control of actin mRNAs in crawling cells.

Following injury, fibroblasts migrate to the site of the wound and assist in wound healing. Fibroblasts and many other types of migrating cells control their direction of movement by controlling where within the cell they polymerize new actin microfilaments. The "leading edge" of the cell where this actin polymerization occurs is called the lamellipodium. A series of elegant studies by the Singer lab at the Albert Einstein College of Medicine showed that actin mRNA is localized to lamellipodia and that localization is dependent on a 54-nucleotide element in the actin 3′ untranslated region of the mRNA transcript termed a **zip code.** The actin zip code sequence is a binding site for a RBP called **zip code binding protein 1** (**ZBP1**). ZBP1 binds actin mRNAs and prevents translation initiation as well as promotes the transport of the mRNA to the lamellipodia. Once the mRNAs arrive at the final destination, a kinase called **Src** phosphorylates ZBP1, which disrupts RNA binding and allows translation initiation (ST Figure 2–10). Since Src activity is limited to the cell periphery, this mechanism allows cells to transport mRNAs in a translationally repressed state to the cell periphery where they become translated at the site of actin polymerization. Consistent with this model, mouse fibroblasts lacking ZBP1 have reduced actin mRNA localization and reduced directional motility.



ST FIGURE 2–10 Localization and translational regulation of actin messenger RNA. The RNA binding protein ZBP1 associates with actin mRNA in the nucleus and escorts it to the cytoplasm. ZBP1 binds cytoskeleton motor proteins (MP), which transport ZBP1 and actin mRNA to the cell periphery. At the cell periphery ZBP1 is phosphorylated by Src and dissociates from actin mRNA, allowing it to be translated. Actin translation and polymerization at the leading edge direct cell movement.

There is some evidence suggesting that this actin mRNA localization mechanism is used in other cell types as well. Local translation of actin in the neurites (cellular projections) of neurons is required for neurite outgrowth and axon guidance decisions, and mouse neurons lacking ZBP1 have reduced neurite length and exhibit defects in axon guidance.

mRNA localization and translational control are likely to be widespread mechanisms in diverse species and cell types. For example, *Drosophila nanos* mRNA, which encodes a translational repressor, is localized to the posterior pole of *Drosophila* embryos and to the dendrites of some neurons. Loss of *nanos* mRNA localization results in embryos that develop with anterior/posterior patterning defects and neurons with a reduced number of dendritic branches. Interestingly, larvae mutant for *nanos* also exhibit defects in sensory function. By tagging *nanos* mRNAs with a fluorescent molecule in living cells, the Gavis lab from Princeton University demonstrated that *nanos* mRNAs are transported in **ribonucleoprotein** (**RNP**) **particles** to dendrites where they are locally translated. This transport is dependent on the motor protein dynein, which "walks" along microtubules of the cytoskeleton. These studies and many others have demonstrated how mRNA localization and localized protein synthesis are critical for neuronal function. In fact, defects in mRNA localization in neurons have been implicated in human disorders such as fragile-X syndrome, spinal muscular atrophy, and spinocerebellar ataxia.

Visit the Study Area in MasteringGenetics for a list of further readings on this topic, including journal references and selected Web sites.

Review Questions

1. What are some of the different roles that RNA plays in biological systems?
2. What arguments support the RNA World Hypothesis?
3. What types of reactions do ribozymes catalyze? What types of chemical bonds are formed or broken?
4. How is bacterial DNA methylation and expression of restriction enzymes an innate defense strategy, whereas the CRISPR/Cas system is an adaptive defense strategy?
5. What are the three types of small noncoding RNAs in eukaryotic cells, and how are they different from one another?
6. The mechanism for RNA-induced transcriptional silencing of heterochromatic DNA is paradoxical. For example, how does it make sense that centromeric DNA in fission yeast first needs to be transcribed before it can be transcriptionally silenced?
7. Although exRNAs are found in many fluids within plants and animals, why are they usually found within vesicles or bound by proteins?
8. How and why are eukaryotic mRNAs transported and localized to discrete regions of the cell?

Discussion Questions

1. The RNA World Hypothesis suggests that the earliest forms of life used RNA as a genome instead of DNA. Why then do we not see organisms alive today with RNA genomes?
2. Bacterial sRNAs can bind to mRNAs through complementary binding to regulate gene expression. What determines whether the sRNA/mRNA binding will promote or repress mRNA translation?
3. In many cases, ncRNAs serve as "guides" to direct proteins to other nucleic acids. What are some examples of ncRNAs acting as "guides," and what purposes do they serve?
4. Prokaryotes and eukaryotes have both evolved mechanisms to defend against viral/foreign nucleic acids. How are these mechanisms similar, and how are they different?
5. Extracellular RNAs are abundant in human bodily fluids, but there is currently very little evidence for their potential functions in the body. Speculate on the roles that exRNAs might play in human biology.
6. How is it possible that a given mRNA in a cell is found throughout the cytoplasm but the protein that it encodes is only found in a few specific regions of the cytoplasm? Cite a few different possibilities.

DNA Forensics

Genetics is arguably the most influential science today—dramatically affecting technologies in fields as diverse as agriculture, archaeology, medical diagnosis, and disease treatment.

One of the areas that has been the most profoundly altered by modern genetics is forensic science. **Forensic science** (or *forensics*) uses technological and scientific approaches to answer questions about the facts of criminal or civil cases. Prior to 1986, forensic scientists had a limited array of tools with which to link evidence to specific individuals or suspects. These included some reliable methods such as blood typing and fingerprint analysis, but also many unreliable methods such as bite mark comparisons and hair microscopy.

Since the first forensic use of **DNA profiling** in 1986 (Box 1), **DNA forensics** (also called **forensic DNA fingerprinting** or **DNA typing**) has become an important method for police to identify sources of biological materials. DNA profiles can now be obtained from saliva left on cigarette butts or postage stamps, pet hairs found at crime scenes, or bloodspots the size of pinheads. Even biological samples that are degraded by fire or time are yielding DNA profiles that help the legal system determine identity, innocence, or guilt. Investigators now scan large databases of stored DNA profiles in order to match profiles generated from crime scene evidence. DNA profiling has proven the innocence of hundreds of people who were convicted of serious crimes and even sentenced to death. Forensic scientists have used DNA profiling to identify victims of mass disasters such as the Asian Tsunami of 2004 and the September 11, 2001 terrorist attacks in New York. They have also used forensic DNA analysis to identify endangered species and animals trafficked in the illegal wildlife trade. The power of DNA forensic analysis has captured the public imagination, and DNA forensics is featured in several popular television series.

"Even biological samples degraded by fire or time are yielding DNA profiles that help determine identity, innocence, or guilt."

The applications of DNA profiling extend beyond forensic investigations. These include paternity and family relationship testing, identification of plant materials, verification of military casualties, and evolutionary studies.

It is important for all of us to understand the basics of forensic DNA analysis. As informed citizens, we need to monitor its uses and potential abuses. Although DNA profiling is well validated as a technique and is considered the gold standard of forensic identification, it is not without controversy and the need for legislative oversight.

In this Special Topic chapter, we will explore how DNA profiling works and how the results of profiles are interpreted. We will learn about DNA databases, the potential problems associated with DNA profiling, and the future of this powerful technology.

DNA Profiling Methods

VNTR-Based DNA Fingerprinting

The era of DNA-based human identification began in 1985, with Dr. Alec Jeffreys's publication on DNA loci known as **minisatellites,** or **variable number of tandem repeats** (**VNTRs**). As described earlier in the text (see Chapter 11), VNTRs are located in noncoding regions of the genome and are made up of DNA sequences of between 15 and 100 bp long, with each unit repeated a number of times. The number of repeats found at each VNTR locus varies from person to person, and hence VNTRs can be from 1 to 20 kilobases (kb) in length, depending on the person. For example, the VNTR

5′- GACTGCCTGCTAAGAT**GACTGCCTGCTAAGAT**
GACTGCCTGCTAAGAT-3′

is composed of three tandem repeats of a 16-nucleotide sequence (highlighted in bold).

VNTRs are useful for DNA profiling because there are as many as 30 different possible alleles (repeat lengths) at any VNTR in a population. This creates a large number of possible genotypes. For example, if one examined four different VNTR loci within a population, and each locus had 20 possible alleles, there would be more than 2 billion (4^{20}) possible genotypes in this four-locus profile.

To create a VNTR profile (also known as a DNA fingerprint), scientists extract DNA from a tissue sample and digest it with a restriction enzyme that cleaves

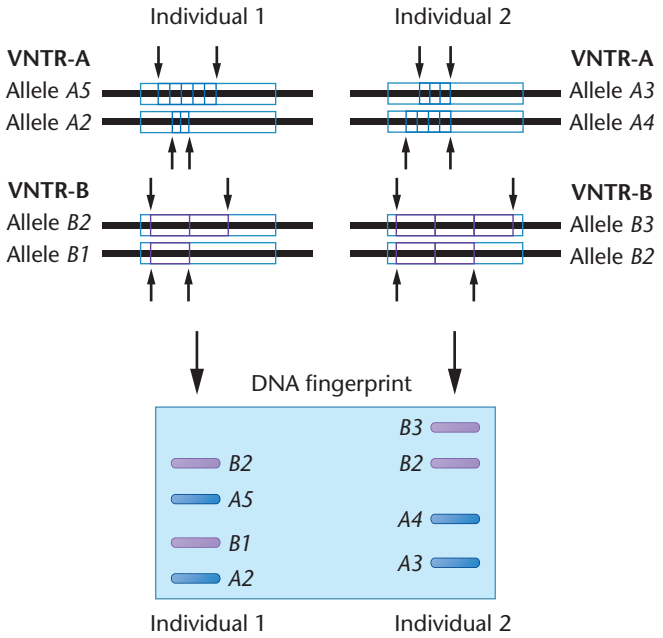
BOX 1

The Pitchfork Case: The First Criminal Conviction Using DNA Profiling

In the mid-1980s, the bodies of two schoolgirls, Lynda Mann and Dawn Ashworth, were found in Leicestershire, England. Both girls had been raped, strangled, and their bodies left in the bushes. In the absence of useful clues, the police questioned a local mentally retarded porter named Richard Buckland. During interrogation, Buckland confessed to the murder of Dawn Ashworth; however, police did not know whether he was also responsible for Lynda Mann's death. In 1986, in order to identify the second killer, the police asked Dr. Alec Jeffreys of the University of Leicester to try a new method of DNA analysis called DNA fingerprinting. Dr. Jeffreys had developed a method of analyzing DNA regions called variable number of tandem repeats (VNTRs), which vary in length between members of a population. Dr. Jeffreys's VNTR analysis revealed a match between the DNA profiles from semen samples obtained from both crime scenes, suggesting that the same person was responsible for both rapes. However, neither of the DNA profiles matched those from a blood sample taken from Richard Buckland. Having eliminated their only suspect, the police embarked on the first mass DNA dragnet in history, by requesting blood samples from every adult male in the region. Although 4000 men offered samples, one did not. Colin Pitchfork, a bakery worker, paid a friend to give a blood sample in his place, using forged identity documents. Their plan was detected when their conversation was overheard at a local pub. The conversation was reported to police, who then arrested Pitchfork, obtained his blood sample, and sent it for analysis. His DNA profile matched the profiles from the semen samples left at both crime scenes. Pitchfork confessed to the murders, pleaded guilty, and was sentenced to life in prison. The Pitchfork Case was not only the first criminal case resolved by forensic DNA profiling, but also the first case in which DNA profiling led to the exoneration of an innocent person.

on either side of the VNTR repeat region (**ST Figure 3–1**). The digested DNA is separated by gel electrophoresis and subjected to Southern blot analysis (which is described in detail in Chapter 17). Briefly, separated DNA is transferred from the gel to a membrane and hybridized with a radioactive probe that recognizes DNA sequences within the VNTR region. After exposing the membrane to X-ray film, the pattern of bands is measured, with larger VNTR repeat alleles remaining near the top of the gel and smaller VNTRs, which migrate more rapidly through the gel, being closer to the bottom. The pattern of bands is the same for a given individual, no matter what tissue is used as the source of the DNA. If enough VNTRs are analyzed, each person's DNA profile will be unique (except, of course, for identical twins) because of the huge number of possible VNTRs and alleles. In practice, scientists analyze about five or six loci to create a DNA profile.

A significant limitation of VNTR profiling is that it requires a relatively large sample of DNA (10,000 cells or about 50 μg of DNA)—more than is usually found at a typical crime scene. In addition, the DNA must be relatively intact (nondegraded). As a result, VNTR profiling has been used most frequently when large tissue samples are available—such as in paternity testing. Although VNTR profiling is still used in some cases, it has mostly been replaced by more sensitive methods, as described next.



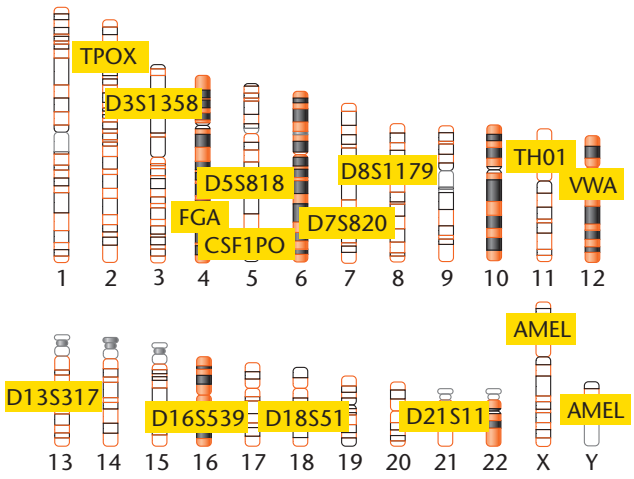
ST FIGURE 3–1 DNA fingerprint at two VNTR loci for two individuals. VNTR alleles at two loci (*A* and *B*) are shown for two different individuals. Arrows mark restriction-enzyme cutting sites that flank the VNTRs. Restriction-enzyme digestion produces a series of fragments that can be separated by gel electrophoresis and detected as bands on a Southern blot (bottom). The number of repeats at each locus is variable, so the overall pattern of bands is distinct for each individual. The DNA fingerprint profile shows that these individuals share one allele (*B2*).

Autosomal STR DNA Profiling

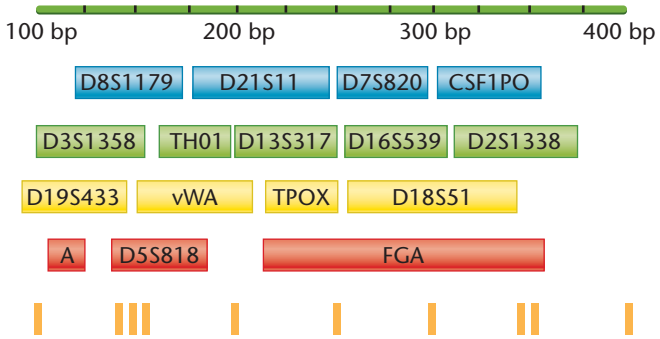
The development of the polymerase chain reaction (PCR) revolutionized DNA profiling. PCR methods are described in detail earlier in the text (see Chapter 17). Using PCR-amplified DNA samples, scientists are able to generate DNA profiles from trace samples (e.g., the bulb of single hairs or a few cells from a bloodstain) and from samples that are old or degraded (such as a bone found in a field or an ancient Egyptian mummy).

The majority of human forensic DNA profiling is now done using commercial kits that amplify and analyze regions of the genome known as **microsatellites,** or **short tandem repeats** (**STRs**). STRs are similar to VNTRs, but the repeated motif is shorter—between two and nine base pairs, repeated from 7 to 40 times. For example, one locus known as D8S1179 is made up of the four base-pair sequence TCTA, repeated 7 to 20 times, depending on the allele. There are 19 possible alleles of the locus that are found within a population. Although hundreds of STR loci are present in the human genome, only a subset is used for DNA profiling. At the present time, the FBI and other U.S. law enforcement agencies use 13 STR loci as a core set for forensic analysis (**ST Figure 3–2**). Most European countries now use 12 STR loci as a core set.

Several commercially available kits are currently used for forensic DNA analysis of STR loci. The methods vary slightly, but generally involve the following steps. As shown in **ST Figure 3–3**, each primer set is tagged by one of four fluorescent dyes—blue, green, yellow, or red. Each primer set is designed to amplify DNA fragments, the sizes of which vary depending on the number of repeats within the region amplified. For example, the primer sets that amplify the D19S433, vWA, TPOX, and D18S51 STR loci are all labeled with a yellow fluorescent tag. The sizes of the amplified DNA fragments produced allow scientists to differentiate between the yellow-labeled products. For example, the amplified products from the D19S433 locus range from about 100 to 150 bp in length, whereas those from the vWA locus range from about 150 to 200 bp, and so on.



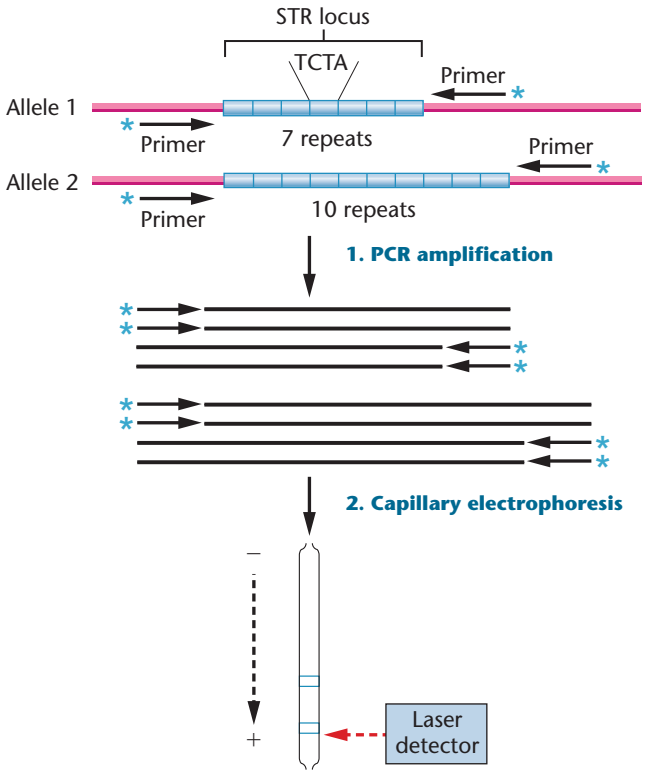
ST FIGURE 3–2 Chromosomal positions of the 13 core STR loci used for forensic DNA profiling. The AMEL (Amelogenin) locus is included with the 13 core loci and is used to determine the gender of the person providing the DNA sample. The AMEL locus on the X chromosome contains a 6-nucleotide deletion compared to that on the Y chromosome.



ST FIGURE 3–3 Relative size ranges and fluorescent dye labeling colors of 16 STR products generated by a commercially available DNA profiling kit. The DNA fragments shown in orange at the bottom of the diagram are DNA size markers. The AMEL locus is indicated as an A.

After amplification, the DNA sample will contain a small amount of the original template DNA sample and a large amount of fluorescently labeled amplification products (**ST Figure 3–4**). The sizes of the amplified fragments are measured by **capillary electrophoresis.** This method uses thin glass tubes that are filled with a polyacrylamide gel material similar to that used in slab gel electrophoresis. The amplified DNA sample is loaded onto the top of the capillary tube, and an electric current is passed through the tube. The negatively charged DNA fragments migrate through the gel toward the positive electrode, according to their sizes. Short fragments move more quickly through the gel, and larger ones more slowly. At the bottom of the tube, a laser detects each fluorescent fragment as it migrates through the tube. The data are analyzed by software that calculates both the sizes of the fragments and their quantities, and these are represented as peaks on a graph (**ST Figure 3–5**). Typically, automated capillary electrophoresis systems analyze as many as 16 samples at a time, and the analysis takes approximately 30 minutes.

After DNA profiling, the profile can be directly compared to a profile from another person, from crime scene evidence, or from other profiles stored in DNA profile databases (**ST Figure 3–6**). The STR profile genotype of an individual is expressed as the number of times the STR sequence is repeated. For example, in the profile shown in



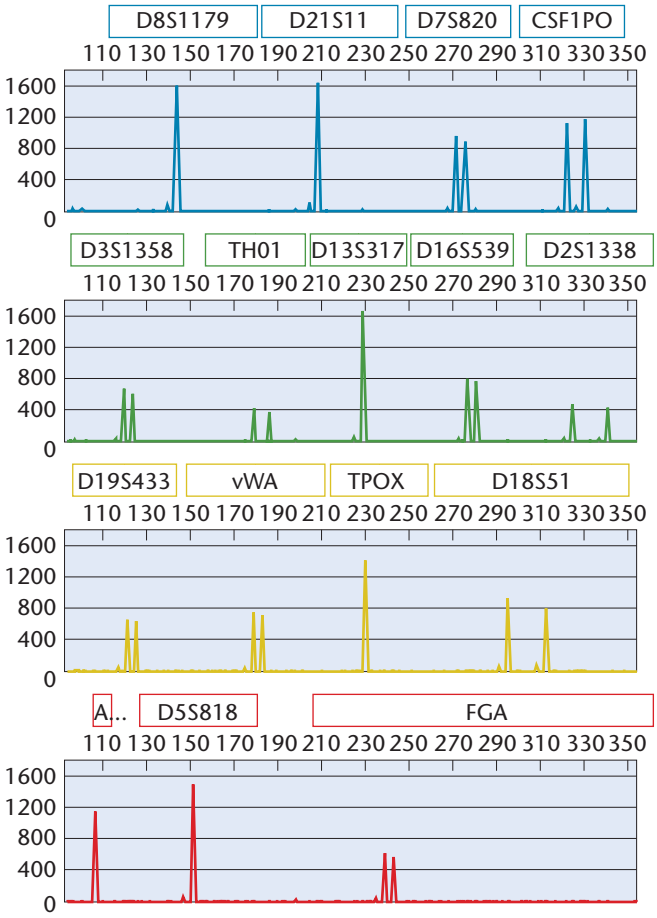
ST FIGURE 3–4 Steps in the PCR amplification and analysis of one STR locus (D8S1179). In this example, the person is heterozygous at the D8S1179 locus: One allele has 7 repeats and one has 10 repeats. Primers are specific for sequences flanking the STR locus and are labeled with a blue fluorescent dye. The double-stranded DNA is denatured, the primers are annealed, and each allele is amplified by PCR in the presence of all four dNTPs and Taq DNA polymerase. After amplification, the labeled products are separated according to size by capillary electrophoresis, followed by fluorescence detection.

ST Figure 3–6, the person's profile would be expressed as shown in **ST Table 3.1**.

Scientists interpret STR profiles using statistics, probability, and population genetics, and these methods will be discussed in the section Interpreting DNA Profiles.

Y-Chromosome STR Profiling

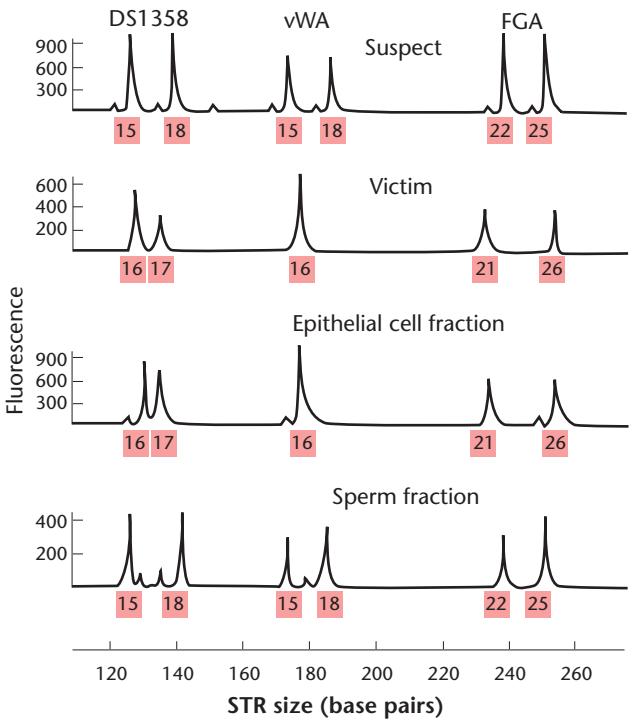
In many forensic applications, it is important to differentiate the DNA profiles of two or more people in a mixed sample. For example, vaginal swabs from rape cases usually contain a mixture of female somatic cells and male sperm cells. In addition, some crime samples may contain evidence material from a number of male suspects. In these types of cases, STR profiling of Y-chromosome DNA is useful. There are more than 200 STR loci on the Y chromosome that are useful for DNA profiling; however, fewer than 20 of these are used routinely for forensic analysis. PCR amplification of Y-chromosome STRs uses specific primers that do not amplify DNA on the X chromosome.



ST FIGURE 3–5 An electropherogram showing the results of a DNA profile analysis using the 16-locus STR profile kit shown in ST Figure 3–3. Heterozygous loci show up as double peaks and homozygous loci as single, higher peaks. The sizes of each allele can be calculated from the peak locations relative to the size axis shown at the top of each panel. The single peak for the AMEL (A) locus indicates that this DNA profile is that of a female individual, as described in ST Figure 3–2.

One limitation of Y-chromosome DNA profiling is that it cannot differentiate between the DNA from fathers and sons or from male siblings. This is because the Y chromosome is directly inherited from the father to his sons, as a single unit. The Y chromosome does not undergo recombination, meaning that less genetic variability exists on the Y chromosome than on autosomal chromosomes. Therefore, all patrilineal relatives share the same Y chromosome-profile. Even two apparently unrelated males may share the same Y profile, if they also share a distant male ancestor.

Although these features of Y-chromosome profiles present limitations for some forensic applications, they are useful for identifying missing persons when a male relative's DNA is available for comparison. They also allow researchers to trace paternal lineages in genetic genealogy studies.



ST FIGURE 3–6 Electropherogram showing the STR profiles of four samples from a rape case. Three STR loci were examined from samples taken from a suspect, a victim, and two fractions from a vaginal swab taken from the victim. The x-axis shows the DNA size ladder, and the y-axis indicates relative fluorescence intensity. The number below each allele indicates the number of repeats in each allele, as measured against the DNA size ladder. Notice that the STR profile of the sperm sample taken from the victim matches that of the suspect.

Mitochondrial DNA Profiling

Another important addition to DNA profiling methods is **mitochondrial DNA** (**mtDNA**) analysis. Between 200 and 1700 mitochondria are present in each human somatic cell. Each mitochondrion contains one or more 16-kb circular DNA chromosomes. Mitochondria divide within cells and are distributed to daughter cells after cell division. Mitochrondria are passed from the human egg cell to the zygote during fertilization; however, as sperm cells contribute few if any mitochondria to the zygote, they do not contribute these organelles to the next generation. Therefore, all cells in an individual contain multiple copies of identical mitochondria derived from the mother. Like Y-chromosome DNA, mtDNA undergoes little if any recombination and is inherited as a single unit.

Scientists create mtDNA profiles by amplifying regions of mtDNA that show variability between unrelated individuals and populations. Two commonly used regions are known as **hypervariable segment I** and **II** (**HVSI** and **HVSII**). After PCR amplification, the DNA sequence within these regions is determined by automated DNA sequencing. Scientists then compare the sequence with sequences from other individuals or crime samples, to determine whether or not they match.

The fact that mtDNA is present in high copy numbers in cells makes its analysis useful in cases where crime samples are small, old, or degraded. mtDNA profiling is particularly useful for identifying victims of mass murders or disasters, such as the Srebrenica massacre of 1995 and the World Trade Center attacks of 2001, where reference samples from relatives are available. The main disadvantage of mtDNA profiling is that it is not possible to differentiate between the mtDNA from maternal relatives or from siblings. Like Y-chromosome profiles, mtDNA profiles may be shared by two apparently unrelated individuals who also share a distant ancestor—in this case a maternal ancestor. Researchers use mtDNA profiles in scientific studies of genealogy, evolution, and human population migrations.

Mitochondrial DNA analyses have also been useful in wildlife forensics cases. Billions of dollars are generated from the illegal wildlife trade, throughout the world. Often, the identification of the species or origin of plant and animal material is the key to successful prosecution of wildlife trafficking cases. A case of illegal smuggling of bird eggs in Australia, solved by mitochondrial sequence analysis, is presented in Box 2.

ST TABLE 3.1 **STR Profile Genotypes from the Four Profiles Shown in ST Figure 3–6**

	Profile Genotype from			
STR Locus	Suspect	Victim	Epithelial Cells	Sperm Fraction
DS1358	15, 18	16, 17	16, 17	15, 18
vWA	15, 18	16, 16	16, 16	15, 18
FGA	22, 25	21, 26	21, 26	22, 25

Single-Nucleotide Polymorphism Profiling

Single-nucleotide polymorphisms (**SNPs**) are single-nucleotide differences between two DNA molecules. They may be base-pair changes or small insertions or deletions (ST Figure 3–7). SNPs occur randomly throughout the genome and on mtDNA, approximately every 500 to 1000 nucleotides. This means that there are potentially millions of loci in the human genome that can be used for profiling. However, as SNPs usually have only two alleles, many SNPs (50 or more) must be used to create a DNA profile that can distinguish between two individuals as efficiently as STRs.

Scientists analyze SNPs by using specific primers to amplify the regions of interest. The amplified DNA

BOX 2

The Pascal Della Zuana Case: DNA Barcodes and Wildlife Forensics

On August 2, 2006, a freelance photographer named Pascal Della Zuana was stopped by customs officers at Australia's Sydney International Airport. While questioning him about his flight from Thailand to Australia, officers noticed that he was wearing an unusual white vest under his outer clothing. Inside the vest, they discovered 23 concealed bird eggs.

Due to Australia's strict quarantine regulations, the eggs had to be treated with radiation in order to sterilize them. Unable to hatch the eggs, authorities turned to DNA typing in an attempt to identify the origin and species of the eggs.

The eggs were sent to Dr. Rebecca Johnson at the DNA Laboratory at the Australian Museum for forensic identification. Dr. Johnson took a small sample from each egg and extracted the DNA. She used PCR methods to amplify an approximately 650-bp region of the mitochondrial genome, within the cytochrome c oxidase 1 gene. She then organized these sequences into a format known as a DNA barcode. In order to identify the species, Dr. Johnson compared each DNA barcode to barcode entries in a large DNA barcode database compiled at the University of Guelph in Canada. The database contains mitochondrial DNA barcode sequences from hundreds of universities and museums throughout the world, cataloging more than 70,000 different species.

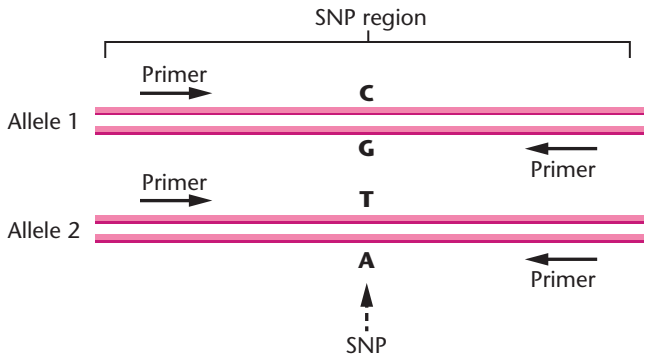
The results of Dr. Johnson's barcode sequence comparisons were dramatic. Della Zuana's vest had concealed eggs of exotic bird species such as macaws, African grey and Eclectus parrots, as well as a rare threatened species, the Moluccan cockatoo.

On January 20, 2007, Pascal Della Zuana was found guilty of contravening the Convention on International Trade in Endangered Species (CITES), as well as three Australian Customs and Quarantine Acts. He was fined $10,000 and sentenced to two years in prison.

During the court case, it was learned that, if hatched, the birds would have fetched about $250,000 on the black market. The worldwide smuggling of wildlife and wildlife parts is thought to be worth as much as US$150 billion each year—surpassed only by drugs and arms in terms of illegal profit.

regions are then analyzed by a number of different methods such as automated DNA sequencing or hybridization to immobilized probes on DNA microarrays that distinguish between DNA molecules with single-nucleotide differences.

Forensic SNP profiling has one major advantage over STR profiling. Because a SNP involves only one nucleotide of a DNA molecule, the theoretical size of DNA required for a PCR reaction is the size of the two primers and one more nucleotide (i.e., about 50 nucleotides). This feature makes SNP analysis suitable for analyzing DNA samples that are severely degraded. Despite this advantage, SNP profiling has not yet become routine in forensic applications. More frequently, researchers use SNP profiling of Y-chromosome and mtDNA loci for lineage and evolution studies.



ST FIGURE 3–7 Example of a single-nucleotide polymorphism (SNP) from an individual who is heterozygous at the SNP locus. The arrows indicate the locations of PCR primers used to amplify the SNP region, prior to DNA sequence analysis. If this SNP locus only had two known alleles—the C and T alleles—there would be three possible genotypes in the population: CC, TT, and CT. The individual in this example has the CT genotype.

Interpreting DNA Profiles

After a DNA profile is generated, its significance must be determined. In a typical forensic investigation, a profile derived from a suspect is compared to a profile from an evidence sample or to profiles already present in a DNA database. If the suspect's profile does not match that of the evidence profile or database entries, investigators can conclude that the suspect is not the source of the sample(s) that generated the other profile(s). However, if the suspect's profile matches the evidence profile or a database entry, the interpretation becomes more complicated. In this case, one could conclude that the two profiles either came from the same person—or they came from two different people who share the same DNA profile by chance. To determine the significance of any DNA profile match, it is necessary to estimate the probability that the two profiles are a random match.

ST TABLE 3.2 A Profile Probability Calculation Based on Analysis of Five STR Loci

STR Locus	Alleles from Profile	Allele Frequency from Population Database*	Genotype Frequency Calculation
D5S818	11	0.361	$2pq = 2 \times 0.361 \times 0.141 = 0.102$
	13	0.141	
TPOX	11	0.243	$p^2 = 0.243 \times 0.243 = 0.059$
	11	0.243	
D8S1179	13	0.305	$2pq = 2 \times 0.305 \times 0.031 = 0.019$
	16	0.031	
CSF1PO	10	0.217	$p^2 = 0.217 \times 0.217 \times 0.047$
	10	0.217	
D19S433	13	0.253	$2pq = 2 \times 0.253 \times 0.369 = 0.187$
	14	0.369	
Genotype frequency from this 5-locus profile = $0.102 \times 0.059 \times 0.019 \times 0.047 \times 0.187 = 0.0000009 = 9 \times 10^{-7}$			

*A U.S. Caucasian population database (Butler, J.M., et al. 2003. *J. Forensic Sci.* 48: 908–911).

© 2003 John Wiley & Sons, Inc.

The **profile probability** or **random match probability** method gives a numerical probability that a person chosen at random from a population would share the same DNA profile as the evidence or suspect profiles. The following example demonstrates how to arrive at a profile probability (ST Table 3.2).

The first locus examined in this DNA profile (D5S818) has two alleles: 11 and 13. Population studies show that the 11 allele of this locus appears at a frequency of 0.361 in this population and the 13 allele appears at a frequency of 0.141. In population genetics, the frequencies of two different alleles at a locus are given the designation p and q, following the Hardy–Weinberg law described earlier in the text (see Chapter 22). We assume that the person having this DNA profile received the 11 and 13 alleles at random from each parent. Therefore, the probability that this person received allele 11 from the mother and allele 13 from the father is expressed as $p \times q = pq$. In addition, the probability that the person received allele 11 from the father and allele 13 from the mother is also pq. Hence, the total probability that this person would have the 11, 13 genotype at this locus, by chance, is $2pq$. As we see from ST Table 3.2, $2pq$ is 0.102 or approximately 10 percent. It is obvious from this sample that using a DNA profile of only one locus would not be very informative, as about 10 percent of the population would also have the D5S818 11, 13 genotype.

The discrimination power of the DNA profile increases when we add more loci to the analysis. The next locus of this person's DNA profile (TPOX) has two identical alleles—the 11 allele. Allele 11 appears at a frequency of 0.243 in this population. The probability of inheriting the 11 allele from each parent is $p \times p = p^2$. As we see in the table, the genotype frequency at this locus would be 0.059, which is about 6 percent of the population. If this DNA profile contained only the first two loci, we could calculate how frequently a person chosen at random from this population would have the genotype shown in the table, by multiplying the two genotype probabilities together. This would be $0.102 \times 0.059 = 0.006$. This analysis would mean that about 6 persons in 1000 (or 1 person in 166) would have this genotype. The method of multiplying all frequencies of genotypes at each locus is known as the **product rule.** It is the most frequently used method of DNA profile interpretation and is widely accepted in U.S. courts.

By multiplying all the genotype probabilities at the five loci, we arrive at the genotype frequency for this DNA profile: 9×10^{-7}. This means that approximately 9 people in every 10 million (or about 1 person in a million), chosen at random from this population, would share this 5-locus DNA profile.

The Uniqueness of DNA Profiles

As we increase the number of loci analyzed in a DNA profile, we obtain smaller probabilities of a random match. Theoretically, if a sufficient number of loci were analyzed, we could be *almost* certain that the DNA profile was unique. At the present time, law enforcement agencies in North America use a core set of 13 STR loci to generate DNA profiles. A hypothetical genotype comprised of the most common alleles of each STR locus in the core STR profile would be expected to occur only once in a population of 10 billion people. Hence, the frequency of this profile would be 1 in 10 billion.

Although this would suggest that most DNA profiles generated by analysis of the 13 core STR loci would be unique on the planet, several situations can alter this interpretation. For example, identical twins share the same DNA, and their DNA profiles will be identical. Identical twins occur at a frequency of about 1 in 250 births.

In addition, siblings can share one allele at any DNA locus in about 50 percent of cases and can share both alleles at a locus in about 25 percent of cases. Parents and children also share alleles, but are less likely than siblings to share both alleles at a locus. When DNA profiles come from two people who are closely related, the profile probabilities must be adjusted to take this into account. The allele frequencies and calculations that we describe here are based on assumptions that the population is large and has little relatedness or inbreeding. If a DNA profile is analyzed from a person in a small interrelated group, allele frequency tables and calculations may not apply.

The Prosecutor's Fallacy

It is sometimes stated, by both the legal profession and the public, that "the suspect must be guilty given that the chance of a random match to the crime scene sample is 1 in 10 billion—greater than the population of the planet." This type of statement is known as the **prosecutor's fallacy** because it equates guilt with a numerical probability derived from one piece of evidence, in the absence of other evidence. A match between a suspect's DNA profile and crime scene evidence does not necessarily prove guilt, for many reasons such as human error or contamination of samples, or even deliberate tampering. In addition, a DNA profile that does not match the evidence does not necessarily mean that the suspect is innocent. For example, a suspect's profile may not match that from a semen sample at a rape scene, but the suspect could still have been involved in the crime, perhaps by restraining the victim. For these and other reasons, DNA profiles must be interpreted in the context of all the evidence in a case. A more detailed description of problems with DNA profiles is given in the next section.

DNA Profile Databases

Many countries throughout the world maintain national DNA profile databases. The first of these databases was established in the UK in 1995 and now contains approximately 6 million profiles—representing almost 10 percent of the population. In the United States, both state and federal governments have DNA profile databases. The entire system of databases along with tools to analyze the data is known as the **Combined DNA Index System** (**CODIS**) and is maintained by the FBI. As of August 2013, there were more than 11 million DNA profiles stored within the CODIS system. The two main databases in CODIS are the **convicted offender database,** which contains DNA profiles from individuals convicted of certain crimes, and the **forensic database,** which contains profiles generated from crime scene evidence. In addition, some states have DNA profile databases containing profiles from suspects and from unidentified human remains and missing persons.

SPECIAL TOPIC 3

BOX 3

The Kennedy Brewer Case: Two Bite-Mark Errors and One Hit

In 1992 in Mississippi, Kennedy Brewer was arrested and charged with the rape and murder of his girlfriend's 3-year-old daughter, Christine Jackson. Although a semen sample had been obtained from Christine's body, there was not sufficient DNA for profiling. Forensic scientists were also unable to identify the ABO blood group from the bloodstains left at the crime scene. The prosecution's only evidence came from a forensic bite-mark specialist who testified that the 19 "bite marks" found on Christine's body matched imprints made by Brewer's two top teeth. Even though the specialist had recently been discredited by the American Board of Forensic Odontology, and the defense's expert dentistry witness testified that the marks on Christine's body were actually postmortem insect bites, the court convicted Brewer of capital murder and sexual battery and sentenced him to death.

In 2001, more sensitive DNA profiling was conducted on the 1992 semen sample. The profile excluded Brewer as the donor of the semen sample. It also excluded two of Brewer's friends, and Y-chromosome profiles excluded Brewer's male relatives. Despite these test results, Brewer remained in prison for another five years, awaiting a new trial. In 2007, the Innocence Project took on Brewer's case and retested the DNA samples. The profiles matched those of another man, Justin Albert Johnson, a man with a history of sexual assaults who had been one of the original suspects in the case. Johnson subsequently confessed to Christine Jackson's murder, as well as to another rape and murder—that of a 3-year-old girl named Courtney Smith. Levon Brooks, the ex-boyfriend of Courtney's mother, had been convicted of murder in the Smith case, also based on bite-mark testimony by the same discredited expert witness.

On February 15, 2008, all charges against Kennedy Brewer were dropped, and he was exonerated of the crimes. Levon Brooks was subsequently exonerated of the Smith murder in March of 2008.

Since 1989, more than 320 people in the United States have been exonerated of serious crimes, based on DNA profile evidence. Seventeen of these people had served time on death row. In more than 100 of these exoneration cases, the true perpetrator has been identified, often through searches of DNA databases.

BOX 4

A Case of Transference: The Lukis Anderson Story

On November 30, 2012, police discovered the body of Raveesh Kumra at his home in Monte Sereno, California. Kumra's house had been ransacked, and he had suffocated from the tape used to gag him. Police collected DNA samples from the crime scene and performed DNA profiling. Several suspects were identified through matches to DNA database entries. One match, to a sample taken from Kumra's fingernails, was that of Lukis Anderson, a homeless man who was known to police. Based on the DNA profile match, Anderson was arrested, charged with murder, and jailed. He remained in jail, with a death sentence over his head, for the next five months.

The authorities believed that they had a solid case. The crime scene DNA profile was a perfect match to Anderson's DNA profile, and the lab results were accurate. Prosecutors planned to pursue the death penalty. The only problem for the prosecution was that Anderson could not have been involved in the murder, or even present at the crime scene.

On the night of the murder, Anderson had been intoxicated and barely conscious on the streets of San Jose and had been taken to the hospital, where he remained for the next 12 hours. Given his iron-clad alibi, authorities were forced to release Anderson. But they remained baffled about how an innocent person's DNA could have been found on a murder victim—one whom Anderson had never even met.

Several months after Anderson's release, prosecutors announced that they had solved the puzzle. The paramedics who had treated Anderson and taken him to the hospital had then responded to the call at Kumra's house, where they had inadvertently transferred Anderson's DNA onto Kumra's fingernails. It is not clear how the transfer had occurred, but likely Anderson's DNA had been present on the paramedics' equipment or clothing.

If Lukis Anderson had not been in the hospital with an irrefutable alibi, he may have faced the death sentence based on DNA evidence. His story illustrates how too much confidence in the power of DNA evidence can lead to false accusations. It also points to the robustness of DNA, which can remain intact, survive disinfection, and be transferred from one location to another, under unlikely circumstances.

Suspects who are not convicted can request that their profiles be removed from the databases.

DNA profile databases have proven their value in many different situations. As of August 2013, use of CODIS databases had resulted in more than 200,000 profile matches that assisted criminal investigations and missing persons searches. (Box 3). Despite the value of DNA profile databases, they remain a concern for many people who question the privacy and civil liberties of individuals versus the needs of the state.

Technical and Ethical Issues Surrounding DNA Profiling

Although DNA profiling is sensitive, accurate, and powerful, it is important to be aware of its limitations. One limitation is that most criminal cases have either no DNA evidence for analysis or DNA evidence that would not be informative to the case. In some cases, potentially valuable DNA evidence exists but remains unprocessed and backlogged. Another serious problem is that of human error. There are cases in which innocent people have been convicted of violent crimes based on DNA samples that had been inadvertently switched during processing. DNA evidence samples from crime scenes are often mixtures derived from any number of people present at the crime scene or even from people who were not present, but whose biological material (such as hair or saliva) was indirectly introduced to the site (Box 4). Crime scene evidence is often degraded, yielding partial DNA profiles that are difficult to interpret.

One of the most disturbing problems with DNA profiling is its potential for deliberate tampering. DNA profile technologies are so sensitive that profiles can be generated from only a few cells—or even from fragments of synthetic DNA. There have been cases in which criminals have introduced biological material to crime scenes, in an attempt to affect forensic DNA profiles. It is also possible to manufacture artificial DNA fragments that match STR loci of a person's DNA profile. In 2010, a research paper[1] reported methods for synthesizing DNA of a known STR profile, mixing the DNA with body fluids, and depositing the sample on crime scene items. When subjected to routine forensic analysis, these artificial samples generated perfect STR profiles. In the future, it may be necessary to develop methods to detect the presence of synthetic or cloned DNA in crime scene samples. It has been suggested that such detections could be done, based on the fact that natural DNA contains epigenetic markers such as methylation.

[1]Frumkin, D., et al. 2010. Authentication of forensic DNA samples. *Forensic Sci Int Genetics* 4: 95–103.

Many of the ethical questions related to DNA profiling involve the collection and storage of biological samples and DNA profiles. Such questions deal with who should have their DNA profiles stored on a database and whether police should be able to collect DNA samples without a suspect's knowledge or consent.

Another ethical question involves the use of DNA profiles that partially match those of a suspect. There are cases in which investigators search for partial matches between the suspect's DNA profile and other profiles in a DNA database. On the assumption that the two profiles arise from two genetically related individuals, law enforcement agencies pursue relatives of the person whose profile is stored in the DNA database. Testing in these cases is known as *familial DNA testing*. Should such searches be considered scientifically valid or even ethical?

It is now possible to accurately predict the eye and hair color of persons based on information in their DNA sample—a method known as *DNA phenotyping*. In addition, scientists are devising DNA-based tests that could provide estimates of age, height, racial ancestry, hairline, facial width, and nose size. Should this type of information be used to identify or convict a suspect?

As DNA profiling becomes more sophisticated and prevalent, we should carefully consider both the technical and ethical questions that surround this powerful new technology.

Visit the Study Area in MasteringGenetics for a list of further readings on this topic, including journal references and selected web sites.

Review Questions

1. What is VNTR profiling, and what are the applications of this technique?
2. Why are short tandem repeats (STRs) the most commonly used loci for forensic DNA profiling?
3. Describe capillary electrophoresis. How does this technique distinguish between input DNA and amplified DNA?
4. What are the advantages and limitations of Y-chromosome STR profiling?
5. How does the AMEL gene locus allow investigators to tell whether a DNA sample comes from a male or a female?
6. Explain why mitochondrial DNA profiling is often the method of choice for identifying victims of massacres and mass disasters.
7. What is a "profile probability," and what information is required in order to calculate it?
8. Describe the database system known as CODIS. What determines whether a person's DNA profile will be entered into the CODIS system?
9. What is DNA barcoding, and what types of cases use this profiling method?
10. Why is it important to understand the prosecutor's fallacy?

Discussion Questions

1. Given the possibility that synthetic DNA could be purposely introduced to a crime scene in order to implicate an innocent person, what methods could be developed to distinguish between synthetic and natural DNA?
2. Different countries and jurisdictions have different regulations regarding the collection and storage of DNA samples and profiles. What are the regulations within your region? Do you think that these regulations sufficiently protect individual rights?
3. If you were acting as a defense lawyer in a murder case that used DNA profiling as evidence against the defendant, how would you explain to the jury the limitations that might alter their interpretation of the crime scene DNA profile?
4. The phenomena of somatic mosaicism and chimerism are more prevalent than most people realize. For example, pregnancy and bone marrow transplantation may lead to a person's genome becoming a mixture of two different genomes. Describe how DNA forensic analysis may be affected by chimerism and what measures could be used to mitigate any confusion during DNA profiling. Find out more about genetic chimerism in an article by Zimmer, C., DNA double take, *New York Times*, September 16, 2013.

Genomics and Personalized Medicine

Physicians have always practiced personalized medicine in order to make effective treatment decisions for their patients. Doctors take into account a patient's symptoms, family history, lifestyle, and data derived from many types of medical tests. However, within the last 20 years, personalized medicine has taken a new and potentially powerful direction based on genetics and genomics. Today, the phrase *personalized medicine* is used to describe the application of information from a patient's unique genetic profile in order to select effective treatments that have minimal side-effects and to detect disease susceptibility prior to development of the disease.

Despite the immense quantities of medical information and pharmaceuticals that are available, the diagnosis and treatment of human disease remain imperfect. It is sometimes difficult or impossible to accurately diagnose some conditions. In addition, some patients do not respond to treatments, while others may develop side-effects that can be annoying or even life-threatening. As much of the basis for disease susceptibility and the variation that patients exhibit toward drug treatments are genetically determined, progress in genetics, genomics, and molecular biology has the potential to significantly advance medical diagnosis and treatment.

The sequencing of the human genome, the cataloging of genetic sequence variants, and the linking of sequence variants with disease susceptibility form the basis of the newly emerging field of personalized medicine. In addition, a rapidly growing list of genetic tests helps physicians determine whether a patient will have an adverse drug reaction and whether a particular pharmaceutical will be effective for that patient.

Although much of the promise of personalized medicine remains in the future, significant progress is underway. As genome technologies advance and the cost of sequencing personal genomes declines, it is becoming easier to examine a patient's unique genomic profile in order to diagnose diseases and prescribe treatments. Proponents of personalized medicine foresee a future in which each person will have his or her genome sequence determined at birth and will have the sequence stored in a digital form within a personal computerized medical file. Medical practitioners will use automated methods to scan the sequence information within these files for clues to disease susceptibility and reactions to drugs. In the future, genomic profiling and personalized medicine may allow physicians to predict which diseases you will develop, which therapeutics will work for you, and which drug dosages are appropriate.

In this Special Topic chapter, we will outline the current uses of genetic and genomic-based personalized medicine in disease diagnosis and drug selection. In addition, we will outline the future directions for personalized medicine, as well as some ethical and technical challenges associated with it.

"In the future, personalized medicine may allow physicians to predict which diseases you will develop, which therapeutics will work for you, and which drug dosages are appropriate."

Personalized Medicine and Pharmacogenomics

Perhaps the most developed area of personalized medicine is in the field of pharmacogenomics. **Pharmacogenomics** is the study of how an individual's entire genetic makeup determines the body's response to drugs. The term *pharmacogenomics* is used interchangeably with *pharmacogenetics,* which refers to the study of how sequence variation within specific candidate genes affects an individual's drug responses.

In pharmacogenomics, scientists take into account many aspects of drug metabolism and how genetic traits affect these aspects. When a drug enters the body, it interacts with various proteins, including carriers, cell-surface receptors, transporters, and metabolizing enzymes. These proteins affect a drug's target site of action, absorption, pharmacological response, breakdown, and excretion. Because so many interactions occur between a drug and proteins within the patient, many genes and many different genetic polymorphisms can affect a person's response to a drug.

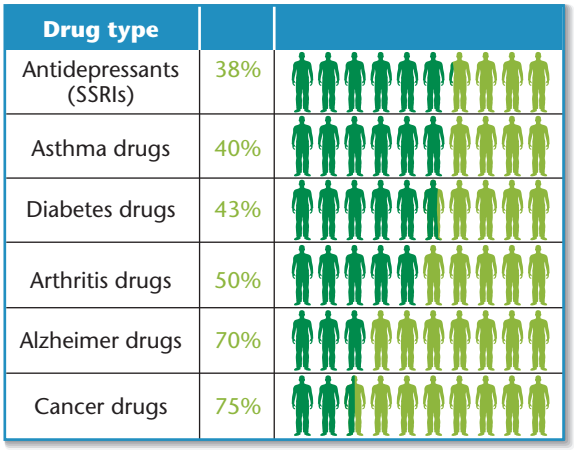
In this subsection, we examine two ways in which genomics and personalized medicine are changing the field

of pharmacogenomics: by optimizing drug therapies and by reducing adverse drug reactions.

Optimizing Drug Therapies

When it comes to drug therapy, it is clear that "one size does not fit all." On average, a drug will be effective in only about 50 percent of patients who take it (**ST Figure 4–1**). This situation means that physicians often must switch their patients from one drug to another until they find one that is effective. Not only does this waste time and resources, but also it may be dangerous to the patient who is exposed to a variety of different pharmaceuticals and who may not receive appropriate treatment in time to combat a progressive illness.

Pharmacogenomics increases the efficacy of drugs by targeting those drugs to subpopulations of patients who will benefit. One of the most common current applications of personalized pharmacogenomics is in the diagnosis and treatment of cancers. Large-scale sequencing studies show that each tumor is genetically unique, even though it may fall into a broad category based on cytological analysis or knowledge of its tissue origin. Given this genomic variability, it is important to understand each patient's mutation profile to select an appropriate treatment—particularly those newer treatments based on the molecular characteristics of tumors (Box 1).



© 2011 Personalized Medicine Coalition

ST FIGURE 4–1 Variations in patient response to drugs. This figure gives a general summary of the percentages of patients for which a particular class of drugs is ineffective.

One of the first success stories in personalized medicine was that of the ***HER-2*** gene and the use of the drug **Herceptin**® in breast cancer. The human epidermal growth factor receptor 2 (*HER-2*) gene is located on chromosome 17 and codes for a transmembrane tyrosine kinase receptor protein called HER-2. These receptors are located within the cell membranes of normal breast epithelial cells and, when bound to an extracellular growth factor (ligand),

SPECIAL TOPIC 4

BOX 1

The Story of Pfizer's Crizotinib

In 2007, Beverly Sotir was diagnosed with advanced non-small cell lung cancer (NSCLC). Beverly, a 68-year-old grandmother and nonsmoker, received standard chemotherapy, but her cancer continued to proliferate. She was given six months to live. At this same time, an apparently unrelated scientific study was underway by the pharmaceutical company, Pfizer. Pfizer had developed a compound called crizotinib, which was designed to inhibit the activity of MET, a tyrosine kinase that is abnormal in a number of tumors. Although crizotinib also inhibited another kinase called ALK (anaplastic lym-phoma kinase), scientists did not consider it significant. After clinical trials for crizotinib began, an article was published* describing a chromosomal translocation found in a small number of NSCLCs. This translocation fused the *ALK* gene to another gene called *EML4,* leading to production of a fusion protein that stimulated cancer cell growth. Pfizer immediately changed its clinical trial to include NSCLC patients. Beverly's doctors at the Dana-Farber Cancer Institute in Boston tested her tumors, discovered that they contained the *ALK/EML4* fusion gene, and enrolled Beverly in the trials. The results were dramatic. Within six months, Beverly's tumors shrunk by more than 50 percent and some disappeared entirely. As of 2011, Beverly continued to do well.

Results of the clinical trials for crizotinib showed that tumors shrank or stabilized in 90 percent of the 82 patients whose tumors contained the *ALK* fusion gene. Those patients who responded well to treatment had positive responses for up to 15 months. Scientists report that the *ALK* fusion gene tends to occur most frequently in young NSCLC patients who have never smoked. Approximately 4 percent of patients with NSCLC have this translocation in their tumor cells. Although only a small percentage of people might benefit from crizotinib, this means that about 45,000 people a year, worldwide, may be eligible for this treatment. Crizotinib is now approved in the United States for treatment of NSCLCs.

* Choi, S.M., et al. 2007. Identification of the transforming EML4-ALK fusion gene in non-small cell lung cancer. *Nature* 448: 561–566.

send signals to the cell nucleus that result in the transcription of genes involved in cell growth and division.

In about 25 percent of invasive breast cancers, the *HER-2* gene is amplified and the protein is overexpressed on the cell surface. In some breast cancers, the *HER-2* gene is present in as many as 100 copies per cell. The presence of *HER-2* overexpression is associated with increased tumor invasiveness, metastasis, and cell proliferation, as well as a poorer patient prognosis.

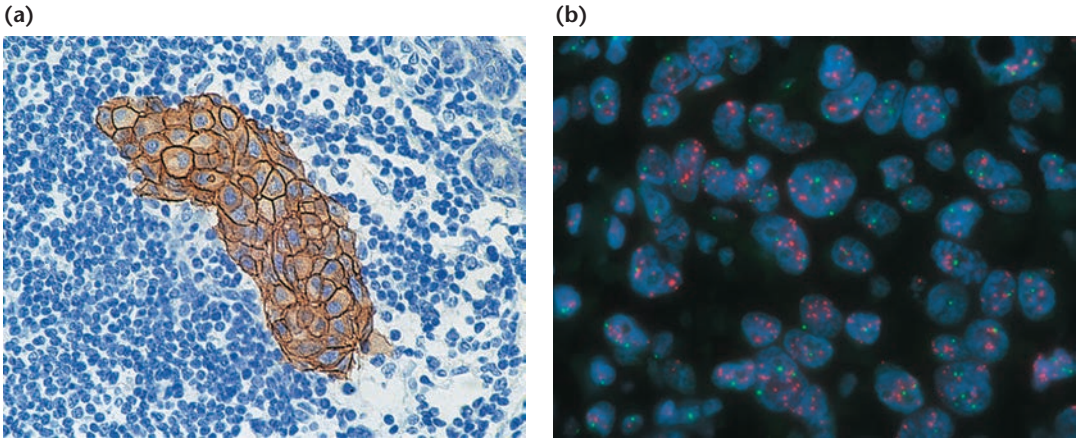
Using recombinant DNA technology, Genentech Corporation in California developed a monoclonal antibody known as trastuzumab (or Herceptin) that is designed to bind specifically to the extracellular region of the HER-2 receptor. When bound to the receptor, Herceptin appears to inhibit the signaling capability of HER-2 and may also flag the HER-2-expressing cell for destruction by the patient's immune system. In cancer cells that overexpress HER-2, Herceptin treatment causes cell-cycle arrest, and in some cases, death of the cancer cells.

Because Herceptin will only act on breast cancer cells that have amplified *HER-2* genes, it is important to know the HER-2 phenotype of each cancer. In addition, Herceptin has potentially serious side-effects. Hence, its use must be limited to those who could benefit from the treatment. A number of molecular assays have been developed to determine the gene and protein status of breast cancer cells. Two types of tests are used routinely to determine the amount of *HER-2* overexpression in cancer cells: immunohistochemistry (IHC) and fluorescence *in situ* hybridization (FISH). In IHC assays, an antibody that binds to the HER-2 protein is added to fixed tissue on a slide. The presence of bound antibody is then detected with a stain and observed under the microscope [**ST Figure 4–2(a)**]. The FISH assay (which is described in Chapter 9) assesses the number of *HER-2* genes by comparing the fluorescence signal from a HER-2 probe with a control signal from another gene that is not amplified in the cancer cells [**ST Figure 4–2(b)**].

Herceptin has had a major effect on the treatment of HER-2 positive breast cancers. When Herceptin is used in combination with chemotherapy, there is a 25 to 50 percent increase in survival, compared with the use of chemotherapy alone. Herceptin is now one of the biggest selling biotechnology products in the world, generating more than $5 billion in annual sales.

There are now dozens of drugs whose prescription and use depend on the genetic status of the target cells. Approximately 10 percent of FDA-approved drugs have labels that include pharmacogenomic information (**ST Table 4.1**). For example, about 40 percent of colon cancer patients respond to the drugs **Erbitux®** (cetuximab) and **Vectibix®** (panitumumab). These two drugs are monoclonal antibodies that bind to **epidermal growth factor receptors** (**EGFRs**) on the surface of cells and inhibit the EGFR signal transduction pathway. In order to work, cancer cells must express EGFR on their surfaces and must also have a wild-type *K-RAS* gene. The presence of EGFR can be assayed using a staining test and observation of cancer cells under a microscope. Mutations in the *K-RAS* gene can be detected using assays based on the polymerase chain reaction (PCR) method, which is described earlier in the text (see Chapter 17).

Another example of treatment decisions being informed by genetic tests is that of the **Oncotype DX®** Assay (Genomic Health Inc.). This assay analyzes the expression (amount of mRNA) from 21 genes in breast cancer samples, in order to



ST FIGURE 4–2 Protein and gene-amplification assays to determine HER-2 levels in cancer cells. (a) Normal and breast cancer cells within a biopsy sample, stained by HER-2 immunohistochemistry. Cell nuclei are stained blue. Cancer cells that overexpress HER-2 protein stain brown at the cell membrane. (b) Cancer cells assayed for *HER-2* gene copy number by fluorescence *in situ* hybridization. Cell nuclei are stained blue. *HER-2* gene DNA appears bright red. Chromosome 17 centromeres stain green. The degree of *HER-2* gene amplification is expressed as the ratio of red staining foci to green staining foci.

ST TABLE 4.1 Examples of Personalized Medicine Drugs and Diagnostics

Therapy	Gene Test	Description
Herceptin® (trastuzumab)	*HER-2* amplification	Breast cancer test to accompany Herceptin use
Erbitux® (cetuximab)	*EGFR* expression, *K-RAS* mutations	Protein and mutation analysis prior to treatment
Gleevec® (imatinib)	*BCR/ABL* fusion	Gleevec used in treatment of Philadelphia chromosome-positive chronic myelogenous leukemia
Gleevec® (imatinib)	*C-KIT*	Gleevec used in stomach cancers expressing mutated *C-KIT*
Tarceva® (erlotinib)	*EGFR* expression	Lung cancer for EGFR-positive tumors
Drugs/surgery	*MLH1, MSH2, MSH6*	Gene mutations related to colon cancers
Hormone/chemotherapies	Oncotype DX® test	Selection of breast cancer patients for chemotherapy
Chemotherapies	Aviara Cancer TYPE ID®	Classifies 39 tumor types using gene-expression assays

© 2011 Personalized Medicine Coalition

help physicians select appropriate treatments and predict the course of the disease. These genes were chosen because their levels of gene expression correlate with breast cancer recurrence after initial treatment. Based on the mRNA expression levels revealed in the assay results, scientists calculate a "Recurrence Score," estimating the likelihood that the cancer will recur within a ten-year period. Those patients with a low-risk rating would likely not benefit by adding chemotherapy to their treatment regimens and so can be treated with hormones alone. Those with higher risk scores would likely benefit from more aggressive therapies.

Reducing Adverse Drug Reactions

Every year, about 2 million people in the United States have serious side-effects from pharmaceutical drugs, and approximately 100,000 people die. The costs associated with these **adverse drug reactions** (**ADRs**) are estimated to be $136 billion annually. Although some ADRs result from drug misuse, others result from a patient's inherent physiological reactions to a drug.

Sequence variations in a large number of genes can affect drug responsiveness. Of particular significance are the genes that encode the cytochrome P450 families of enzymes. These family members are encoded by 57 different genes. People with some cytochrome P450 gene variants metabolize and eliminate drugs slowly, which can lead to accumulations of the drug and overdose side-effects. In contrast, other people have variants that cause drugs to be eliminated quickly, leading to reduced effectiveness. An example of gene variants that affect drug responses is that of *CYP2D6* gene. This member of the cytochrome P450 family encodes the debrisoquine hydroxylase enzyme, which is involved in the metabolism of approximately 25 percent of all pharmaceutical drugs, including acetaminophen, clozapine, beta blockers, tamoxifen, and codeine. There are more than 70 variant alleles of this gene. Some mutations in this gene reduce the activity of the encoded enzyme, and others can increase it. Approximately 80 percent of people are homozygous or heterozygous for the wild-type *CYP2D6* gene and are known as extensive metabolizers. Approximately 10 to 15 percent of people are homozygous for alleles that decrease activity (poor metabolizers), and the remainder of the population have duplicated genes (ultra-rapid metabolizers). Poor metabolizers are at increased risk for ADRs, whereas ultra-rapid metabolizers may not receive sufficient dosages to have an effect on their conditions.

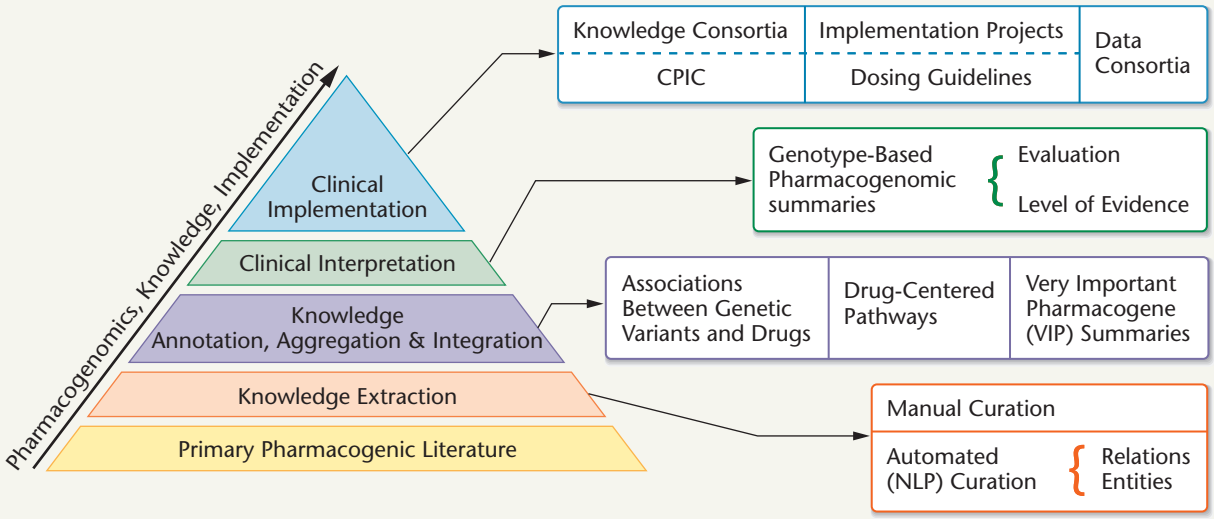
In 2005, the FDA approved a microarray gene test called the **AmpliChip® CYP450** assay (Roche Diagnostics) that detects 29 genetic variants of two cytochrome P450 genes—*CYP2D6* and *CYP2C19*. This test detects single-nucleotide polymorphisms (SNPs) as well as gene duplications and deletions. The AmpliChip CYP450 assay is an example of a genotyping microarray, such as those described earlier in the text (see Chapter 19). After scanning with an automated scanner, the data are analyzed by computer software, and the *CYP2D6/CYP2C19* genotype of the individual is generated.

Another example of pharmacogenomics in personalized medicine is that of the *CYP2C9* and *VKORC1* genes and the drug **warfarin.** Warfarin (also known as Coumadin) is an anticoagulant drug that is prescribed to prevent blood clots after surgery and to aid people with cardiovascular conditions who are prone to clots. Warfarin inhibits the vitamin K-dependent synthesis of several clotting factors. There is a more than ten-fold variability between patients in the doses of warfarin that have a therapeutic response. In the past, physicians attempted to adjust the doses of warfarin through a trial-and-error process during the first year of treatment. If the dosage of warfarin is too high, the patient may experience serious hemorrhaging; if it is too

BOX 2

The Pharmacogenomics Knowledge Base (PharmGKB): Genes, Drugs, and Diseases on the Web

The Pharmacogenomics Knowledge Base (PharmGKB) is a publicly available Internet database and information source developed by Stanford University. It is funded by the National Institutes of Health (NIH) and forms part of the NIH Pharmacogenomics Research Network, a U.S. research consortium. The goal of PharmGKB is to provide researchers and the general public with information that will increase the understanding of how genetic variation contributes to an individual's reaction to drugs. On the PharmGKB Web site (see **ST Figure 4–3**), you may search for genes and variants that affect drug reactions, information on a large number of drugs, diseases and their genetic links, pharmacogenomic pathways, gene tests, and relevant publications. Visit the PharmGKB Web site at http://www.pharmgkb.org.



ST FIGURE 4–3 The PharmGKB Knowledge Pyramid. A visual representation of the types of information available at www.pharmgkb.org.

low, the patient may develop life-threatening blood clots. It is estimated that 20 percent of patients are hospitalized during their first six months of treatment due to warfarin side-effects.

Variations in warfarin activity are affected by polymorphisms in several genes, particularly *CYP2C9* and *VKORC1*. Two single-nucleotide polymorphisms in *CYP2C9* lead to reduced elimination of warfarin and increased risk of hemorrhage. About 25 percent of Caucasians are heterozygous for one of these polymorphisms, and 5 percent appear to be homozygous. About 5 percent of patients of Asian and African descent carry these variants. Patients who are heterozygous or homozygous for some alleles of *CYP2C9* require a 10 to 90 percent lower dose of warfarin.

The FDA recommends the use of *CYP2C9* and *VKORC1* genetic tests to predict the likelihood that a patient may have an adverse reaction to warfarin. Several companies offer tests to detect polymorphisms in these genes, using methods based on PCR amplification and allele-specific primers. It is estimated that the use of warfarin genetic tests could prevent 17,000 strokes and 85,000 serious hemorrhages per year. The savings in health care could reach $1.1 billion per year.

Pharmacogenomic tests and treatments, and the genetic information on which they are based, are rapidly advancing. Updated information on all aspects of pharmacogenomics can be found on the Pharmacogenomics Knowledge Base, which is described in Box 2.

Personalized Medicine and Disease Diagnostics

The ultimate goal of personalized medicine is to apply information from a patient's full genome to help physicians diagnose disease and select treatments tailored to that particular patient. Not only will this information be gleaned from genome sequencing, but it will also be informed by gene-expression information derived from transcriptomic, proteomic, metabolomic, and epigenetic tests.

At the present time, the most prevalent use of genomic information for disease diagnostics is genetic testing that examines specific disease-related genes and gene variants. Most existing genetic tests detect the presence of mutations in single genes that are known to be linked to a disease. Currently, more than 1600 such genetic tests are available. A comprehensive list of genetic tests can be viewed on the NIH Genetic Testing Registry at www.ncbi.nlm.nih.gov/gtr/. The technologies used in many of these genetic tests are presented earlier in the text (see Chapter 19).

Genetic tests are classified according to their uses, and they fall into one or more groups. *Diagnostic tests* are designed to detect the presence or absence of gene variants or mutations linked to a suspected genetic disorder in a symptomatic patient. *Predictive tests* detect mutations and variants in patients with a family history of a known genetic disorder—for example, Huntington disease or *BRCA*-linked breast cancer. *Carrier tests* help physicians identify patients who carry a gene mutation linked to a disorder that might be passed on to their offspring—such as Tay–Sachs or cystic fibrosis. *Preimplantation tests* are performed on early embryos in order to select embryos for implantation that do not carry a suspected disease. *Prenatal tests* detect potential genetic diseases in a fetus. The test for Down syndrome is a well-known example.

Over the last decade, genome sequencing methods have progressed rapidly in speed, accuracy, and cost-effectiveness. In addition, other "omics" technologies such as transcriptomics and proteomics are providing major insights into how DNA sequences lead to gene expression and, ultimately, to phenotype. (Refer to Chapter 18 for descriptions of techniques and data emerging from human "omics" technologies.) As these technologies become more rapid and cost-effective, they will begin to make important contributions to personalized medicine.

Although the application of "omics" to personalized medicine has not yet entered routine medical care, several proof-of-principle cases illustrate the way in which whole-genome analysis may develop in the future. They also reveal some of the limitations that must be overcome before genome-based medicine becomes commonplace and practical. In the next two sections, we will describe several of these studies as they pertain to the diagnosis of cancers and other diseases.

Personal Genomics and Cancer

As we learned earlier in the text (see Chapter 16), cancer is a genetic disease at the level of somatic cells. High-throughput sequencing of normal and cancer genomes, along with RNA sequencing and protein profiling of normal and cancer cells, has revealed more of the mutations and gene rearrangements associated with specific cancers. Studies such as the Cancer Genome Atlas project are amassing data equivalent to 20,000 genome projects on normal and tumor DNA from patients with more than 20 different types of cancer. Such studies are revealing that cancers once classified in general terms (such as "prostate cancer") are in fact many different diseases based on their genetic profiles. For example, in the past, blood cancers were categorized into two large groups: leukemias and lymphomas. Today, we know that each category can be broken down into dozens of different types, based on gene mutation and expression characteristics. Similarly, breast cancer is now thought to be at least 10 separate diseases, based on genomic and gene-expression data. The recognition that tumors differ significantly in gene expression will likely be used in the future to tailor therapies to attack or modify specific gene-expression aberrations.

Another significant discovery from cancer genome research is that every tumor is genetically unique, even though common cellular pathways are involved. This realization indicates that each cancer may require a personalized treatment and that the genomic "net" that is cast to detect altered gene function must be wide enough to capture all relevant defects within each cancer. The potential for whole-genome sequencing and gene-expression assays in cancer diagnosis and treatment is illustrated by a case described in Box 3.

This story illustrates the enormous quantities of resources involved in genomic sequencing and gene-expression assays, as well as the interpretation of the resulting data. It also shows that genomic sequencing alone may not be sufficient to detect the most important defects in cancer cells, including those that would be suitable targets for therapy. The story points out that few gene-specific drugs are currently available and those that do exist are expensive and may not be covered by medical insurance. The patient in this story had been fortunate that a key defect in his cancer had been detectable using genomic techniques and could be targeted by an existing drug. As most cancers contain dozens to hundreds of genetic and gene-expression defects, and more than one gene product may drive the cells to form cancers, the goal of developing drugs for each of these defects remains a challenging one. Despite these challenges, it is a story of future promise for the role of cancer diagnosis and gene-specific treatments in personalized medicine.

Personal Genomics and Disease Diagnosis: Analyzing One Genome

In 2010, the journal *Lancet* published a report illustrating the type of information that we can currently obtain from personal whole-genome sequencing.[1] The personal genome

[1]Ashley, E.A., et al. 2010. Clinical assessment incorporating a personal genome. *Lancet* 375: 1525–1535.

BOX 3

Personalized Cancer Diagnostics and Treatments: The Lukas Wartman Story[1]

During his final year of medical school in 2002, Dr. Lukas Wartman began to experience symptoms of fatigue, fever, and bone pain. After months of tests, he was given a diagnosis of adult acute lymphoblastic leukemia (ALL). Following two years of chemotherapies, his cancer went into remission for three years. When the ALL recurred, his doctors treated him with intensive chemotherapy and a bone marrow transplant, which put him back into remission for another three years. After his second relapse, all attempts at treatment failed and he was rapidly deteriorating.

At the time of his second relapse, Dr. Wartman was working as a physician-scientist at Washington University, researching the genetics of leukemias. His colleagues, including Dr. Timothy Ley, associate director of the Washington University Genome Institute, decided to rush into a last-minute effort to save him. Using the university's sequencing facilities and supercomputers, the research team sequenced the entire genomes of his normal and cancer cells. They also analyzed his RNA types and expression levels using RNAseq technologies.

As they had expected, Dr. Wartman's cancer cells contained many gene mutations. Unfortunately, there were no known drugs that would attack the products of these mutated genes. The RNA sequence analysis, however, revealed unexpected results. It showed that the fms-related tyrosine kinase 3 (*FLT3*) gene, although having a normal DNA sequence, was overexpressed in his cancer cells—perhaps due to mutations in the gene's regulatory regions. The *FLT3* gene encodes a protein kinase that is involved in normal hematopoietic cell growth and differentiation, and its overexpression would be a potentially important contributor to Dr. Wartman's cancer. Equally interesting, and fortunate, was that the drug sunitinib (Sutent) was known to inhibit the FLT3 kinase and had been approved for use in the treatment of some kidney and gastrointestinal cancers.

Dr. Wartman decided to try sunitinib. Unfortunately, the drug cost $330 per day, and Dr. Wartman's insurance company refused to pay for it. In addition, the drug company Pfizer refused to supply the drug to him under its compassionate use program. Despite these setbacks, he collected enough money to buy a week's worth of sunitinib. Within days of starting treatment, his blood counts were approaching normal. Within two weeks his bone marrow was free of cancer cells. At this point Pfizer reversed its decision and supplied Dr. Wartman with the drug. In addition, he underwent a second bone marrow transplant to help ensure that the cancer would not return. Although Dr. Wartman's long-term prognosis is still uncertain, his successful experience with personalized cancer treatment has given him hope and has spurred research into the regulation of the *FLT3* gene in other cancers.

[1]Kolata, G., In treatment for leukemia, glimpses of the future. *New York Times,* July 7, 2012.

sequence in this study was the first one to be sequenced using a method known as true single-molecule sequencing (tSMS™). Some high-throughput methods, such as those described earlier in the text (see Chapters 17 and 18), require cloning or PCR amplification of template DNA prior to sequencing. In contrast, the tSMS method directly sequences individual genomic DNA strands with minimum processing. The sequencing of this genome took about a week, was performed with one machine, used the services of three people, and cost $48,000. The genome sequence was that of Dr. Stephen Quake, a Stanford University professor who developed the technology and headed the research group. He was a healthy 40-year-old male who had a family history of arthritis, aortic aneurysm, coronary artery disease, and sudden cardiac death.

By comparing Dr. Quake's DNA sequence with other human genome sequences in databases, they discovered a total of 2.6 million SNPs and 752 copy number variations. The researchers then sorted through the genome sequence data to determine which of these variants might have an effect on phenotype. This was accomplished by searching known SNPs in several large databases, manually creating their own disease-associated SNP database, and calculating likelihood ratios for various disease risks. The analysis required the combined efforts of more than two dozen scientists and clinicians over a period of about a year, and information gleaned from more than a dozen sequence databases, new and existing sequence analysis tools, and hundreds of individually accessed research papers.

To determine how Dr. Quake may respond to pharmaceutical drugs, the researchers searched the PharmGKB database (see Box 2) for the presence of known variants within pharmacogenomically important genes. He was found to have 63 clinically relevant SNPs within genes associated with drug reactions. In addition, his genome contained six previously unknown SNPs that could alter amino acid sequences in drug-response genes. For example, the genome sequence revealed that Dr. Quake was heterozygous for a null mutation in the *CYP2C19* gene. This mutation could make him sensitive to a range of drugs, including those used to treat aspects of heart disease. He would also be more sensitive than normal to warfarin, based on SNPs within his *VKORC1* and *CYP4F2*

genes. In contrast, Dr. Quake's DNA sequence contained gene variants associated with good responses to statins; however, other gene variants suggested that he might require higher-than-normal statin dosages.

The search for mutations within genes that directly affect disease conditions revealed several potentially damaging variants. Dr. Quake was heterozygous for a SNP within the *CFTR* gene that would change a glycine to arginine at position 458. This mutation could lead to cystic fibrosis if it was passed on to a son or daughter who also inherited a defective *CFTR* gene from the other parent. Similarly, he was heterozygous for a recessive mutation in the hereditary haemochromatosis protein precursor gene (*HFE*), which is associated with the development of haemochromatosis, a serious condition leading to toxic accumulations of iron. Also, Dr. Quake was heterozygous for a recessive mutation in the solute carrier family 3 (*SLC3A1*) gene. This mutation is linked to cystinuria, an inherited disorder characterized by inadequate excretion of cysteine and development of kidney stones. The scientists discovered a heterozygous SNP within the parafibromin (*CDC73*) gene that would create a prematurely terminated protein. This gene is a tumor-suppressor gene linked to the development of hyperparathyroidism and parathyroid tumors. The presence of this SNP increased the risk that Dr. Quake might develop these types of tumors, if any of his cells experienced a loss-of-heterozygosity mutation in the other copy of the gene.

The analysis of Dr. Quake's genome sequence for the purpose of predicting future development of multifactorial disease was more challenging. Genome-wide association studies have revealed large numbers of sequence variants that are associated with complex diseases; however, each of these variants most often contributes only a small part of the susceptibility to disease. Because not all variants have been discovered or characterized, it is difficult to establish a numerical risk score for each of these diseases based on the presence of one or more SNPs. As an example, the researchers discovered SNPs within three genes (*TMEM43*, *DSP*, and *MYBPC3*) that may be associated with sudden cardiac death. However, the exact effects of two of these SNPs are still unclear, and the other SNP had not previously been described. Dr. Quake had five SNPs in genes associated with an increased risk of developing myocardial infarction and two SNPs associated with a lower risk. Among the SNPs associated with increased risk, a variant in the apolipoprotein A precursor (*LPA*) gene is associated with a five-fold increased plasma lipoprotein(a) concentration and a two-fold increased risk of coronary artery disease. By taking into consideration the simultaneous potential effects of many different SNPs, as well as the patient's own environmental and personal lifestyle factors, the researchers concluded that Dr. Quake's genetics contributed to a significantly increased risk for eight conditions (such as Type 2 diabetes, obesity, and coronary artery disease) and a decreased risk for seven conditions (such as Alzheimer disease). Dr. Quake was offered the services of clinical geneticists, counselors, and clinical lab directors in order to help interpret the information generated from the genome sequence. Genetic counseling covered areas such as the psychological and reproductive implications of genetic disease risk, the possibilities of discrimination based on genetic test results, and the uncertainties in risk assessments.

In 2012, another study of personal genome analysis was reported (Box 4). This study combined data from whole-genome sequencing, transcriptomics, proteomics, and metabolomics profiles from a single patient at multiple time points over a 14-month period. This in-depth multilevel personal profiling allowed the patient to be monitored through both healthy and diseased states, as he contracted two virus infections and a period of Type 2 diabetes. This research points out how complex changes in gene expression may affect phenotype and shows the importance of looking beyond the raw sequence of an individual DNA. It also indicates that gene-expression profiles can be monitored by current technologies and may be applied in the future as part of personalized medical testing.

Technical, Social, and Ethical Challenges

There are still many technical hurdles to overcome before personalized medicine will become a standard part of medical care. The technologies of genome sequencing, "omics" profiling, microarray analysis, and SNP detection need to be faster, more accurate, and cheaper. Scientists expect that these challenges will be overcome in the near future; however, genome analysis needs to be used with caution until the technology becomes highly accurate and reliable. Even a low rate of error in genetic sequences or test results could lead to misdiagnoses and inappropriate treatments. Another challenge will be the storage and interpretation of vast amounts of genomic sequence data. Each personal genome generates the letter-equivalent of 200 large phone books, which must be stored in databases, mined for relevant sequence variants, and meaning assigned to each sequence variant. To undertake these kinds of analyses, scientists need to gather data from large-scale population genotyping studies that will link sequence variants to phenotype, disease, or drug responses. Experts suggest that such studies will take the coordinated efforts of public and private research teams and more than a decade to complete. Scientists will also need to develop efficient automated

BOX 4

Beyond Genomics: Personal Omics Profiling

A study published by a research team led by Dr. Michael Snyder of Stanford University provides an example of how multiple "omics" technologies can be used to examine one person's healthy and diseased states.[1]

Blood samples were taken from a healthy individual (Dr. Snyder) at 20 time points over a 14-month study period. Dr. Snyder's whole-genome sequence was generated at each time point using two different methods and backed up by exome sequencing using three different methods. In addition, his genome sequence was compared to that of his mother. Concurrently, whole-transcriptome sequencing, proteomic profiling, and metabolomics assays were performed.

Dr. Snyder's genome sequence revealed a number of SNPs that are known to be associated with elevated risks for coronary artery disease, basal cell carcinoma, hypertriglyceridemia, and Type 2 diabetes. A mutation in the *TERT* gene, which is involved in telomere replication, gave an increased risk for aplastic anemia. These data were followed by a series of medical tests. Dr. Snyder had no signs of aplastic anemia, and his telomere lengths were close to normal. Similarly, his mother, who shared his mutation in the *TERT* gene, had no symptoms of aplastic anemia. Medical tests revealed he did have elevated triglyceride levels, which he subsequently controlled using medication. Blood glucose levels were initially normal but became abnormally high after he became infected with respiratory syncytial virus (RSV). In response to these data, Dr. Snyder modified his diet and exercise regime and later brought his blood glucose down to normal levels. An analysis of drug response gene variants revealed that he should have good responses to diabetic drugs.

Using RNAseq technologies, the researchers monitored the numbers and types of more than 19,000 mRNAs and miRNAs transcribed from more than 12,000 genes over 20 time points. The data showed that sets of genes were coordinately regulated in response to conditions such as RSV infection and glucose levels. The researchers also found that RNA species underwent differential splicing and editing during changes in physiological states. Editing events included changes of adenosine to inosine and cytidine to uridine, and many of these RNA edits altered the amino acid sequences of translated proteins.

The researchers also profiled the levels of more than 6000 proteins and metabolites over the time course of the study. Like the RNA data, the protein and metabolite data showed coordinated changes that occur through virus infections and glucose-level changes. Some of these changes were shared between RNA, protein, and metabolites, and others were unique to each category. The medical significance of these patterns will be addressed in future studies.

[1]Chen, R. et al. 2012. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 148: 1293–1307.

systems and algorithms to deal with this massive amount of information. Moreover, these data analyses will have to consider that genetic variants contribute only partially to personal phenotype. Personalized medicine will also need to integrate information about environmental, personal lifestyle, and epigenetic factors.

Another technical challenge for personalized medicine is the development of automated health information technologies. Health-care providers will need to use electronic health records to store, retrieve, and analyze each patient's genomic profile, as well as to compare this information with constantly advancing knowledge about genes and disease. Currently, fewer than 10 percent of hospitals and physicians in the United States have access to these types of information technologies.

Personalized medicine has a number of societal implications. To make personalized medicine available to everyone, the costs of genetic tests, as well as the genetic counseling that accompanies them, must be reimbursed by insurance companies, even in cases where there are no prior diseases or symptoms. Regulatory changes are required to ensure that genetic tests and genomic sequencing are accurate and that the data generated are reliably stored in databases that guarantee the patient's privacy. At the present time, less than 1 percent of genetic tests are regulated by agencies such as the FDA.

Personalized medicine also requires changes to medical education. In the future, physicians will be expected to use genomics information as part of their patient management. For this to be possible, medical schools will need to train future physicians to interpret and explain genetic data. In addition, more genetic counselors and genomics specialists will be required. These specialists will need to understand genomics and disease, as well as to manipulate bioinformatic data. As of 2010, there were only about 2500 genetic counselors and 1100 clinical geneticists in North America.

The ethical aspects of the new personalized medicine are also diverse and challenging. For example, it is sometimes argued that the costs involved in the development of genomics and personalized medicine are a misallocation of limited resources. Some argue that science should solve larger problems facing humanity, such as the distribution of food and clean water, before embarking on personalized medicine. Similarly, some critics argue that such highly

specialized and expensive medical care will not be available to everyone and represents a worsening of economic inequality. There are also concerns about how we will protect the privacy of genome information that is contained in databases and private health-care records. In addition, there need to be effective ways to prevent discrimination in employment or insurance coverage, based on information derived from genomic analysis.

Most experts agree that we are at the beginning of a personalized medicine revolution. Information from genetics and genomics research is already increasing the effectiveness of drugs and enabling health-care providers to predict diseases prior to their occurrence. In the future, personalized medicine will touch almost every aspect of medical care. By addressing the upcoming challenges of the new personalized medicine, we can guide its use for the maximum benefit to the greatest number of people.

Visit the Study Area in MasteringGenetics for a list of further readings on this topic, including journal references and selected web sites.

Review Questions

1. What is pharmacogenomics, and how does it differ from pharmacogenetics?
2. Describe how the drug Herceptin works. What types of gene tests are ordered prior to treatment with Herceptin?
3. What is the Oncotype DX Assay, and how is it used?
4. How do the cytochrome P450 proteins affect drug responses? Give two examples.
5. What types of genetic tests are currently available, and how are they classified?
6. Give two examples of how genomic studies have altered our understanding of cancers.
7. Why is it necessary to examine gene-expression profiles, in addition to genome sequencing, for effective personalized medicine?
8. Using the PharmGKB database, explain the relationship between *CYP2D6* variants and the response of patients to the breast cancer drug, tamoxifen.

Discussion Questions

1. In this chapter, we present three case studies that use personalized genomics analysis to predict and treat diseases. Although these cases have shown how personalized medicine may evolve in the future, they have triggered controversy. What are some objections to these types of studies, and how can these objections be addressed?
2. What are the biggest challenges that must be overcome before personalized medicine becomes a routine component of medical care? What do you think is the most difficult of these challenges and why?
3. How can we ensure that a patient's privacy is maintained as genome information accumulates within medical records? How would you feel about allowing your genome sequence to be available for use in research?
4. As gene tests and genomic sequences become more commonplace, how can we prevent the emergence of "genetic discrimination" in employment and medical insurance?

Genetically Modified Foods

Throughout the ages, humans have used selective breeding techniques to create plants and animals with desirable genetic traits. By selecting organisms with naturally occurring or mutagen-induced variations and breeding them to establish the phenotype, we have evolved varieties that now feed our growing populations and support our complex civilizations.

Although we have had tremendous success shuffling genes through selective breeding, the process is a slow one. When recombinant DNA technologies emerged in the 1970s and 1980s, scientists realized that they could modify agriculturally significant organisms in a more precise and rapid way by identifying and cloning genes that confer desirable traits, then introducing these genes into organisms. Genetic engineering of animals and plants promised an exciting new phase in scientific agriculture, with increased productivity, reduced pesticide use, and enhanced flavor and nutrition.

Beginning in the 1990s, scientists created a large number of genetically modified (GM) food varieties. The first one, approved for sale in 1994, was the Flavr Savr tomato—a tomato that stayed firm and ripe longer than non-GM tomatoes. Soon afterward, other GM foods were developed: papaya and zucchini with resistance to virus infection, canola containing the tropical oil laurate, corn and cotton plants with resistance to insects, and soybeans and sugar beets with tolerance to agricultural herbicides. By 2012, more than 200 different GM crop varieties had been created. Worldwide, GM crops are planted on 170 million hectares of arable land, with a global value of $15 billion for GM seeds.

"Genetic engineering of animals and plants promised an exciting new phase in scientific agriculture, with increased productivity, reduced pesticide use, and enhanced flavor and nutrition."

Although many people see great potential for GM foods—to help address malnutrition in a world with a growing human population and climate change—others question the technology, oppose GM food development, and sometimes resort to violence to stop the introduction of GM varieties (ST Figure 5–1). Even Golden Rice—a variety of rice that contains the vitamin A precursor and was developed on a humanitarian nonprofit basis to help alleviate vitamin A deficiencies in the developing world—has been the target of opposition and violence.

Some countries have outright bans on all GM foods, whereas others embrace the technologies. Opponents cite safety and environmental concerns, whereas some scientists and commercial interests extol the almost limitless virtues of GM foods. The topic of GM food attracts hyperbole and exaggerated rhetoric, information, and misinformation—on both sides of the debate.

So, what are the truths about GM foods? In this Special Topic chapter, we will introduce the science behind GM foods and examine the promises and problems of the new technologies. We will look at some of the controversies and present information to help us evaluate the complex questions that surround this topic.

What Are GM Foods?

GM foods are derived from **genetically modified organisms** (**GMOs**), specifically plants and animals of agricultural importance. GMOs are defined as organisms whose genomes have been altered in ways that do not occur naturally. Although the definition of GMOs sometimes includes organisms that have been genetically modified by selective breeding, the most commonly used definition refers to organisms modified through genetic engineering or recombinant DNA technologies. Genetic engineering allows one or more genes to be cloned and transferred from one organism to another—either between individuals of the same species or between those of unrelated species. It also allows an organism's endogenous genes to be altered in ways that lead to enhanced or reduced expression levels. When genes are transferred between unrelated species, the resulting organism is called **transgenic.** The term **cisgenic** is sometimes used to describe gene transfers within a species. In contrast, the term **biotechnology** is a more general one, encompassing a wide range of methods that manipulate organisms or their components—such as isolating enzymes or producing wine, cheese, or yogurt. Genetic modification of plants or animals is one aspect of biotechnology.

ST FIGURE 5–1 Anti-GM protesters attacking a field of genetically modified maize in southwestern France. In July 2004, hundreds of activists opposed to GM crops destroyed plants being tested by the U.S. biotech company Pioneer Hi-Bred International.

In 2012, it was estimated that GM crops were grown in approximately 30 countries on 11 percent of the arable land on Earth. The majority of these GM crops (almost 90 percent) are grown in five countries—the United States, Brazil, Argentina, Canada, and India. Of these five, the United States accounts for approximately half of the acreage devoted to GM crops. According to the U.S. Department of Agriculture, 93 percent of soybeans and 90 percent of maize grown in the United States are from GM crops. In the United States, more than 70 percent of processed foods contain ingredients derived from GM crops.

Soon after the release of the Flavr Savr tomato in the 1990s, agribusinesses devoted less energy to designing GM foods to appeal directly to consumers. Instead, the market shifted toward farmers, to provide crops that increased productivity. By 2012, approximately 200 different GM crop varieties were approved for use as food or livestock feed in the United States. However, only about two dozen are widely planted. These include varieties of soybeans, corn, sugar beets, cotton, canola, papaya, and squash. ST Table 5.1 lists some of the common GM food crops available for planting in the United States. Of these GM crops, by far the most widely planted are varieties that are herbicide tolerant or insect resistant. At the time of writing this chapter, no GM food animal was approved for consumption, although a GM salmon variety was nearing market approval in the United States (Box 1). A number of agriculturally important animals such as goats and sheep have been genetically modified to produce pharmaceutical products in their milk. The use of transgenic animals as bioreactors is discussed earlier in the text (see Chapter 19).

ST TABLE 5.1 Some GM Crops Approved for Food, Feed, or Cultivation in the United States*

Crop	Number of Varieties	GM Characteristics
Soybeans	19	Tolerance to glyphosate herbicide Tolerance to glufosinate herbicide Reduced saturated fats Enhanced oleic acid Enhanced omega-3 fatty acid
Maize	68	Tolerance to glyphosate herbicide Tolerance to glufosinate herbicide Bt insect resistance Enhanced ethanol production
Cotton	30	Tolerance to glyphosate herbicide Bt insect resistance
Potatoes	28	Bt insect resistance
Canola	23	Tolerance to glyphosate herbicide Tolerance to glufosinate herbicide Enhanced lauric acid
Papaya	4	Resistance to papaya ringspot virus
Sugar beets	3	Tolerance to glyphosate herbicide
Rice	3	Tolerance to glufosinate herbicide
Zucchini squash	2	Resistance to zucchini, watermelon, and cucumber mosaic viruses
Alfalfa	2	Tolerance to glyphosate herbicide
Plum	1	Resistance to plum pox virus

* Information from the International Service for the Acquisition of Agri-Biotech Applications, www.isaaa.org.

Herbicide-Resistant GM Crops

Weed infestations destroy about 10 percent of crops worldwide. To combat weeds, farmers often apply herbicides before seeding a crop and between rows after the crops are growing. As the most efficient broad-spectrum herbicides also kill crop plants, herbicide use may be difficult and limited. Farmers also use tillage to control weeds; however, tillage damages soil structure and increases erosion.

Herbicide-tolerant varieties are the most widely planted of GM crops, making up approximately 70 percent of all GM crops. The majority of these varieties contain a bacterial gene that confers tolerance to the broad-spectrum herbicide **glyphosate**—the active ingredient in commercial herbicides such as Roundup®. Studies have shown that glyphosate is effective at low concentrations, is degraded rapidly in soil and water, and is not toxic to humans.

Farmers who plant glyphosate-tolerant crops can treat their fields with glyphosate, even while the GM crop is growing. This approach is more efficient and economical than mechanical weeding and reduces soil damage caused by repeated tillage. It is suggested that there is less environmental

BOX 1

The Tale of GM Salmon—Downstream Effects?

It took 18 years and about $60 million, but the first GM animal to be approved as human food—the AquAdvantage salmon—may soon hit the U.S. market.

The AquAdvantage salmon is an Atlantic salmon that is genetically modified to grow twice as fast as its non-GM cousins, reaching marketable size in one and a half years rather than the usual three years. Scientists at AquaBounty Technologies in Massachusetts created the variety by transforming an Atlantic salmon with a single gene encoding the Chinook salmon growth hormone. The gene was cloned downstream of the antifreeze protein gene promoter from an eel. This promoter stimulates growth hormone synthesis in the winter, a time when the fish's own growth hormone gene is not expressed. The rapid growth of the GM salmon allows fish farmers to double their productivity.

AquaBounty intends to sell GM fish eggs to two facilities—one in Canada and one in Panama—that will raise the salmon and market them. To ensure that the fish will not escape the facilities, the company promises to sell only fertilized eggs that are female, triploid, and sterile. The facilities are to be approved only if the tanks are located inland and have sufficient filters to ensure that eggs and small fish cannot escape.

Despite these assurances, environmental groups are planning to fight the sale of GM salmon. Some grocery chains in the United States have banned GM fish, and legislators in several western U.S. states are trying to block the approval of the AquAdvantage salmon based on fears that the accidental release of these fish could contaminate wild salmon populations with transgenes and disturb normal ecosystems.

Supporters of GM fish point out that the GM salmon are very unlikely to escape their facilities, and if any did escape, they would be poorly adapted to wild conditions. Critics of the new GM salmon point out that the technique used to create sterile triploids (pressure-shocking the fertilized eggs) still allows a small percentage of fertile diploids to remain in the stock. They state that even a few fertile fish, if they escaped into the wild, could have long-term effects on wild populations. A study published in 2013 shows that it is possible for the AquAdvantage salmon to breed successfully with a close relative, the brown trout.* In laboratory conditions, the hybrids grew more quickly than either the GM or non-GM varieties, and in closed stream-like systems, the hybrids outcompeted both parental fish varieties for food supplies. The authors point out that these results should be taken into account during environmental assessments, although it is still not known whether the hybrid salmon—trout variety could successfully breed in the wild. If GM salmon could escape, breed, and introduce transgenes into wild populations, there could be unknown negative downstream effects on fish ecosystems.

* Oke, K.B., et al. 2013. Hybridization between genetically modified Atlantic salmon and wild brown trout reveals novel ecological interactions. *Proc. R. Soc. B*. 280 (1763): 20131047.

The AquAdvantage salmon grows twice as fast as a non-GM Atlantic salmon, reaching market size in half the time.

impact when using glyphosate, compared with having to apply higher levels of other, more toxic, herbicides.

Recently, evidence suggests that some weeds may be developing resistance to glyphosate, thereby reducing the effectiveness of glyphosate-tolerant crops. (This and other concerns about herbicide-tolerant GM plants are described later in this chapter.) One method used to engineer a glyphosate-tolerant plant is described in the next section.

Insect-Resistant GM Crops

The second most prevalent GM modifications are those that make plants resistant to agricultural pests. Insect damage is one of the most serious threats to worldwide food production. Farmers combat insect pests using crop rotation and predatory organisms, as well as applying insecticides.

The most widely used GM insect-resistant crops are the **Bt crops.** ***Bacillus thuringiensis*** (Bt) is group of soil-dwelling bacterial strains that produce crystal (Cry) proteins that are toxic to certain species of insects. These Cry proteins are encoded by the bacterial *cry* genes and form crystal structures during sporulation. The Cry proteins are toxic to Lepidoptera (moths and butterflies), Diptera (mosquitoes and flies), Coleoptera (beetles), and Hymenoptera (wasps and ants). Insects must ingest the bacterial spores or Cry proteins in order for the toxins to act. Within the high pH of the insect gut, the crystals dissolve and are cleaved by insect protease enzymes. The Cry proteins bind

to receptors on the gut wall, leading to breakdown of the gut membranes and death of the insect.

Each insect species has specific types of gut receptors that will match only a few types of Bt Cry toxins. As there are more than 200 different Cry proteins, it is possible to select a Bt strain that will be specific to one pest type.

Bt spores have been used for decades as insecticides in both conventional and organic gardening, usually applied in liquid sprays. Sunlight and soil rapidly break down the Bt insecticides, which have not shown any adverse effects on groundwater, mammals, fish, or birds. Toxicity tests on humans and animals have shown that Bt causes few negative effects.

To create Bt crops, scientists introduce one or more cloned *cry* genes into plant cells using methods described in the next section. The GM crop plants will then manufacture their own Bt Cry proteins, which will kill the target pest species when it eats the plant tissues.

Although Bt crops have been successful in reducing crop damage, increasing yields, and reducing the amounts of insecticidal sprays used in agriculture, they are also controversial. Early studies suggested that Bt crops harmed Monarch butterfly populations, although more recent studies have drawn opposite conclusions. Other concerns still exist, and these will be discussed in subsequent sections of this chapter.

BOX 2

The Success of Hawaiian GM Papaya

In the mid-1990s, the papaya ringspot virus (PRSV) spread rapidly throughout Hawaii's papaya fields and threatened to destroy the industry within a few years. To try to stop the destruction of Hawaiian papaya, a team of scientists from the University of Hawaii, the USDA Agricultural Research Center in Hawaii, and the Upjohn Company cloned the coat protein gene of PRSV and introduced it into cultured papaya cells using biolistic transformation. The goal was to create PRSV resistance using a mechanism known as pathogen-derived resistance. The presence of virus coat proteins within the plant is thought to interfere with the disassembly and movement of an infecting virus, slowing or preventing infection. Researchers tested resistance to PRSV in the transformed papaya plants and developed two GM varieties—SunUp and Rainbow. SunUp was homozygous for the PRSV coat protein gene, and Rainbow was an F_1 hybrid of SunUp and a non-GM variety Kapoho. After three years of field testing and two years of moving through federal regulatory processes, GM papaya was approved for use. Seeds were given for free to farmers, who immediately planted them to replace their virus-devastated fields. Within three years, papaya harvests in Hawaii doubled and consumer acceptance was positive. Virus-resistant GM papaya is credited with saving the Hawaiian papaya industry.

An interesting side-effect of the presence of GM papaya in Hawaii was the recovery of non-GM and organically grown papaya. Because PRSV levels declined due to the presence of virus-resistant fields and the abandoning of infected fields, some growers can now produce non-GM papaya, albeit on a smaller scale than before the virus spread throughout Hawaii. At the present time, more than 70 percent of Hawaiian papaya is genetically modified. GM papaya is approved for sale in the United States, Canada, and Japan.

Since the development of GM papaya in Hawaii, efforts to develop similar varieties in other parts of the world have stalled because of increasing public resistance to GM foods.

Since 2010, thousands of GM papaya trees in Hawaii have been cut down and destroyed by anonymous attackers. Efforts to introduce GM papaya in Thailand have failed, and the government recently banned GM foods. Japan has approved the sale of GM papaya, but only if it is labeled as genetically modified.

A papaya fruit, grown on a non-GM papaya plant infected with PRSV.

GM Crops for Direct Consumption

To date, most GM crops have been designed to help farmers increase yields. Also, most GM food crops are not consumed directly by humans, but are used as animal feed or as sources of processed food ingredients such as oils, starches, syrups, and sugars. For example, 98 percent of the U.S. soybean crop is used as livestock feed. The remainder is processed into a variety of food ingredients, such as lecithin, textured soy proteins, soybean oil, and soy flours. However, a few GM foods have been developed for direct consumption. Examples are rice, squash, and papaya (Box 2).

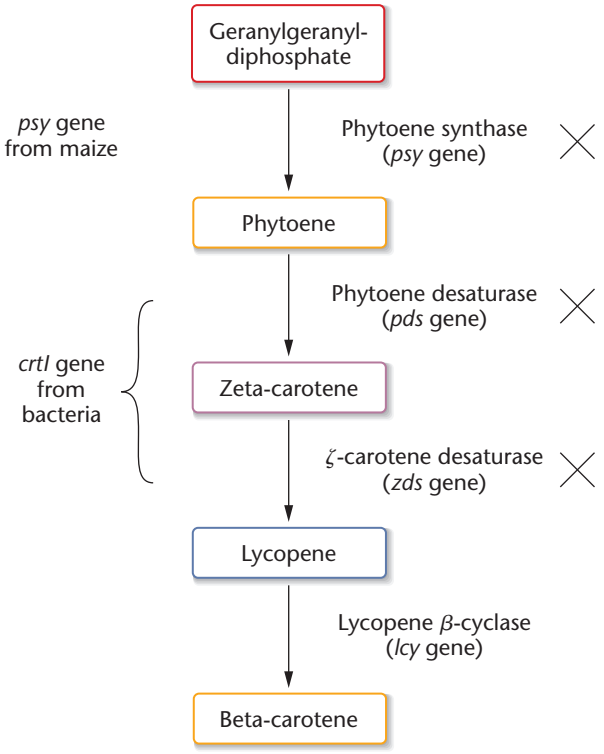
One of the most famous and controversial examples of GM foods is **Golden Rice**—a rice variety designed to synthesize beta-carotene (the precursor to **vitamin A**) in the rice grain endosperm.

Vitamin A deficiency is a serious health problem in more than 60 countries, particularly countries in Asia and Africa. The World Health Organization estimates that 190 million children and 19 million pregnant women are vitamin A deficient. Between 250,000 and 500,000 children with vitamin A deficiencies become blind each year, and half of these will die within a year of losing their sight. As vitamin A is also necessary for immune system function, deficiencies lead to increases in many other conditions, including diarrhea and virus infections. The most seriously affected people live in the poorest countries and have a basic starch-centered diet, often mainly rice. Vitamin A is normally found in dairy products and can be synthesized in the body from beta-carotene found in orange-colored fruits and vegetables and in green leafy vegetables.

Several approaches are being taken to alleviate the vitamin A deficiency status of people in developing countries. These include supplying high-dose vitamin A supplements and growing fresh fruits and vegetables in home gardens. These initiatives have had partial success, but the expense of delivering education and supplementation has impeded the effectiveness of these programs.

In the 1990s, scientists began to apply recombinant DNA technology to help solve vitamin A deficiencies in people with rice-based diets. Although the rice plant naturally produces beta-carotene in its leaves, it does not produce it in the rice grain endosperm, which is the edible part of the rice. The beta-carotene precursor, geranylgeranyl-diphosphate, is present in the endosperm, but the enzymes that convert it to beta-carotene are not synthesized (**ST Figure 5–2**).

In the first version of Golden Rice, scientists introduced the genes *phytoene synthase* (*psy*) cloned from the daffodil plant and *carotene desaturase* (*crtI*) cloned from the bacterium *Erwinia uredovora* into rice plants. The bacterial *crtI* gene was chosen because the enzyme encoded by this gene can perform the functions of two of the missing rice enzymes, thereby simplifying the transformation process.



ST FIGURE 5–2 Beta-carotene pathway in Golden Rice 2. Rice plant enzymes and genes involved in beta-carotene synthesis are shown on the right. The enzymes that are not expressed in rice endosperm are indicated with an "X." The genes inserted into Golden Rice 2 are shown on the left.

The resulting plant produced rice grains that were a yellow color due to the presence of beta-carotene (**ST Figure 5–3**). This strain synthesized modest levels of beta-carotene—but only enough to potentially supply 15–20 percent of the

ST FIGURE 5–3 Non-GM and Golden Rice 2. Golden Rice 2 contains high levels of beta-carotene, giving the rice endosperm a yellow color. The intensity of the color reflects the amount of beta-carotene in the endosperm.

recommended daily allowance of vitamin A. In the second version of the GM plant, called Golden Rice 2, the daffodil *psy* gene was replaced with the *psy* gene from maize. Golden Rice 2 produced beta-carotene levels that were more than 20-fold greater than those in Golden Rice. In the next section we describe the methods used to create Golden Rice 2.

Clinical trials have shown that the beta-carotene in Golden Rice 2 is efficiently converted into vitamin A in humans and that about 150 grams of uncooked Golden Rice 2 (which is close to the normal daily rice consumption of children aged 4–8 years) would supply all of the childhood daily requirement for vitamin A.

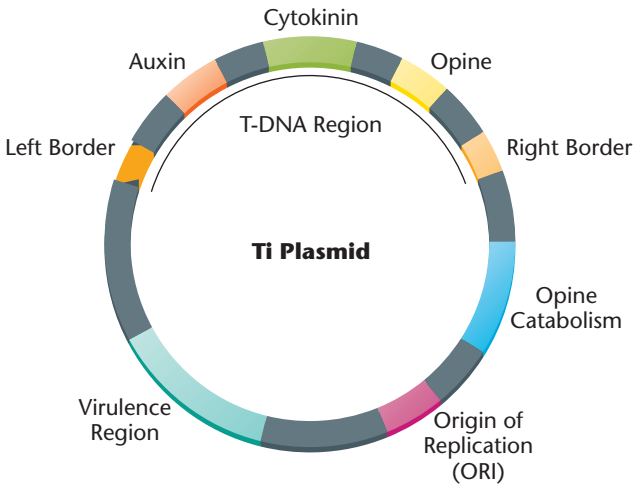
At the present time, Golden Rice 2 is undergoing field, biosafety, and efficacy testing in preparation for approval by government regulators in Bangladesh and the Philippines. If Golden Rice 2 proves useful in alleviating vitamin A deficiencies and is approved for use, seed will be made available at the same price as non-GM seed and farmers will be allowed to keep and replant seed from their own crops.

Despite the promise of Golden Rice 2, controversies remain. Critics of GM foods suggest that Golden Rice could make farmers too dependent on one type of food or might have long-term health or environmental effects. These and other controversies surrounding GM foods are discussed in subsequent sections of this chapter.

Methods Used to Create GM Plants

Most GM plants are created using one of two approaches: the **biolistic method** or ***Agrobacterium tumefaciens*-mediated transformation** technology. Both methods target plant cells that are growing *in vitro*. Scientists can generate plant tissue cultures from various types of plant tissues, and these cultured cells will grow either in liquid cultures or on the surface of solid growth media. When grown in the presence of specific nutrients and hormones, these cultured cells will form clumps of cells called calluses, which, when transferred to other types of media, will form roots. When the rooted plantlets are mature, they are transferred to soil medium in greenhouses where they develop into normal plants.

The *biolistic method* is a physical method of introducing DNA into cells. Particles of heavy metals such as gold are coated with the DNA that will transform the cells; these particles are then fired at high speed into plant cells *in vitro,* using a device called a **gene gun.** Cells that survive the bombardment may take up the DNA-coated particles, and the DNA may migrate into the cell nucleus and integrate into a plant chromosome. Plants that grow from the bombarded cells are then selected for the desired phenotype.



ST FIGURE 5–4 Structure of the Ti plasmid. The 250-kb Ti plasmid from *Agrobacterium tumefaciens* inserts the T-DNA portion of the plasmid into the host cell's nuclear genome and induces tumors. Genes within the virulence region code for enzymes responsible for transfer of T-DNA into the plant genome. The T-DNA region contains auxin and cytokinin genes that encode hormones responsible for cell growth and tumor formation. The opine genes encode compounds used as energy sources for the bacterium. The T-DNA region of the Ti plasmid is replaced with the gene of interest when the plasmid is used as a transformation vector.

Although biolistic methods are successful for a wide range of plant types, a much improved transformation rate is achieved using *Agrobacterium-mediated technology*. *Agrobacterium tumefaciens* (also called *Rhizobium radiobacter*) is a soil microbe that can infect plant cells and cause tumors. These characteristics are conferred by a 200-kb tumor-inducing plasmid called a **Ti plasmid.** After infection with *Agrobacterium*, the Ti plasmid integrates a segment of its DNA known as transfer DNA (T-DNA) into random locations within the plant genome (**ST Figure 5–4**). To use the Ti plasmid as a transformation vector, scientists remove the T-DNA segment and replace it with cloned DNA of the genes to be introduced into the plant cells.

In order to have the newly introduced gene expressed in the plant, the gene must be cloned next to an appropriate promoter sequence that will direct transcription in the required plant tissue. For example, the beta-carotene pathway genes introduced into Golden Rice were cloned next to a promoter that directs transcription of the genes in the rice endosperm. In addition, the transformed gene requires appropriate transcription termination signals and signal sequences that allow insertion of the encoded protein into the correct cell compartment.

Selectable Markers

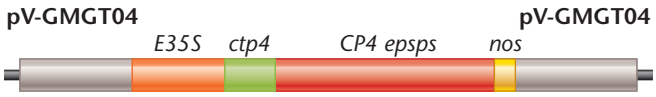
The rates at which T-DNA successfully integrates into the plant genome and becomes appropriately expressed are

low. Often, only one cell in 1000 or more will be successfully transformed. Before growing cultured plant cells into mature plants to test their phenotypes, it is important to eliminate the background of nontransformed cells. This can be done using either positive or negative selection techniques.

An example of negative selection involves use of a **marker gene** such as the hygromycin-resistance gene. This gene, together with an appropriate promoter, can be introduced into plant cells along with the gene of interest. The cells are then incubated in culture medium containing hygromycin—an antibiotic that also inhibits the growth of eukaryotic cells. Only cells that express the hygromycin-resistance gene will survive. It is then necessary to verify that the resistant cells also express the cotransformed gene. This is often done by techniques such as PCR amplification using gene-specific primers. Plants that express the gene of interest are then tested for other characteristics, including the phenotype conferred by the introduced gene of interest.

An example of positive selection involves the use of a selectable marker gene such as that encoding **phosphomannose isomerase** (**PMI**). This enzyme is common in animals but is not found in most plants. It catalyzes the interconversion of mannose 6-phosphate and fructose 6-phosphate. Plant cells that express the *pmi* gene can survive on synthetic culture medium that contains only mannose as a carbon source. Cells that are cotransformed with the *pmi* gene under control of an appropriate promoter and the gene of interest can be positively selected by growing the plant cells on a mannose-containing medium. This type of positive selection was used to create Golden Rice 2. Studies have shown that purified PMI protein is easily digested, nonallergenic, and nontoxic in mouse oral toxicity tests. A variation in positive selection involves use of a marker gene whose expression results in a visible phenotype, such as deposition of a colored pigment.

The following descriptions illustrate the methods used to engineer two GM crops: Roundup-Ready soybeans and Golden Rice 2.



ST FIGURE 5–5 Portion of plasmid pV-GMGT04 used to create Roundup-Ready soybeans. A 1365-bp fragment encoding the EPSPS enzyme from *Agrobacterium CP4* was cloned downstream from the cauliflower mosaic virus *E35S* promoter and the petunia chloroplast transit peptide signal sequence (*ctp4*). CTP4 signal sequences direct the EPSPS protein into chloroplasts, where aromatic amino acids are synthesized. The *CP4 epsps* coding region was cloned upstream of the nopaline synthase (*nos*) transcription termination and polyadenylation sequences. The *CP4 epsps* sequences encode a 455-amino-acid 46-kDa ESPSP protein.

Roundup-Ready® Soybeans

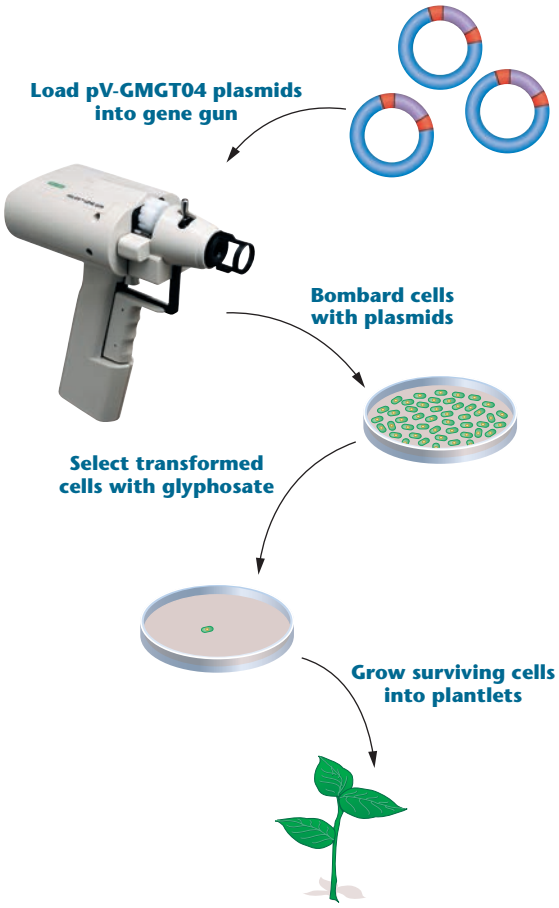
The Roundup-Ready soybean GM variety received market approval in the United States in 1996. It is a GM plant with resistance to the herbicide glyphosate, the active ingredient in Roundup, a commercially available broad-spectrum herbicide. Glyphosate interferes with the enzyme 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS), which is present in all plants and is necessary for plant synthesis of the aromatic amino acids phenylalanine, tyrosine, and tryptophan. EPSPS is not present in mammals, which obtain aromatic amino acids from their diets.

To produce a glyphosate-resistant soybean plant, researchers cloned an *epsps* gene from the *Agrobacterium* strain CP4. This gene encodes an EPSPS enzyme that is resistant to glyphosate. They then cloned the *CP4 epsps* gene downstream of a constitutively expressed promoter from the cauliflower mosaic virus to allow gene expression in all plant tissues. In addition, a short peptide known as a chloroplast transit peptide (in this case from petunias) was cloned onto the 5′-end of the *epsps* gene-coding sequence. This allowed newly synthesized EPSPS protein to be inserted into the soybean chloroplast (**ST Figure 5–5**). The final plasmid contained two *CP4 epsps* genes and, for the initial experiments, a *beta-glucuronidase* (*GUS*) gene from *E. coli*. The *GUS* gene acted as a positive marker, as cells that expressed the plasmid after transformation could be detected by the presence of a blue precipitate. The final cell line chosen for production of Roundup-Ready soybeans did not contain the *GUS* gene.

The plasmids were introduced into cultured soybean cells using biolistic bombardment. Afterward, cells were treated with glyphosate to eliminate any nontransformed cells (**ST Figure 5–6**). The resulting calluses were grown into plants, which were then field tested for glyphosate resistance and a large number of other parameters, including composition, toxicity, and allergenicity.

Golden Rice 2

To create Golden Rice 2, scientists cloned three genes into the T-DNA region of a Ti plasmid. The Ti plasmid, called pSYN12424, is shown in **ST Figure 5–7**. The first gene was the *carotene desaturase* (*crtI*) gene from *Erwinia uredovora,* fused between the rice *glutelin* gene promoter (*Glu*) and the *nos* gene terminator region (*nos*). The *Glu* promoter directs transcription of the fusion gene specifically in the rice endosperm. The *nos* terminator was cloned from the *Agrobacterium tumefaciens nopaline synthase* gene and supplies the transcription termination and polyadenylation sequences



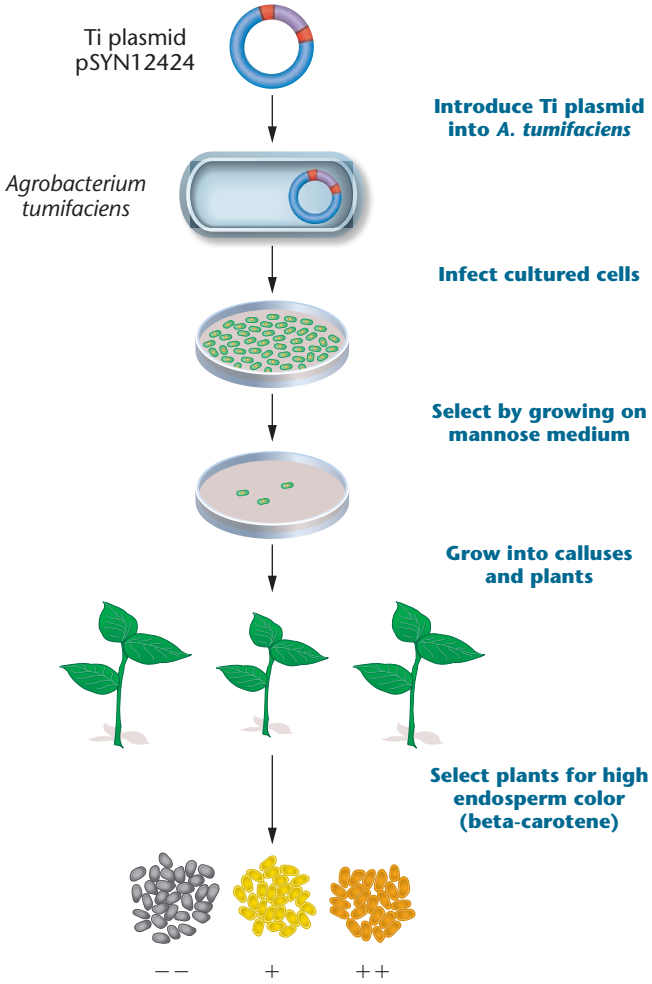
ST FIGURE 5–6 Method for creating Roundup-Ready soybeans. Plasmids were loaded into the gene gun and fired at high pressure into cells growing in tissue cultures. Cells were grown in the presence of glyphosate to select those that had integrated and expressed the *epsps* gene. Surviving cells were stimulated to form calluses and to grow into plantlets.

required at the 3′-end of plant genes. The second gene was the *phytoene synthase* (*psy*) gene cloned from maize. The maize *psy* gene has approximately 90 percent sequence similarity to the rice *psy* gene and is involved in carotenoid synthesis in maize endosperm. This gene was also fused to the *Glu* promoter and the *nos* terminator sequences in order to obtain proper transcription initiation and termination in rice endosperm. The third gene was the selectable marker

ST FIGURE 5–7 T-DNA region of Ti plasmid pSYN12424. The Ti plasmid used to create Golden Rice 2 contained the *carotene desaturase* (*crtI*) gene cloned from bacteria, the *phytoene synthase* (*psy*) gene cloned from maize, and the *phosphomannose isomerase* (*pmi*) gene cloned from *E. coli*. The *glutelin* (*Glu*) gene promoter directs transcription in rice endosperm, and the *polyubiquitin* (*Ubi1*) promoter directs transcription in all tissues. Transcription termination signals were provided by the *nopaline synthase* (*nos*) gene 3′ region.

gene *phosphomannose isomerase* (*pmi*), cloned from *E. coli*. In the Golden Rice 2 Ti plasmid, the *pmi* gene was fused to the maize *polyubiquitin* gene promoter (*Ubi1*) and the *nos* terminator sequences. The *Ubi1* promoter is a constitutive promoter, directing transcription of the *pmi* gene in all plant tissues.

To introduce the pSYN12424 plasmid into rice cells, researchers established embryonic rice cell cultures and infected them with *Agrobacterium tumefaciens* that contained pSYN12424 (**ST Figure 5–8**). The cells were then placed under selection, using culture medium containing only mannose as a carbon source. Surviving cells expressing the *pmi* gene were then stimulated to form calluses that were grown into plants. To confirm that all three genes were present in the transformed rice plants, samples were taken and analyzed by the polymerase chain reaction (PCR) using gene-specific primers. Plants that



ST FIGURE 5–8 Method for creating Golden Rice 2. Rice plant cells were transformed by pSYN12424 and selected on mannose-containing medium, as described in the text. Plants that produced high levels of beta-carotene in rice grain endosperm (++), based on the intensity of the grain's yellow color, were selected for further analysis.

contained one integrated copy of the transgenic construct and synthesized beta-carotene in their seeds were selected for further testing.

GM Foods Controversies

GM foods may be the most contentious of all products of modern biotechnology. Advocates of GM foods state that the technologies have increased farm productivity, reduced pesticide use, preserved soils, and have the potential to feed growing human populations. Critics claim that GM foods are unsafe for both humans and the environment; accordingly, they are applying pressure on regulatory agencies to ban or severely limit the extent of GM food use. These campaigns have affected regulators and politicians, resulting in a patchwork of regulations throughout the world. Often the debates surrounding GM foods are highly polarized and emotional, with both sides in the debate exaggerating their points of view and selectively presenting the data. So, what are the truths behind these controversies?

One point that is important to make as we try to answer this question is that *it is not possible to make general statements about all "GM foods."* Each GM crop or organism contains different genes from different sources, attached to different expression sequences, accompanied by different marker or selection genes, inserted into the genome in different ways and in different locations. GM foods are created for different purposes and are used in ways that are both planned and unplanned. Each construction is unique and therefore needs to be assessed separately.

We will now examine two of the main GM foods controversies: those involving human health and safety, and environmental effects.

Health and Safety

GM food advocates often state that there is no evidence that GM foods currently on the market have any adverse health effects, either from the presence of toxins or from potential allergens. These conclusions are based on two observations. First, humans have consumed several types of GM foods for more than 20 years, and no reliable reports of adverse effects have emerged. Second, the vast majority of toxicity tests in animals, which are required by government regulators prior to approval, have shown no negative effects. A few negative studies have been published, but these have been criticized as poorly executed or nonreproducible.

Critics of GM foods counter the first observation in several ways. First, as described previously, few GM foods are eaten directly by consumers. Instead, most are used as livestock feed, and the remainder form the basis of purified food ingredients. Although no adverse effects of GM foods in livestock have been detected, the processing of many food ingredients removes most, if not all, plant proteins and DNA. Hence, ingestion of GM food-derived ingredients may not be a sufficient test for health and safety. Second, GM foods critics argue that there have been few human clinical trials to directly examine the health effects of most GM foods. One notable exception is Golden Rice 2, which has undergone two small clinical trials. They also say that the toxicity studies that have been completed are performed in animals—primarily rats and mice—and most of these are short-term toxicity studies.

Supporters of GM foods answer these criticisms with several other arguments. The first argument is that short-term toxicity studies in animals are well-established methods for detecting toxins and allergens. The regulatory processes required prior to approval of any GM food demand data from animal toxicity studies. If any negative effects are detected, approval is not given. Supporters also note that several dozen long-term toxicity studies have been published that deal with GM crops such as glyphosate-resistant soybeans and Bt corn, and none of these has shown long-term negative effects on test animals. A few studies that report negative long-term effects have been criticized as poorly designed and unreliable. GM food advocates note that human clinical trials are not required for any other food derived from other genetic modification methods such as selective breeding. During standard breeding of plants and animals, genomes may be mutagenized with radiation or chemicals to enhance the possibilities of obtaining a desired phenotype. This type of manipulation has the potential to introduce mutations into genes other than the ones that are directly selected. Also, plants and animals naturally exchange and shuffle DNA in ways that cannot be anticipated. These include interspecies DNA transfers, transposon integrations, and chromosome modifications. These events may result in unintended changes to the physiology of organisms—changes that could potentially be as great as those arising in GM foods.

Environmental Effects

Critics of GM foods point out that GMOs that are released into the environment have both documented and potential consequences for the environment—and hence may indirectly affect human health and safety. GM food advocates argue that these potential environmental consequences can be identified and managed. Here, we will describe two different aspects of GM foods as they may affect the natural environment and agriculture.

1. Emerging herbicide and insecticide resistance. Many published studies report that the planting of herbicide-tolerant and insect-resistant GM crops has reduced the quantities of herbicides and insecticides that are

ST FIGURE 5–9 Herbicide-resistant weeds. Water hemp weeds, resistant to glyphosate herbicide, growing in a field of Roundup-Ready soybeans.

broadly applied to agricultural crops. As a result, the effects of GM crops on the environment have been assumed to be positive. However, these positive effects may be transient, as herbicide and insecticide resistance is beginning to emerge (**ST Figure 5–9**).

Since glyphosate-tolerant crops were introduced in the mid-1990s, more than 24 glyphosate-resistant weed species have appeared in the United States. Resistant weeds have been found in 18 other countries, and in some cases, the presence of these weeds is affecting crop yields. One reason for the rapid rise of resistant weeds is that farmers have abandoned other weed-management practices in favor of using a single broad-spectrum herbicide. This strong selection pressure has brought the rapid evolution of weed species bearing gene variants that confer herbicide resistance. In response, biotechnology companies are developing new GM crops with tolerance to multiple herbicides. However, scientists argue that weeds will also develop resistance to the use of multiple herbicides, unless farmers vary their weed management practices and incorporate tillage, rotation, and other herbicides along with using the GM crop. Scientists point out that herbicide resistance is not limited to the use of GM crops. Weed populations will evolve resistance to any herbicide used to control them, and the speed of evolution will be affected by the extent to which the herbicide is used.

Since 1996, more than eight different species of insect pests have evolved some level of resistance to Bt insecticidal proteins. For example, in 2011 scientists reported the first cases of resistance of the western corn rootworm to Bt maize expressing the *cry3Bb1* gene, in maize fields in Iowa. In 2010, scientists from Monsanto detected large numbers of pink bollworms with resistance to the toxin expressed from the *cry1Ac* gene in one variety of Bt cotton. In order to slow down the development of Bt resistance, several strategies are being followed. The first is to develop varieties of GM crops that express two Bt toxins simultaneously. Several of these varieties are already on the market and are replacing varieties that express only one Bt *cry* gene. The second strategy involves the use of "refuges" surrounding fields that grow Bt crops. These refuges contain non-GM crops. Insect pests grow easily within the refuges, which place no evolutionary pressure on the insects for resistance to Bt toxins. The idea is for these nonselected insects to mate with any resistant insects that appear in the Bt crop region of the field. The resulting hybrid offspring will be heterozygous for any resistance gene variant. As long as the resistance gene variant is recessive, the hybrids will be killed by eating the Bt crop. In fields that use refuges and plant GM crops containing two Bt genes, resistance to Bt toxins has been delayed or is absent. As with emerging herbicide resistance, farmers are also encouraged to combine the use of Bt crops with conventional pest control methods.

2. The spread of GM crops into non-GM crops. There have been several documented cases of GM crop plants appearing in uncultivated areas in the United States, Canada, Australia, Japan, and Europe. For example, GM sugar beet plants have been found growing in commercial top soils. GM canola plants have been found growing in ditches and along roadways, railway tracks, and in fill soils, far from the fields in which they were grown. A 2011 study[1] found "feral" GM canola plants growing in 288 of 634 sample sites along roadways in North Dakota. Of these plants, 41 percent contained the CP4 EPSPS protein (conferring glyphosate resistance), and 39 percent contained the PAT protein (conferring resistance to the herbicide glufosinate). In addition, two of the plants (0.7 percent of the sample) expressed both proteins (resistant to both herbicides). GM plants that express both proteins have not been created by genetic modification and were assumed to have arisen by cross-fertilization of the other two GM crops. The researchers who conducted this survey were not surprised to find GM canola along transportation routes, as seeds are often spilled during shipping. More surprising was the extent of the distribution and the presence of hybridized GM canola plants.

[1]Schafer, M.G. et al. 2011. *PLoS One* 6:e25736.

One of the major concerns about the escape of GM crop plants from cultivation is the possibility of **outcrossing** or **gene flow**—the transfer of transgenes from GM crops into sexually compatible non-GM crops or wild plants, conferring undesired phenotypes to the other plants. Gene flow between GM crops and adjacent non-GM crops is of particular concern for farmers who want to market their crops as "GM-free" or "organic" and for farmers who grow seed for planting.

Gene flow of GM transgenes has been documented in GM and non-GM canola as well as sugar beets, and in experiments using rice, wheat, and maize. GM critics often refer to controversial studies about GM outcrossing in Oaxaca, Mexico. In the first study in 2001, it was reported that the local maize crops contained transgenes from Monsanto's Roundup-Ready and Bt insect-resistant maize. As GM crops were not approved for use in Mexico, it was thought that the transgenes came from maize that had been imported from the United States as a foodstuff, and then had been planted by farmers who were not aware that the seeds were transgenic. Over the next ten years, subsequent studies reported mixed results. In some studies, the transgenes were not detected, and in others, the same transgenes were detected. There is still no consensus about whether gene flow has occurred between the GM and non-GM maize in Mexico.

It is thought that the presence of glyphosate-resistant transgenes in wild plant populations is not likely to be an environmental risk and would confer no positive fitness benefits to the hybrids. The presence of glyphosate-resistant genes in wild populations would, however, make it more difficult to eradicate the plants. This is illustrated in a case of escaped GM bentgrass in Oregon, where it has been difficult to get rid of the plants because it is no longer possible to use the relatively safe herbicide glyphosate. The potential for environmental damage may be greater if the GM transgenes did confer an advantage—such as insect resistance or tolerance to drought or flooding.

In an attempt to limit the spread of transgenes from GM crops to non-GM crops, regulators are considering a requirement to separate the crops so that pollen would be less likely to travel between them. Each crop plant would require different isolation distances to take into account the dynamics of pollen spreading. Several other methods are being considered. For example, one proposal is to make all GM plants sterile using RNAi technology. Another is to introduce the transgenes into chloroplasts. As chloroplasts are inherited maternally, their genomes would not be transferred via pollen. All of these containment methods are in development stages and may take years to reach the market.

The Future of GM Foods

Over the last 20 years, GM foods have revealed both promise and problems. GM advocates are confident that the next generation of GM foods will show even more promising prospects—and may also address many of the problems.

Research is continuing on ways to fortify staple crops with nutrients to address diet problems in poor countries. For example, Australian scientists are adding genes to bananas that will not only provide resistance to Panama disease—a serious fungal disease that can destroy crops—but also increase the levels of beta-carotene and other nutrients, including iron. Other GM crops in the pipeline include plants engineered to resist drought, high salinity, nitrogen starvation, and low temperatures.

Scientists hope that new genome information and more precise technologies will allow them to accurately edit a plant's endogenous genes—decreasing, increasing, or eliminating expression of one or more of the plant's genes in order to create a desirable phenotype. These approaches avoid the use of transgenes and address some of the concerns about GM foods. The current techniques that researchers use to introduce genes into plant cells result in random insertions into the genome. New techniques are being devised that will allow genes to be inserted into precise locations in the genome, avoiding some of the potential unknown effects of disrupting a plant's normal genome with random integrations.

Researchers are also devising more creative ways to protect plants from insects and diseases. One intriguing project involves introducing into wheat a gene that encodes a pheromone that acts as a chemical alarm signal to aphids. If successful, this approach could protect the wheat plants from aphids without using toxins. Another project involves cassava, which is a staple crop for many Africans and is afflicted by two viral diseases—cassava mosaic virus and brown streak virus—that stunt growth and cause root rot. Although some varieties of cassava are resistant to these viruses, the life cycle of cassava is so long that it would be difficult to introduce resistance into other varieties using conventional breeding techniques. Scientists plan to transform plants with genes from resistant cassava. This type of cisgenic gene transfer is more comparable to traditional breeding than transgenic techniques.

In the future, GM foods will likely include additional GM animals. As described in Box 1, a transgenic Atlantic salmon variety is likely to receive marketing approval in the near future. In another project, scientists have introduced a DNA sequence into chickens that protects the birds from spreading avian influenza. The sequence encodes a hairpin RNA molecule with similarity to a normal viral RNA that binds to the viral polymerase. The presence of the hairpin RNA inhibits

the activity of the viral polymerase and interferes with viral propagation. If this strategy proves useful *in vivo*, the use of these GM chickens would not only reduce the incidence of avian influenza in poultry production, but also reduce the transmissibility of avian influenza viruses to humans.

Although these and other GM foods show promise for increasing agricultural productivity and decreasing disease, the political pressure from anti-GM critics remains a powerful force. An understanding of the science behind these technologies will help us all to evaluate the future of GM foods.

Visit the Study Area in MasteringGenetics for a list of further readings on this topic, including journal references and selected web sites.

Review Questions

1. How do genetically modified organisms compare with organisms created through selective breeding?
2. Can current GM crops be considered as transgenic or cisgenic? Why?
3. Of the approximately 200 GM crop varieties that have been developed, only a few are widely used. What are these varieties, and how prevalent are they?
4. How does glyphosate work, and how has it been used with GM crops to increase agricultural yields?
5. Describe the mechanisms by which the Cry proteins from *Bacillus thuringiensis* act as insecticides.
6. What measures have been taken to alleviate vitamin A deficiencies in developing countries? To date, how successful have these strategies been?
7. What is Golden Rice 2, and how was it created?
8. Describe how plants can be transformed using biolistic methods. How does this method compare with *Agrobacterium tumefaciens*-mediated transformation?
9. How do positive and negative selection techniques contribute to the development of GM crops?
10. Describe how the Roundup-Ready soybean variety was developed, and what genes were used to transform the soybean plants.

Discussion Questions

1. What are the laws regulating the development, approval, and use of GM foods in your region and nationally?
2. Do you think that foods containing GM ingredients should be labeled as such? What would be the advantages and disadvantages to such a strategy?
3. One of the major objections to GM foods is that they may be harmful to human health. Do you agree or disagree, and why?

Gene Therapy

Although drug treatments can be effective in controlling symptoms of genetic disorders, the ideal outcome of medical treatment is to cure a disease. This is the goal of **gene therapy**—the delivery of therapeutic genes into a patient's cells to correct genetic disease conditions caused by a faulty gene or genes. The earliest attempts at gene therapy focused on the delivery of normal, *therapeutic* copies of a gene to be expressed in such a way as to override or negate the effects of the disease gene and thus minimize or eliminate symptoms of the genetic disease. But in recent years newer methods for inhibiting or silencing defective genes, and even approaches for targeted removal of defective genes, have increasingly emerged as potential mechanisms for gene therapy.

Gene therapy is one of the goals of **translational medicine**—taking a scientific discovery, such as the identification of a disease-causing gene, and translating the finding into an effective therapy, thus moving from the laboratory bench to a patient's bedside to treat a disease. In theory, the delivery of a therapeutic gene is rather simple, but in practice, gene therapy has been very difficult to execute. In spite of over 20 years of trials, this field has not lived up to its expectations. However, gene therapy is currently experiencing a fast-paced resurgence of sorts, with several high-profile new successes and potentially exciting new technologies sitting on the horizon. It is hoped that gene therapy will soon become part of mainstream medicine. The treatment of a human genetic disease by gene therapy is the ultimate application of genetic technology. In this Special Topic chapter we will explore how gene therapy is executed, and we will highlight selected examples of successes and failures as well as discuss new approaches to gene therapy. Finally, we will consider ethical issues regarding gene therapy.

"The treatment of a human genetic disease by gene therapy is the ultimate application of genetic technology."

What Genetic Conditions Are Candidates for Treatment by Gene Therapy?

Two essential criteria for gene therapy are that the gene or genes involved in causing a particular disease have been identified and that the gene can be cloned or synthesized in a laboratory. As a result of the Human Genome Project, the identification of human disease genes and their specific DNA sequences has greatly increased the number of candidate genes for gene therapy trials. Almost all of the early gene therapy trials and most gene therapy approaches have focused on treating conditions caused by a single gene. This has been the case because theoretically it is technically easier to affect one gene than disease conditions caused by multiple genes and potentially multiple mutations.

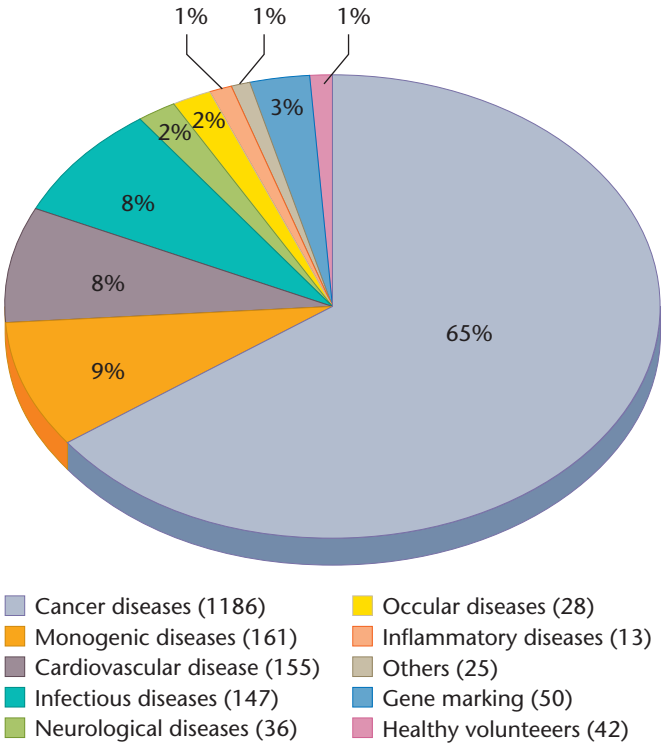
The cells affected by the genetic condition must be readily accessible for treatment by gene therapy. For example, blood disorders such as leukemia, hemophilia, and other conditions have been major targets of gene therapy because it is relatively routine to manipulate blood cells outside of the body and return them to the body in comparison to treating cells in the brain and spinal cord, skeletal or cardiac muscle, and organs with heterogeneous populations of cells such as the pancreas.

In the past decade, every major category of genetic diseases has been targeted by gene therapy (**ST Figure 6–1**). A majority of recently approved clinical trials are for cancer treatment. Gene therapy approaches are currently being investigated for the treatment of hereditary blindness, neurological (neurodegenerative) diseases including Alzheimer disease, Parkinson disease and amyotrophic lateral sclerosis (ALS), cardiovascular disease, muscular dystrophy, hemophilia, a variety of cancers, and infectious diseases, such as HIV, among many other conditions, including depression and drug and alcohol addiction. Over 2000 approved gene therapy clinical trials have occurred or recently been initiated worldwide.

In the United States, proposed gene therapy clinical trials must first be approved by review boards at the institution where they will be carried out, and then the protocols must be approved by the Food and Drug Administration (FDA).

How Are Therapeutic Genes Delivered?

In general, there are two broad approaches for delivering therapeutic genes to a patient being treated by gene therapy, *ex vivo gene therapy* and *in vivo gene therapy*



ST FIGURE 6–1 Graphic representation of different genetic conditions being treated by gene therapy clinical trials worldwide. Notice that cancers are the major target for treatment.

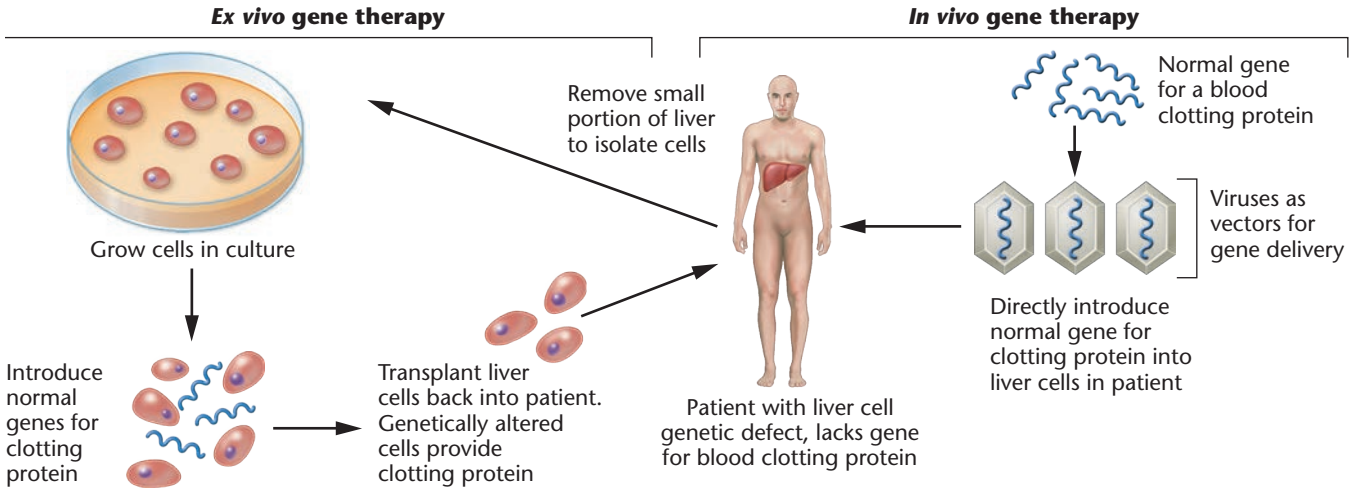
(ST Figure 6–2). In *ex vivo* gene therapy, cells from a person with a particular genetic condition are removed, treated in a laboratory by adding either normal copies of a therapeutic gene or a DNA or RNA sequence that will inhibit expression of a defective gene, and then these cells are transplanted back into the person where the therapeutic gene will express normal copies of the required protein. Genetically altered cells treated in this manner can be transplanted back into the patient without fear of immune system rejection because these cells were derived from the patient initially.

In vivo gene therapy does not involve removal of a person's cells. Instead, therapeutic DNA is introduced directly into affected cells of the body. One of the major challenges of *in vivo* gene therapy is restricting the delivery of therapeutic genes to only the intended tissues and not to all tissues throughout the body.

Viral Vectors for Gene Therapy

For both *in vitro* and *ex vivo* approaches, the key to successful gene therapy is having a delivery system to transfer genes into a patient's cells. Because of the relatively large molecular size and electrically charged properties of DNA, most human cells do not take up DNA easily. Therefore, delivering therapeutic DNA molecules into human cells is challenging. Since the early days of gene therapy, genetically engineered viruses as vectors have been the main tools for delivering therapeutic genes into human cells. Viral vectors for gene therapy are engineered to carry therapeutic DNA as their payload so that the virus infects target cells and delivers the therapeutic DNA without causing damage to cells.

In a majority of gene therapy trials around the world, scientists have used genetically modified *retroviruses* as vectors. Recall from earlier in the text (see Chapter 9) that retroviruses (HIV is a retrovirus) contain an RNA genome that scientists use as a template for the synthesis of a complementary DNA molecule. **Retroviral vectors** are created by removing replication and disease-causing genes from



ST FIGURE 6–2 *Ex vivo* and *in vivo* gene therapy for a patient with a liver disorder. *Ex vivo* gene therapy involves isolating cells from the patient, introducing normal copies of a therapeutic gene (encoding a blood clotting protein in this example) into these cells, and then returning cells to the body where they will produce the required clotting protein. *In vivo* approaches involve introducing DNA directly into cells while they are in the body.

BOX 1

ClinicalTrials.gov

One of the best resources on the Web for learning about ongoing clinical trials, including current gene therapy trials, is ClinicalTrials.gov. The site can easily be searched to find a wealth of resources about ongoing gene therapy trials throughout the United States that are of interest to you. To find a gene therapy clinical trial, use the "Search for Studies" box and type in the name of a disease and "gene therapy." This search string will take you to a page listing active gene therapy clinical trials, with links to detailed information about the trial.

the virus and replacing them with a cloned human gene. After the altered RNA has been packaged into the virus, the recombinant viral vector containing the therapeutic human gene is used to infect a patient's cells. Technically, virus particles are carrying RNA copies of the therapeutic gene. Once inside a cell, the virus cannot replicate itself, but the therapeutic RNA is reverse transcribed into DNA, which enters the nucleus of cells and *integrates* into the genome of the host cells' chromosome. If the inserted therapeutic gene is properly expressed, it produces a normal gene product that may be able to ameliorate the effects of the mutation carried by the affected individual.

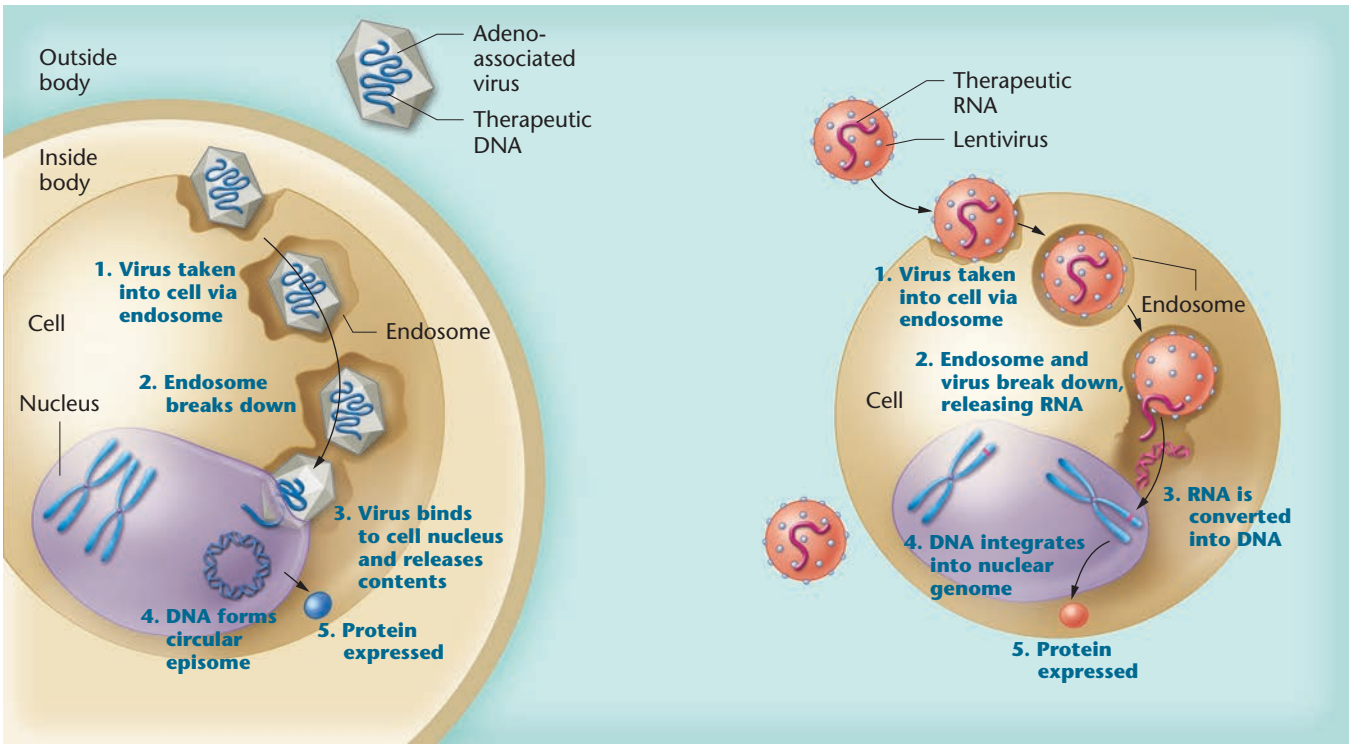
One advantage of retroviral vectors is that they provide long-term expression of delivered genes because they integrate the therapeutic gene into the genome of the patient's cells. But a major problem with retroviral vectors is that they have produced severe toxicity in some cases due to *insertional mutations*. Retroviral vectors generally integrate their genome into the host-cell genome at random sites. Thus, there is the potential for retroviral integration that randomly inactivates genes in the genome or gene-regulatory regions such as a promoter sequence.

Adenovirus vectors were used in many early gene therapy trials. An advantage of these vectors is that they are capable of carrying large therapeutic genes. But because many humans produce antibodies to adenovirus vectors they can mount immune reactions that can render the virus and its therapeutic gene ineffective or cause significant side-effects to the patient. A related virus called **adeno-associated virus (AAV)** is now widely used as a gene therapy vector [**ST Figure 6–3(a)**]. In its native form, AAV infects about 80–90 percent of humans during childhood, causing symptoms associated with the common cold. Disabled forms of AAV are popular for gene therapy because the virus is nonpathogenic, so it usually does not elicit a major response from the immune system of treated patients. AAV also does not typically integrate into the host-cell genome, so there is little risk of the insertional mutations that have plagued retroviruses, although modified forms of AAV have been used to deliver genes to specific sites on individual chromosomes. Most forms of AAV deliver genes into the host-cell nucleus where it forms small hoops of DNA called *episomes* that are expressed under the control of promoter sequences contained within the viral genome. But because therapeutic DNA delivered by AAV does not usually become incorporated into the genome, it is not replicated when host cells divide, and so the gene therapy approach may require repeated, ongoing applications to be successful [ST Figure 6–3(a)].

Work with **lentivirus vectors** is an active area of gene therapy research [**ST Figure 6–3(b)**]. Lentivirus is a retrovirus that can accept relatively large pieces of genetic material. Another positive feature of lentivirus is that it is capable of infecting nondividing cells, whereas other viral vectors often infect cells only when they are dividing. It is still not possible to control where lentivirus integration occurs in the host-cell genome, but the virus does not appear to gravitate toward gene-regulatory regions the way that other retroviruses do. Thus the likelihood of causing insertional mutations appears to be much lower than for other vectors.

The human immunodeficiency virus (HIV) responsible for acquired immunodeficiency syndrome (AIDS) is a type of lentivirus. It may surprise you that HIV could be used as a vector for gene therapy. For any viral vector, scientists must be sure that the vector has been genetically engineered to render it inactive so that the virus cannot produce disease or spread throughout the body and infect other tissues. In the case of HIV, modified forms of HIV, strains lacking the genes necessary for reconstitution of fully functional viral particles, are being used for gene therapy trials. HIV has evolved to infect certain types of T lymphocytes (T cells) and macrophages, making it a good vector for delivering therapeutic genes into the bloodstream.

Increasingly, viral vectors and nonviral vector approaches are being used to deliver therapeutic genes into *stem cells*, usually *in vitro*, and then the stem cells are either reintroduced into the patient or differentiated *in vitro* into mature cell types before being transplanted into the correct organ of a patient being treated.



ST FIGURE 6–3 Delivering therapeutic genes. (a) Nonintegrating viruses such as modified adeno-associated virus (AAV) deliver therapeutic genes without integrating them into the genome of target cells. Delivered DNA resides as minichromosomes (episomes), but over time as cells divide, these nonintegrating hoops of DNA are gradually lost. (b) Integrating viruses include lentivirus, an RNA retrovirus that delivers therapeutic genes into the cytoplasm where reverse transcriptase converts RNA into DNA. DNA then integrates into the genome, ensuring that therapeutic DNA will be passed into daughter cells during cell division.

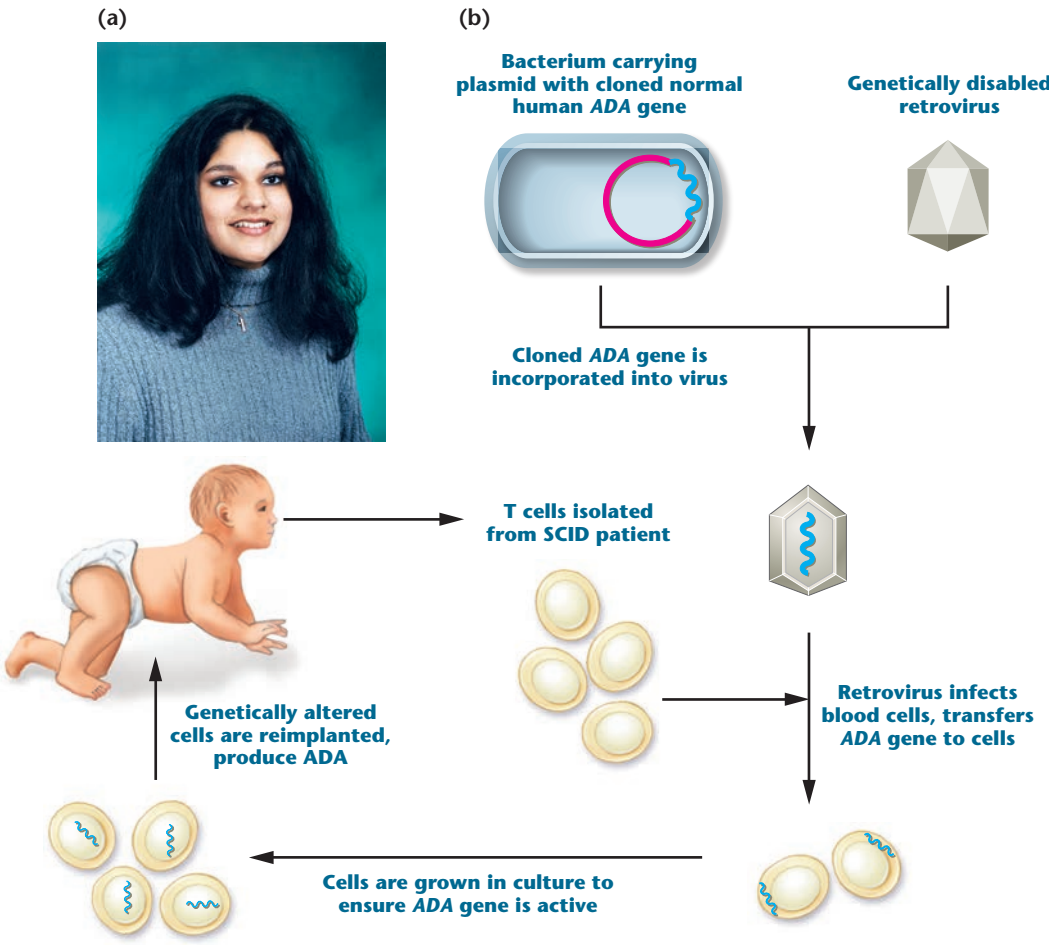
Nonviral Delivery Methods

Scientists continue to experiment with various *in vivo* and *ex vivo* strategies for trying to deliver so called naked DNA into cells without the use of viral vectors. Nonviral methods that are being used to transfer genes into cells include chemically assisted transfer of genes across cell membranes, nanoparticle delivery of therapeutic genes, and fusion of cells with artificial lipid vesicles called *liposomes* that contain cloned DNA sequences. Short-term expression of genes through "gene pills" is being explored. In this concept, a pill delivers therapeutic DNA to the intestines where the DNA is absorbed by intestinal cells that then express the therapeutic protein and secrete the protein into the bloodstream.

The First Successful Gene Therapy Trial

In 1990 the FDA approved the first human gene therapy trial, which began with the treatment of a young girl named Ashanti DeSilva [**ST Figure 6–4(a)**], who has a heritable disorder called **severe combined immunodeficiency** (**SCID**). Individuals with SCID have no functional immune system and usually die from what would normally be minor infections. Ashanti has an autosomal form of SCID caused by a mutation in the gene encoding the enzyme *adenosine deaminase (ADA)*. Her gene therapy began when clinicians isolated some of her white blood cells, called T cells [**ST Figure 6–4(b)**]. These cells, which are key components of the immune system, were mixed with a retroviral vector carrying an inserted copy of the normal *ADA* gene. The virus infected many of the T cells, and a normal copy of the *ADA* gene was inserted into the genome of some T cells.

After being mixed with the vector, the T cells were grown in the laboratory and analyzed to make sure that the transferred *ADA* gene was expressed (ST Figure 6–4). Then a billion or so genetically altered T cells were injected into Ashanti's bloodstream. Repeated treatments were required to produce a sufficient number of functioning T cells. In addition, Ashanti also periodically received injections of purified ADA protein throughout this process so the exact effects of gene therapy were difficult to discern. Ashanti continues to receive supplements of the ADA enzyme to allow her to lead a normal life.



ST FIGURE 6–4 The first successful gene therapy trial. (a) Ashanti DeSilva, the first person to be successfully treated by gene therapy. (b) To treat SCID using gene therapy, a cloned human ADA gene is transferred into a viral vector, which is then used to infect white blood cells removed from the patient. The transferred ADA gene is incorporated into a chromosome and becomes active. After growth to enhance their numbers, the cells are inserted back into the patient, where they produce ADA, allowing the development of an immune response.

Subsequent gene therapy treatments for SCID have focused on using bone marrow stem cells and *in vitro* approaches to repopulate the number of ADA-producing T cells. To date, gene therapy has restored the health of about 20 children affected by SCID. SCID treatment is still considered the most successful example of gene therapy.

Gene Therapy Setbacks

From 1990 to 1999, more than 4000 people underwent gene therapy for a variety of genetic disorders. These trials often failed and thus led to a loss of confidence in gene therapy. In the United States, gene therapy plummeted even further in 1999 when teenager Jesse Gelsinger died while undergoing a test for the safety of gene therapy to treat a liver disease called ornithine transcarbamylase (OTC) deficiency. Large numbers of adenovirus vectors bearing the *OTC* gene were injected into his hepatic artery. The vectors were expected to target his liver, enter liver cells, and trigger the production of OTC protein. In turn, it was hoped that the OTC protein might correct his genetic defect and cure him of his liver disease.

Researchers had previously treated 17 people with the therapeutic virus, and early results from these patients were promising. But as the 18th patient, Jesse Gelsinger, within hours of his first treatment, developed a massive immune reaction. He developed a high fever, his lungs filled with fluid, multiple organs shut down, and he died four days later of acute respiratory failure. Jesse's severe response to the adenovirus may have resulted from how his body reacted to a previous exposure to the virus used as the vector for this protocol.

In the aftermath of the tragedy, several government and scientific inquiries were conducted. Investigators

learned that in the clinical trial scientists had not reported other adverse reactions to gene therapy and that some of the scientists were affiliated with private companies that could benefit financially from the trials. It was determined that serious side-effects seen in animal studies were not explained to patients during informed-consent discussions. The FDA subsequently scrutinized gene therapy trials across the country, halted a number of them, and shut down several gene therapy programs. Other groups voluntarily suspended their gene therapy studies. Tighter restrictions on clinical trial protocols were imposed to correct some of the procedural problems that emerged from the Gelsinger case. Jesse's death had dealt a severe blow to the struggling field of gene therapy—a blow from which it was still reeling when a second tragedy hit.

The outlook for gene therapy brightened momentarily in 2000, when a group of French researchers reported what was hailed as the first large-scale success in gene therapy. Children with a fatal X-linked form of SCID (X-SCID, also known as "bubble boy" disease) developed functional immune systems after being treated with a retroviral vector carrying a normal gene. But elation over this study soon turned to despair, when it became clear that 5 of the 20 patients in the trial developed leukemia as a direct result of their therapy. One of these patients died as a result of the treatment, while the other four went into remission from the leukemia. In two of the children examined, their cancer cells contained the retroviral vector, inserted near or into a gene called *LMO2*. This *insertional mutation* activated the *LMO2* gene, causing uncontrolled white blood cell proliferation and development of leukemia. The FDA immediately halted 27 similar gene therapy clinical trials, and once again gene therapy underwent a profound reassessment.

On a positive note, long-term survival data from trials in the UK to treat X-SCID and SCID using hematopoietic stem cells from the patients' bone marrow for gene therapy have shown that 14 of 16 children have had their immune system restored at least 9 years after the treatment. These children formerly had life expectancies of less than 20 years. Nevertheless, the above events had major negative impacts on the progress of gene therapy.

Problems with Gene Therapy Vectors

Critics of gene therapy have berated research groups for undue haste, conflicts of interest, sloppy clinical trial management, and for promising much but delivering little. Most of the problems associated with gene therapy, including the Jesse Gelsinger case and the French X-SCID trial, have been traced to the viral vectors used to transfer therapeutic genes into cells. These vectors have been shown to have several serious drawbacks.

- First, integration of retroviral genomes, including the human therapeutic gene into the host cell's genome, occurs only if the host cells are replicating their DNA. In the body, only a small number of cells in any tissue are dividing and replicating their DNA.
- Second, the injection of large amounts of most viral vectors, but particularly adenovirus vectors, is capable of causing an adverse immune response in the patient, as happened in Jesse Gelsinger's case.
- Third, insertion of viral genomes into host chromosomes can activate or mutate an essential gene, as in the case of the French patients. Viral integrase, the enzyme that allows for viral genome integration into the host genome, interacts with chromatin-associated proteins, often steering integration toward transcriptionally active genes.
- Fourth, AAV vectors cannot carry DNA sequences larger than about 5 kb, and retroviruses cannot carry DNA sequences much larger than 10 kb. Many human genes exceed the 5—10 kb size range.
- Finally, there is a possibility that a fully infectious virus could be created if the inactivated vector were to recombine with another unaltered viral genome already present in the host cell.

To overcome these problems, new viral vectors and strategies for transferring genes into cells are being developed in an attempt to improve the action and safety of vectors. Fortunately, gene therapy has experienced resurgence in part because of several promising new trials and successful treatments.

Recent Successful Trials

Treating Retinal Blindness

In recent years, patients being treated for blindness have greatly benefited from gene therapy approaches. Congenital retinal blinding conditions affect about 1 in 2000 people worldwide, many of which are the result of a wide range of genetic defects. Over 165 different genes have been implicated in various forms of retinal blindness.

Successful gene therapy has been achieved in subsets of patients with *Leber congenital amaurosis (LCA)*, a degenerative disease of the retina that affects 1 in 50,000 to 1 in 100,000 infants each year and causes severe blindness.

Gene therapy treatments for LCA were originally pioneered in dogs. Based on the success of these treatments, the protocols were adapted and applied to human gene therapy trials.

LCA is caused by alterations to photoreceptor cells (rods and cones), light-sensitive cells in the retina, due to 18 or more genes. One gene in particular, *RPE65*, has been the gene therapy target of choice. The protein product of the *RPE65* gene metabolizes retinol, which is a form of vitamin A that allows the rod and cone cells of the retina to detect light and transmit electrical signals to the brain. In one of the earliest trials, young adult patients with defects in the *RPE65* gene were given injections of the normal gene. Several months after a single treatment, many adult patients, while still legally blind, could detect light, and some of them could read lines of an eye chart. This treatment approach for LCA was based on injecting AAV-carrying *RPE65* at the back of the eye directly under the retina. The therapeutic gene enters about 15 to 20 percent of cells in the retinal pigment epithelium, the layer of cells just beneath the visual cells of the retina. Adults treated by this approach have shown substantial improvements in a variety of visual functions tests, but the greatest improvement has been demonstrated in children, all of whom have gained sufficient vision to allow them to be ambulatory. Researchers think the success in children has occurred because younger patients have not lost as many photoreceptor cells as older patients.

Over two dozen gene therapy trials have been completed or are ongoing for various forms of blindness, including age-related degenerative causes of blindness. Because of the small size of the eye and the relatively small number of cells that need to be treated, the prospects for gene therapy to become routine treatment for eye disorders appears to be very good. Retinal cells are also very long-lived; thus, AAV delivery approaches can be successful for long periods of time even if the gene does not integrate.

Successful Treatment of Hemophilia B

A very encouraging gene therapy trial in England successfully treated a small group of adults with hemophilia B, a blood disorder caused by a deficiency in the coagulation protein human factor IX. Currently, hemophilia B patients are treated several times each week with infusions of concentrated doses of the factor IX protein. In the gene therapy trial, six adult patients received, *in vivo*, a single dose of an adenovirus vector (AAV8) carrying normal copies of the human factor IX gene introduced into liver cells. Of six patients treated, four were able to stop factor IX infusion treatments after the gene therapy trial. Several other trials of this AAV treatment approach are underway, and expectations are high that a gene therapy cure for hemophilia B is close to becoming a routine reality.

HIV as a Vector Shows Promise in Recent Trials

Researchers at the University of Paris and Harvard Medical School reported that two years after gene therapy treatment for ***β*-thalassemia,** a blood disorder involving the *β*-globin gene that reduces the production of hemoglobin, a young man no longer needed transfusions and appeared to be healthy. A modified, disabled HIV was used to carry a copy of the normal ***β-globin*** gene. Although this trial resulted in activation of the growth factor gene called *HMGA2*, reminiscent of what occurred in the French X-SCID trials, activation of the transcription factor did not result in an overproduction of hematopoietic cells or create a condition of preleukemia.

In 2013, researchers at the San Raffaele Telethon Institute for Gene Therapy in Milan, Italy, reported two studies using lentivirus vectors derived from HIV in combination with hematopoietic stem cells (HSCs) to successfully treat children with either **metachromatic leukodystrophy** (**MLD**) or **Wiskott-Aldrich syndrome** (**WAS**). MLD is a neurodegenerative disorder affecting storage of enzymes in lysosomes and is caused by mutation in the arylsulfatase A (*ARSA*) gene that results in an accumulation of fats called sulfatides. These are toxic to neurons, causing progressive loss of the myelin sheath (demyelination) surrounding neurons in the brain, leading to a loss of cognitive functions and motor skills. There is no cure for MLD. Children with MLD appear healthy at birth but eventually develop MLD symptoms.

In this trial, researchers used an *ex vivo* approach with a lentivirus vector to introduce a functional *ARSA* gene into bone marrow-derived HSCs from each patient and then infused treated HSCs back into patients. Three years after the start of a trial involving a total of 16 patients, 10 patients with MLD and 6 with WAS, data from six patients analyzed 18 to 24 months after gene therapy indicated that the trials are safe and effective. These initial reports are based on studying three children from each study because these are the first patients for whom sufficient time has passed after gene therapy treatment to make significant conclusions regarding the safety and effectiveness of the trials. It took over 15 years of research to get to this point. These trials involved a team of over 70 people, including researchers and clinicians, which is indicative of the teamwork approach typical of gene therapy trials.

In three children with MLD, gene therapy treatment halted disease progression for 18 to 24 months after therapy, as determined by magnetic resonance images of

BOX 2

Glybera Is the First Commercial Gene Therapy to Be Approved in the West

In late 2012, a gene therapy product called Glybera (alipogene tiparvovec) made history when the European Medicines Agency of the European Union approved it as the first gene therapy trial to win commercial approval in the Western world. Glybera is an AAV vector system for delivering therapeutic copies of the *LPL* gene to treat patients with a rare disease called *lipoprotein lipase deficiency* (LPLD, also called familial hyperchylomicronemia). LPLD patients have high levels of triglycerides in their blood. Elevated serum triglycerides are toxic to the pancreas and cause a severe form of pancreatic inflammation called pancreatitis. Developed by Amsterdam-based company uniQure BV, it is still unclear if Glybera will be approved by the U.S. FDA. Nonetheless, the success of Glybera trials in Europe signals what many gene therapy researchers hope will be the beginning of a wave of approvals for gene therapy treatments in Europe and the United States.

the brain and through tests of cognitive and motor skills. Because disease onset is predicted at 7 to 21 months, scientists are very encouraged by the outcomes of this trial. The trial was technically complicated because it required that HSCs travel through the bloodstream and release the ARSA protein that is taken up into neurons. A major challenge was to create enough engineered cells to produce a sufficient quantity of therapeutic ARSA protein to counteract the neurodegenerative process.

Similar results were reported for treating patients with WAS, an X-linked condition resulting in defective platelets that make patients more vulnerable to infections, frequent bleeding, autoimmune diseases, and cancer. Genome sequencing of MLD and WAS patients treated in these trials showed no evidence of genome integration near oncogenes. Similarly, patients showed no evidence of hematopoietic stem cell overproduction, suggesting that this lentivirus delivery protocol produced a safe and stable delivery of the therapeutic genes.

Targeted Approaches to Gene Therapy

The gene therapy approaches and examples we have highlighted thus far have focused on the addition of a therapeutic gene that functions along with the defective gene. However, the removal, correction, and/or replacement of a mutated gene and silencing expression of a defective gene are two other approaches being developed. Rapid progress is being made with these approaches.

DNA-Editing Nucleases for Gene Targeting

For nearly 20 years, scientists have been working on modifications of restriction enzymes and other nucleases to engineer proteins capable of **gene targeting** or **gene editing**—replacing specific genes in the genome. The concept is to combine a nuclease with a sequence-specific DNA binding domain that can be precisely targeted for digestion. In 1996 researchers fused DNA-binding proteins with a zinc-finger motif and DNA cutting domain from the restriction enzyme *Fok*I to create enzymes called **zinc-finger nucleases** (**ZFNs;** ST Figure 6–5). The zinc-finger motif is found in many transcription factors and consists of a cluster of two cysteine and two histidine residues that bind zinc atoms and interact with specific DNA sequences. By coupling zinc-finger motifs to DNA cutting portions of a polypeptide, ZFNs provide a mechanism for modifying sequences in the genome in a sequence-specific *targeted way*.

The DNA-binding domain of the ZFN can be engineered to attach to any sequence in the genome. The zinc fingers bind with a spacing of 5–7 nucleotides, and the nuclease domain of the ZFN cleaves between the binding sites.

Another category of DNA-editing nucleases called **TALENs (transcription activator-like effector nucleases**) was created by adding a DNA-binding motif identified in transcription factors from plant pathogenic bacteria known as transcription activator-like effectors (TALEs) to nucleases to create TALENS. TALENS also cleave as dimers. The DNA-binding domain is a tandem array of amino acid repeats, with each TALEN repeat binding to a specific single base pair. The nuclease domain then cuts the sequence between the dimers, a stretch that spans about 13 bp.

ZFNs and TALENS have shown promise in animal models and cultured cells for gene replacement approaches that involve removing a defective gene from the genome. These enzymes can create site-specific cleavage in the genome. When coupled with certain integrases, ZFNs and TALENs may lead to gene editing by cutting out defective sequences and using recombination to introduce homologous sequences into the genome that replace defective sequences. Although this technology has not yet advanced sufficiently for reliable use in humans, there have been several promising trials.

For example, ZFNs are actively being used in clinical trials for treating patients with HIV. Scientists are

ST FIGURE 6–5 Zinc-finger nucleases and TALENs bind and cut DNA at specific sequences.

exploring ways to deliver immune system-stimulating genes that could make individuals resistant to HIV infection or cripple the virus in HIV-positive persons. In 2007, Timothy Brown, a 40-year-old HIV-positive American, had a relapse of acute myeloid leukemia and received a stem cell (bone-marrow) transplant. Because he was HIV-positive, Brown's physician selected a donor with a mutation in both copies of the *CCR5* gene, which encodes an HIV co-receptor carried on the surface of T cells to which HIV must bind to enter T cells (specifically CD4+ cells). People with naturally occurring mutations in both copies of the *CCR5* gene are resistant to most forms of HIV. Brown relapsed again and received another stem cell transplant from the *CCR5*-mutant donor. Eventually, the cancer was contained, and by 2010, levels of HIV in his body were still undetectable even though he was no longer receiving immune-suppressive treatment. Brown is generally considered to be the first person to have been cured of an HIV infection.

This example encouraged researchers to press forward with a gene therapy approach to modify the *CCR5* gene of HIV patients. In one promising trial, T cells were removed from HIV-positive men, and ZFNs were used to disrupt the *CCR5* gene. The modified cells were then reintroduced into patients. In five of six patients treated, immune-cell counts rose substantially and viral loads also decreased following the therapy. What percentage of immune cells would have to be treated this way to significantly inhibit spread of the virus is not known, but initial results are very promising.

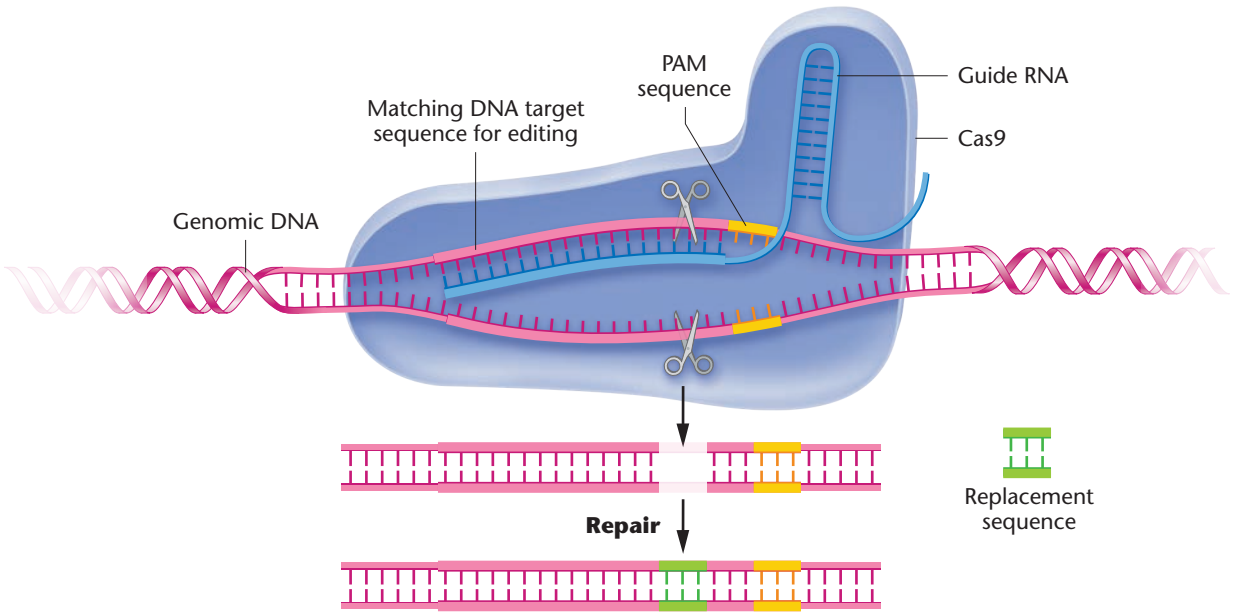
Recently, researchers working with human cells used TALENs to remove defective copies of the *COL7A1* gene, which causes recessive dystrophic epidermolysis bullosa (RDEB), an incurable and often fatal disease that causes excessive blistering of the skin, pain, and severely debilitating skin damage. Researchers at the University of Minnesota used a TALEN to cut DNA near a mutation in *COL7A1* gene in skin cells taken from an RDEB patient. These cells were then converted into a type of stem cell called *induced pluripotent stem cells (iPSCs)*. The iPSCs were treated with therapeutic copies of the *COL7A1* gene and then differentiated into skin cells that expressed the correct protein. This is a promising result, and researchers now plan to transplant these skin cells into patients in an attempt to cure them of RDEB. Another group has recently taken a similar approach using TALENS to repair cultured cells in order to correct the mutation in Duchenne muscular dystrophy (DMD). Researchers are optimistic that this approach can soon be adapted to treat patients.

CRISPR/Cas Method Revolutionizes Gene Editing Applications

No gene targeting method has created more excitement than the gene editing technique known as **CRISPR/Cas** (clustered regularly interspaced short palindrome repeats/CRISPR-associated proteins) or simply **CRISPR.** Identified in bacterial cells, the CRISPR system functions to provide bacteria and archaea immunity against invading bacteriophages and foreign plasmids. First introduced in 2013, a CRISPR craze unfolded that has revolutionized genome-engineering applications including gene editing for gene therapy. Because CRISPR works in bacteria, animal, and plant cells, the method offers diverse applications for genetic engineering by targeted gene editing.

CRISPR is based on delivering a single-stranded "guided" RNA sequence (sgRNA) that is complementary to the target gene sequence in the genome and attached to the endonuclease called Cas9 (**ST Figure 6–6**). Compared to TALEN approaches, sgRNAs are relative easy to design and synthesize. At the same time as the sgRNA sequence is delivered, a DNA template strand coding for a replacement sequence is delivered. The sgRNA-Cas9 complex binds to the target DNA sequence, and Cas9 generates a blunt, double-stranded break in the DNA. CRISPR recognition of DNA cleavage sites is determined by RNA-DNA base pairing and a protospacer-adjacent motif (PAM), a three-nucleotide sequence adjacent to the complementary sequence. As cells repair the DNA damage caused by Cas9, repair enzymes incorporate template DNA into the genome at the CRISPR/Cas site, thus replacing the target DNA sequence.

Part of the power of CRISPR is that editing can be done directly in a living, adult animal. Within months of the technique being widely available, researchers around



ST FIGURE 6–6 The CRISPR/Cas system allows for gene editing by targeting specific sequences in the genome.

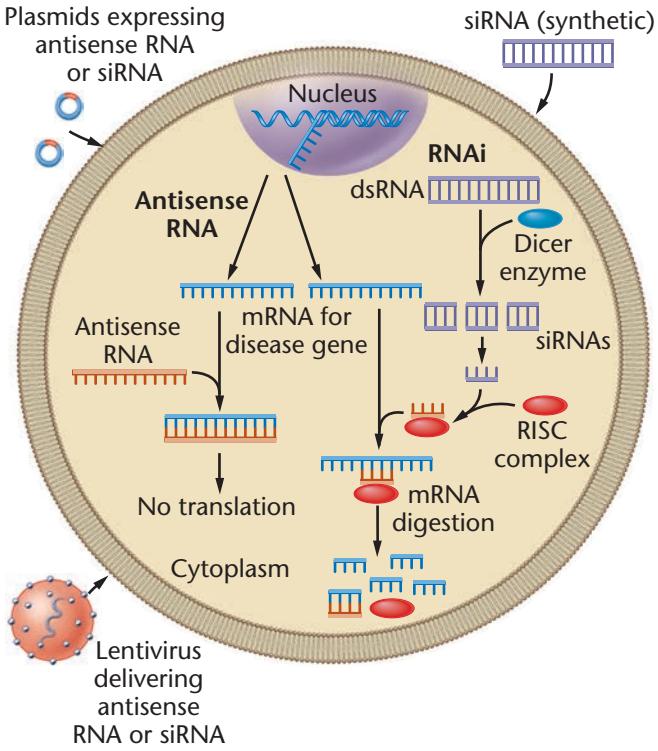
the world used CRISPR to target specific genes in human cells, mice, rats, bacteria, fruit flies, yeast, zebrafish, and dozens of other organisms. A team from the Massachusetts Institute of Technology (MIT) recently cured mice of a rare liver disorder, type I tyrosinemia, through gene editing by CRISPR. In tyrosinemia, a condition affecting about 1 in 100,000 people, mutation of the *Fuh* gene encoding the enzyme fumarylacetoacetase prevents breakdown of the amino acid tyrosine. After an *in vivo* approach with a one-time treatment, roughly 1 in 250 liver cells accepted the CRISPR-delivered replacement of the mutant gene with a normal copy of the gene. But about 1 month later these cells proliferated and replaced diseased cells, taking over about one-third of the liver, which was sufficient to allow mice to metabolize tyrosine and show no effects of disease. Mice were subsequently taken off a low-protein diet and a drug normally used to disrupt tyrosine production. Other headline-grabbing examples of successful CRISPR applications in mice and humans include therapies for β-thalassemia, cancer genes, and HIV. Based on the rapid development and effectiveness of the CRISPR method, CRISPR is clearly the most current, promising tool for gene editing in the future. Stay tuned!

RNA Silencing for Gene Inhibition

Attempts have been made to use **antisense oligonucleotides** to inhibit translation of mRNAs from defective genes, but this approach to gene therapy has generally not yet proven to be reliable. Nonetheless, the emergence of RNA interference as a powerful gene-silencing tool has reinvigorated gene therapy approaches by gene silencing. As you learned in earlier in the text (see Chapter 15), **RNA interference** (**RNAi**) is a form of gene-expression regulation. In animals short, double-stranded RNA molecules are delivered into cells where the enzyme Dicer chops them into 21- to 25-nt long pieces called **small interfering RNAs** (**siRNAs**). siRNAs then join with an enyzme complex called the **RNA-inducing silencing complex** (**RISC**), which shuttles the siRNAs to their target mRNA, where they bind by complementary base pairing. The RISC complex can block siRNA-bound mRNAs from being translated into protein or can lead to degradation of siRNA-bound mRNAs so that they cannot be translated into protein (**ST Figure 6–7**).

A main challenge to RNAi-based therapeutics so far has been *in vivo* delivery of double-stranded RNA or siRNA. RNAs degrade quickly in the body. It is also hard to get RNA to penetrate cells in the target tissue. For example, how does one deliver RNA-based therapies to cancer cells but not to noncancerous, healthy cells? Two common delivery approaches are to inject the siRNA directly or to deliver them via a DNA plasmid vector that is taken in by cells and transcribed to make double-stranded RNA which Dicer can cleave into siRNAs. Lentivirus, liposome, and attachment of siRNAs to cholesterol and fatty acids are other approaches being used to deliver siRNAs (ST Figure 6–7).

More than a dozen clinical trials involving RNAi are underway in the United States. Several RNAi clinical trials to treat blindness are showing promising results. One RNAi strategy to treat a form of blindness called macular degeneration targets a gene called *VEGF*. The VEGF protein promotes blood vessel growth. Overexpression of this gene, causing excessive production of blood vessels in the retina, leads to impaired vision and eventually blindness. Many



ST FIGURE 6–7 Antisense RNA and RNA interference (RNAi) approaches to silence genes for gene therapy. Antisense RNA technology and RNAi are two ways to silence gene expression and turn off disease genes.

expect that this disease will soon become the first condition to receive approval for treatment by RNAi therapy. Other disease candidates for treatment by RNAi include several different cancers, diabetes, liver diseases, multiple sclerosis, and arthritis.

Future Challenges and Ethical Issues

Despite the progress that we have noted thus far, many questions remain to be answered before we can hope for widespread application of the gene therapy methodology in the treatment of genetic disorders:

- What is the proper route for gene delivery in different kinds of disorders? For example, what is the best way to treat brain or muscle tissues?
- What percentage of cells in an organ or a tissue need to express a therapeutic gene to alleviate the effects of a genetic disorder?
- What amount of a therapeutic gene product must be produced to provide lasting improvement of the condition, and how can sufficient production be ensured? Currently, many approaches provide only short-lived delivery of the therapeutic gene and its protein.
- Will it be possible to use gene therapy to treat diseases that involve multiple genes?
- Can expression or the timing of expression of therapeutic genes be controlled in a patient so that genes can be turned on or off at a particular time or as necessary?
- Will targeted gene delivery approaches become more widely used for gene therapy trials?

For many people, the question remains whether gene therapy can ever recover from past setbacks and fulfill its promise as a cure for genetic diseases. Clinical trials for any new therapy are potentially dangerous, and often, animal studies will not accurately reflect the reaction of individual humans to the methodology leading to the delivery of new genes. However, as the history of similar struggles encountered with such life-saving developments such as the use of antibiotics and organ transplants has shown, there will be setbacks and even tragedies, but step by small step, we will move toward a technology that could—someday—provide reliable and safe treatment for severe genetic diseases.

Ethical Concerns Surrounding Gene Therapy

Gene therapy raises several ethical concerns, and many forms of gene therapy are sources of intense debate. At present, all gene therapy trials are restricted to using somatic cells as targets for gene transfer. This form of gene therapy is called **somatic gene therapy;** only one individual is affected, and the therapy is done with the permission and informed consent of the patient or family.

Two other forms of gene therapy have not been approved, primarily because of the unresolved ethical issues surrounding them. The first is called **germ-line therapy,** whereby germ cells (the cells that give rise to the gametes—i.e., sperm and eggs) or mature gametes are used as targets for gene transfer. In this approach, the transferred gene is incorporated into all the future cells of the body, including the germ cells. As a result, individuals in future generations will also be affected, without their consent. Is this kind of procedure ethical? Do we have the right to make this decision for future generations? Thus far, the concerns have outweighed the potential benefits, and such research is prohibited.

Box 3 mentioned gene doping, which is also an example of **enhancement gene therapy,** whereby people may be "enhanced" for some desired trait. This is another unapproved form of gene therapy—which is extremely controversial and is strongly opposed by many people. Should genetic technology be used to enhance human potential? For example, should it be permissible to use gene therapy

BOX 3

Gene Doping for Athletic Performance?

Gene therapy is intended to provide treatments or cures for genetic diseases, but it can also apply for those seeking genetic enhancements to improve athletic performance. As athletes seek a competitive edge, will gene therapy as a form of "gene doping" to improve performance be far behind?

We already know that in animal models enhanced muscle function can be achieved by gene addition. For example, adding copies of the insulin-like growth factor (*IGF-1*) gene to mice improves aspects of muscle function. The kidney hormone erythropoietin (EPO) increases red blood cell production, which leads to a higher oxygen content of the blood and thus improved endurance. Synthetic forms of EPO are banned in Olympic athletes. Several groups have proposed using gene therapy to deliver the *EPO* gene into athletes "naturally."

Since 2004 the World Anti-Doping Agency (WADA) has included gene doping through gene therapy as a prohibited method in sanctioned competitions. However, methods to detect gene doping are not well established. If techniques for gene therapy become more routine, many feel it is simply a matter of time before gene doping through gene therapy will be the next generation of performance-enhancement treatments. Obviously, many legal and ethical questions will arise if gene doping becomes a reality.

to increase height, enhance athletic ability, or extend intellectual potential? Presently, the consensus is that enhancement therapy, like germ-line therapy, is an unacceptable use of gene therapy. However, there is an ongoing debate, and many issues are still unresolved.

Gene therapy is currently a fairly expensive treatment. But what is the right price for a cure? It remains to be seen how health-care insurance providers will view gene therapy. But if gene therapy treatments provide a health-care option that drastically improves the quality of life for patients for whom there are few other options, it is likely that insurance companies will reimburse patients for treatment costs.

Finally, *whom* to treat by gene therapy is yet another ethically provocative consideration. In the Jesse Gelsinger case mentioned earlier, the symptoms of his OTC deficiency were minimized by a low protein diet and drug treatments. Whether it was necessary to treat Jesse by gene therapy is a question that has been widely debated.

Visit the Study Area in MasteringGenetics for a list of further readings on this topic, including journal references and selected web sites.

Review Questions

1. What is gene therapy?
2. Compare and contrast *ex vivo* and *in vivo* gene therapy as approaches for delivering therapeutic genes.
3. When treating a person by gene therapy, is it necessary that the therapeutic gene becomes part of a chromosome (integration) when inserted into cells? Explain your answer.
4. Describe two ways that therapeutic genes can be delivered into cells.
5. Explain how viral vectors can be used for gene therapy and provide two examples of commonly used viral vectors. What are some of the major challenges that must be overcome to develop safer and more effective viral vectors for gene therapy?
6. During the first successful gene therapy trial in which Ashanti DeSilva was treated for SCID, did the therapeutic gene delivered to Ashanti replace the defective copy of the ADA gene? Why were white blood cells chosen as the targets for the therapeutic gene?
7. Explain an example of successful gene therapy trial. In your answer be sure to consider: a description of the disease condition that was treated, the mutation or disease gene affected, the therapeutic gene delivered, and the method of delivery use for the therapy.
8. What is targeted gene therapy or gene editing, and how does this approach differ from traditional gene therapy approaches?
9. How do ZFNs work?
10. Describe two gene-silencing techniques and explain how they may be used for gene therapy.

Discussion Questions

1. Discuss the challenges scientists face in making gene therapy a safe, reliable, and effective technique for treating human disease conditions.
2. Who should be treated by gene therapy? What criteria are used to determine if a person is a candidate for gene therapy? Should gene therapy be used for cosmetic purposes or to improve athletic performance?
3. Describe future challenges and ethical issues associated with gene therapy.

APPENDIX

Solutions to Selected Problems and Discussion Questions

Chapter 1

2. Your essay should include a description of the impact of recombinant DNA technology on the following: plant and animal husbandry and production, drug development, medical advances, forensics, and understanding gene function.

4. The genotype of an organism is defined as the specific allelic or genetic constitution of an organism, or, often, the allelic composition of one or a limited number of genes under investigation. The observable feature of those genes is called the phenotype. A gene variant is called an allele.

6. A gene is a portion of DNA that encodes the information required to make a specific protein. The DNA is transcribed to make a messenger RNA, which is then translated by ribosomes into a chain of amino acids. The chain then folds up into a specific structure and is processed into the finished protein.

8. The central dogma of molecular genetics refers to the relationships among DNA, RNA, and proteins. The processes of transcription and translation are integral to understanding these relationships. Because DNA and RNA are discrete chemical entities, they can be isolated, studied, and manipulated in a variety of experiments that define modern genetics.

10. Restriction enzymes (endonucleases) cut double-stranded DNA at particular base sequences. When a vector is cleaved with the same enzyme, complementary ends are created such that ends, regardless of their origin, can be combined and ligated to form intact double-stranded structures. Such recombinant forms are often useful for industrial, research, and/or pharmaceutical efforts.

12. Unique transgenic plants and animals can be patented, as ruled by the United States Supreme Court in 1980. Supporters of organismic patenting argue that it is needed to encourage innovation and allow the costs of discovery to be recovered. Capital investors assume that there is a likely chance that their investments will yield positive returns. Others argue that natural substances should not be privately owned and that once they are owned by a small number of companies, free enterprise will be stifled.

14. All life has a common origin, so genes with similar functions tend to be similar in structure and nucleotide sequence in different organisms. That is why what scientists learn by studying the genetics of model organisms can be used to understand human diseases. The use of model organisms like the mouse (*Mus musculus*) also allows scientists to carry out genetic studies that would be unethical to carry out in humans; for e.g., setting up specific matings or creating knockouts or transgenic individuals.

16. Advances in bioinformatics are limited only by the advances in information technology in general, which is expanding exponentially due to extensive financial investment in research by the industries. Policies and legislation regarding the ethical issues will always lag behind because they are in response to public opinion about the new advances in biotechnology.

Chapter 2. Answers to Now Solve This

2-1. 32 chromatids, 16 chromosomes moving to each pole

2-2. **(a)** eight tetrads **(b)** eight dyads **(c)** eight monads

2-3. Not necessarily. If crossing over occurred in meiosis I, then the chromatids in the secondary oocyte are not identical.

Solutions to Problems and Discussion Questions

2. Compared with mitosis, which maintains chromosomal constancy, meiosis provides for a reduction in chromosome number and an opportunity for the exchange of genetic material between homologous chromosomes. In mitosis there is no change in chromosome number or kind in the two daughter cells, whereas in meiosis numerous potentially different haploid (n) cells are produced. During oogenesis, only one of the four meiotic products is functional; however, four of the four meiotic products of spermatogenesis are potentially functional.

4. Chromosomes that are homologous share many properties including *overall length, position of the centromere (metacentric, submetacentric, acrocentric, telocentric), banding patterns, type and location of genes, and autoradiographic pattern*. *Diploidy* is a term often used in conjunction with the symbol $2n$. It means that both members of a homologous pair of chromosomes are present. Haploidy refers to the presence of a single copy of each homologous chromosome (n).

12. During meiosis I, chromosome number is reduced to haploid complements. This is achieved by synapsis of homologous chromosomes and their subsequent separation. It would seem to be more mechanically difficult for genetically identical daughters to form from mitosis if homologous chromosomes paired. By having chromosomes unpaired at metaphase of mitosis, only centromere division is required for daughter cells to eventually receive identical chromosomal complements.

14. The stages of cell cycle follow a specific order of events. Moreover, a new stage is never initiated without the completion of the prior stage. This is due to sequential activation of cyclin and cyclin-dependent kinases. For example, mitotic cyclin/CDKs can only be active in mitosis and never in G1, S, or G2. This regulation is established by cell cycle specific transcription, post-translational modifications, regulating cellular localization of cyclin/CDK partners, and cyclin/CDK inhibitors.

16. There would be 16 combinations with the addition of another chromosome pair.

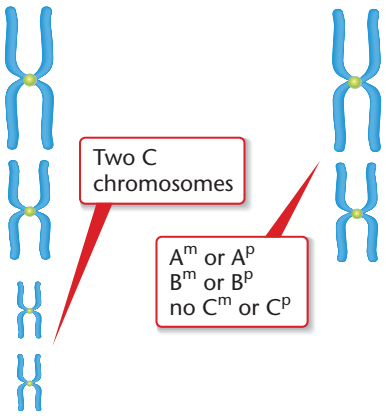
18. One-half of each tetrad will have a maternal homolog: $(1/2)^{10}$.

24. 50, 50, 50, 100, 200

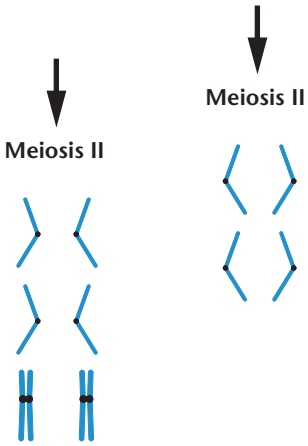
26. Duplicated chromosomes A^m, A^p, B^m, B^p, C^m, and C^p will align at metaphase, with the centromeres dividing and sister chromatids going to opposite poles at anaphase.

28. As long as you have accounted for eight possible combinations in the previous problem, there would be no new ones added in this problem.

30. See the products of nondisjunction of chromosome C at the end of meiosis I as follows.



At the end of meiosis II, assuming that, as the problem states, the C chromosomes separate as dyads instead of monads during meiosis II, you would have monads for the A and B chromosomes and dyads (from the cell on the left) for both C chromosomes as one possibility.



Chapter 3. Answers to Now Solve This

3-1. *P* = checkered; *p* = plain. Checkered is tentatively assigned the dominant function because, especially in cross (b), we see that checkered types are more likely to be produced than plain types.

		Progeny	
P_1 Cross		***Checkered***	***Plain***
(a)	*PP* × *PP*	*PP*	
(b)	*PP* × *pp*	*Pp*	
(c)	*pp* × *pp*		*pp*
(d)	*PP* × *pp*	*Pp*	
(e)	*Pp* × *pp*	*Pp*	*pp*
(f)	*Pp* × *Pp*	*PP, Pp*	*pp*
(g)	*PP* × *Pp*	*PP, Pp*	

3-2. Symbolism as before:

w = wrinkled seeds *g* = green cotyledons
W = round seeds *G* = yellow cotyledons

Examine each characteristic (seed shape vs. cotyledon color) separately. **(a)** Notice a 3:1 ratio for seed shape; therefore, *Ww* × *Ww*, and no green cotyledons; therefore, *GG* × *GG* or *GG* × *Gg*. Putting the two characteristics together gives

WwGG × *WwGG* or *WwGG* × *WwGg*.

(b) *WwGg* × *wwGg*. **(c)** *WwGg* × *WwGg*. **(d)** *WwGg* × *wwgg*.

3-3. **(a)**

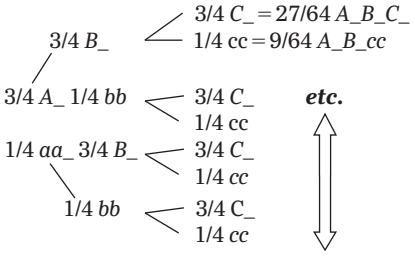
Genotypes	***Ratio***	***Phenotypes***
AABBCC	(1/16)	*A_B_C_* = 12/16
AABBCc	(1/16)	
AABbCC	(1/16)	
AABbCc	(1/16)	
AaBBCC	(2/16)	
AaBBCc	(2/16)	
AaBbCC	(2/16)	
AaBbCc	(2/16)	
aaBBCC	(1/16)	*aaB_C_* = 4/16
aaBBCc	(1/16)	
aaBbCC	(1/16)	
aaBbCc	(1/16)	

(b)

Genotypes	***Ratio***	***Phenotypes***
AaBBCC	1/8	*A_BBC_* = 3/8
AaBBCc	2/8	
AaBBcc	1/8	*A_BBcc* = 1/8
aaBBCC	1/8	*aaBBC_* = 3/8
aaBBCc	2/8	
aaBBcc	1/8	*aaBBcc* = 1/8

(c) There will be eight (2^n) different kinds of gametes from each of the parents and therefore a 64-box Punnett square.

For the phenotypic frequencies, set up the problem in the following manner:



3-4. **(a)** $\chi^2 = 0.47$

The χ^2 value is associated with a probability greater than 0.90 for 3 degrees of freedom (because there are now four classes in the χ^2 test). We fail to reject the null hypothesis and conclude that the observed values do not differ significantly from the expected values.

To deal with parts (b) and (c), it is easier to see the observed values for the monohybrid ratios if the phenotypes are listed:

round, yellow	315	round, green	108
wrinkled, yellow	101	wrinkled, green	32

(b) For the round: wrinkled *monohybrid component*, the smooth types total 423 (315 + 108), while the wrinkled types total 133 (101 + 32).

The χ^2 value is 0.35, and in examining the text for 1 degree of freedom, the *p* value is greater than 0.50 and less than 0.90. We fail to reject the null hypothesis and conclude that the observed values do not differ significantly from the expected values.

(c) For the yellow:green portion of the problem, see that there are 416 yellow plants (315 + 101) and 140 (108 + 32) green plants.

The χ^2 value is 0.01, and in examining the text for 1 degree of freedom, the *p* value is greater than 0.90. We fail to reject the null hypothesis.

Solutions to Problems and Discussion Questions

2. Your essay should include the following points: **1.** Factors occur in pairs. **2.** Some genes have dominant and recessive alleles. **3.** Alleles segregate from each other during gamete formation. When homologous chromosomes separate from each other at anaphase I, alleles will go to opposite poles of the meiotic apparatus. **4.** One gene pair separates independently from other gene pairs. Different gene pairs on the same homologous pair of chromosomes (if far apart) or on non-homologous chromosomes will separate independently from each other during meiosis.

4. **(a)** Only if the abnormal parent is *Mm*, crossing with a normal parent (*mm*) will yield some abnormal (*Mm*) and some normal (*mm*) progeny. **(b)** Since all the children are normal, the abnormal parent has to be heterozygous (*Mm*). Hence, the male is heterozygous, while the female is homozygous recessive.

6. *Pisum sativum* is easy to cultivate. It is naturally self-fertilizing, but it can be crossbred. It has several visible features (e.g., tall or short, red flowers or white flowers) that are consistent under a variety of environmental conditions, yet contrast due

to genetic circumstances. Seeds could be obtained from local merchants.

8. Three of Mendel's postulates are illustrated in this problem. Unit factors occur in pairs (postulate 1) and demonstrate dominance/recessive relationships (postulate 2). The fact that these unit factors separate from each other during gamete formation illustrates postulate 3.

10. 1. Factors occur in pairs. 2. Some genes have dominant and recessive alleles. 3. Alleles segregate from each other during gamete formation. When homologous chromosomes separate from each other at anaphase I, alleles will go to opposite poles of the meiotic apparatus. 4. One gene pair separates independently from other gene pairs.

12. **(a)** $35 = 243$

(b) 1/243

(c) no

(d) yes

14. Symbols:

Seed shape	***Seed color***
W = round	G = yellow
w = wrinkled	g = green

F_1: *WwGg* × *wwgg*
1/4 *WwGg* (round, yellow)
1/4 *Wwgg* (round, green)
1/4 *wwGg* (wrinkled, yellow)
1/4 *wwgg* (wrinkled, green)

16. **(a)** For the 3:1 ratio, the expected numbers should be 270 for phenotype 1 and 90 for phenotype 2. Solving for $x^2 = \Sigma = \frac{(o-e)^2}{e}$, we get $\chi^2 = 72.5$. Based on χ^2 for *df* 1, $p < 0.05$. Hence, for a critical p value of 0.05, null hypothesis is rejected. **(b)** Solve similarly as shown in part (a) above, but for 1:1 ratio.

18. 1/8

20. 3/4

22. *Unit factors in pairs*, *Dominance and recessiveness*, *Segregation*

24. **(a)** There are two possibilities. Either the trait is dominant, in which case I-1 is heterozygous, as are II-2 and II-3, or the trait is recessive and I-1 is homozygous and I-2 is heterozygous. Under the condition of recessiveness, both II-1 and II-4 would be heterozygous; II-2 and II-3 would be homozygous.

(b) Recessive: Parents *Aa, Aa*

(c) Recessive: Parents *Aa, Aa*

(d) Recessive or dominant: if recessive, parents *AA* (probably), *aa*. Second pedigree: recessive or dominant, not sex-linked (because this topic has not been covered as yet), if recessive, parents *Aa*, *aa*

26. **(a)** Notice in cross #1 that the ratio of straight wings to curled wings is 3:1 and the ratio of short bristles to long bristles is also 3:1. This would indicate that straight is dominant to curled and short is dominant to long. Possible symbols would be (using standard *Drosophila* symbolism):

straight wings = w^+ curled wings = w
short bristles = B^+ long bristles = B

(b)
Cross #1: $w^+/w\,;B^+/B \times w^+/w\,;B^+/B$
Cross #2: $w^+/w\,;B/B \times w^+/w\,;B/B$
Cross #3: $w/w\,;B/B \times w^+/w\,;B^+/B$
Cross #4: $w^+/w^+;B^+/B \times w^+/w^+;B^+/B$
(one parent could be w^+/w)
Cross #5: $w/w\,;B^+/B \times w^+/w\,;B^+/B$

Chapter 4. Answers to Now Solve This

4-1. **(a)** $c^kc^a \times c^dc^a \Longrightarrow$ 2/4 sepia; 1/4 cream; 1/4 albino

(b) Because the cream guinea pig had two sepia parents, ($c^kc^d \times c^kc^d$ or $c^kc^d \times c^kc^a$), the cream parent could be c^dc^d or c^dc^a.
$c^kc^a \times c^dc^d \Longrightarrow$ 1/2 sepia; 1/2 cream*
*(If the cream parent is assumed to be homozygous)
or $c^kc^a \times c^dc^a \Longrightarrow$ 1/2 sepia; 1/4 cream; 1/4 albino

(c) Crosses possible: various other possibilities exist depending on the genotypes of the parents.
$c^kc^k \times c^dc^d \Longrightarrow$ all sepia
$c^kc^k \times c^dc^a \Longrightarrow$ all sepia
$c^kc^d \times c^dc^d \Longrightarrow$ 1/2 sepia; 1/2 cream
$c^kc^d \times c^dc^a \Longrightarrow$ 1/2 sepia; 1/2 cream
$c^kc^a \times c^dc^d \Longrightarrow$ 1/2 sepia; 1/2 cream
$c^kc^a \times c^dc^a \Longrightarrow$ 1/2 sepia; 1/4 cream; 1/4 albino

(d) Crosses possible: other possibilities exist, depending on the genotypes of the parents.
$c^kc^a \times c^dc^d \Longrightarrow$ 1/2 sepia; 1/2 cream
$c^kc^a \times c^dc^a \Longrightarrow$ 1/2 sepia; 1/4 cream; 1/4 albino

4-2. A = pigment; a = pigmentless (colorless) B = purple; b = red
AaBb × *AaBb*: *A_B_* = purple; *A_bb* = red; *aaB_* = colorless; *aabb* = colorless

4-3. Let a represent the mutant gene and A represent its normal allele.

(a) This pedigree is consistent with an X-linked recessive trait because the male would contribute an X chromosome carrying the a mutation to the *aa* daughter. The mother would have to be heterozygous *Aa*.

(b) This pedigree is consistent with an X-linked recessive trait because the mother could be *Aa* and transmit her a allele to her one son (a/Y) and her A allele to her other son.

(c) This pedigree is not consistent with an X-linked mode of inheritance because the *aa* mother has an A/Y son.

Solutions to Problems and Discussion Questions

2. Your essay should include a description of alleles that do not function independently of each other or they reduce the viability of a class of offspring. With multiple alleles, there are more than two alternatives of a given gene.

4. 20 (large leaves), 40 (medium leaves), 20 (small leaves); incomplete dominance.

6. Flower color: *RR* = red; *Rr* = pink; *rr* = white
Flower shape: P = personate; p = peloric
Plant height: D = tall; d = dwarf

(a) *RRPPDD* × *rrppdd* ⟹ *RrPpDd* (pink, personate, tall)

(b) 2/4 pink × 3/4 personate × 3/4 tall = 18/64

8. **(a)** Each parent produces only one type of gamete (*CrP* or *cRp*). Hence, all the progeny are *CcRrPp*, that is, all are purple.

(b) Each progeny inherits at least one copy of *C* and one copy of *R*. *C, R*, and *P* are always inherited together. Hence, all the progeny are red.

(c) Using the forked-line method, we find that 9/32 are purple, 9/32 are red, and 14/32 are colorless.

10. **(a)** 1/4 **(b)** 1/2 **(c)** 1/4 **(d)** zero

12. This situation is similar to sex-influenced pattern baldness in humans. Consider two alleles that are autosomal and let

BB = beardless in both sexes
Bb = beardless in females
Bb = bearded in males
bb = bearded in both sexes
P_1: female: *bb* (bearded) × male: *BB* (beardless)
F_1: *Bb* = female beardless; male bearded

Because half of the offspring are males and half are females, one could, for clarity, rewrite the F_2 as:

	1/2 females	*1/2 males*
1/4 *BB*	1/8 beardless	1/8 beardless
2/4 *Bb*	2/8 beardless	2/8 bearded
1/4 *bb*	1/8 bearded	1/8 bearded

One could test the above model by crossing F_1 (heterozygous) beardless females with bearded (homozygous) males. Comparing these results with the reciprocal cross would support the model if the distributions of sexes with phenotypes were the same in both crosses.

14. (a) F_1: carrier female and affected male, F_2: normal male, affected male, carrier female, and affected female.

(b) F_1: carrier female and normal male, F_2: normal female, carrier female, normal male, and affected male.

If the hemophilic gene was autosomal, then *hh* would be an affected individual, while *Hh* and *HH* would be normal.

16. (a) P_1: X^vX^v; +/+ × X^+/Y; *bw*/*bw*

F_1: 1/2 X^+X^v; +/*bw* (female, normal)
1/2 X^v/Y; +/*bw* (male, vermilion)

F_2: 3/16 = female, normal
1/16 = female, brown eyes
3/16 = female, vermilion eyes
1/16 = female, white eyes
3/16 = male, normal
1/16 = male, brown eyes
3/16 = male, vermilion eyes
1/16 = male, white eyes

(b) P_1: X^+X^+; *bw*/*bw* × X^v/Y; +/+

F_1: 1/2 X^+X^v; +/*bw* (female, normal)
1/2 X^+/Y; +/*bw* (male, normal)

F_2: 6/16 = female, normal
2/16 = female, brown eyes
3/16 = male, normal
1/16 = male, brown eyes
3/16 = male, vermilion eyes
1/16 = male, white eyes

(c) P_1: X^vX^v; *bw*/*bw* × X^+/Y; +/+

F_1: 1/2 X^+X^v; +/*bw* (female, normal)
1/2 X^v/Y; +/*bw* (male, vermilion)

F_2: 3/16 = female, normal
1/16 = female, brown eyes
3/16 = female, vermilion eyes
1/16 = female, white eyes
3/16 = male, normal
1/16 = male, brown eyes
3/16 = male, vermilion eyes
1/16 = male, white eyes

18. (a) Because the denominator in the ratios is 64, one would begin to consider that three independently assorting gene pairs were operating in this problem. Because there are only two characteristics (eye color and croaking), however, one might hypothesize that two gene pairs are involved in the inheritance of one trait while one gene pair is involved in the other.

(b) Notice that there is a 48:16 (or 3:1) ratio of rib-it to knee-deep and a 36:16:12 (or 9:4:3) ratio of blue to green to purple eye color. Because of these relationships, one would conclude that croaking is due to one (dominant/recessive) gene pair while eye color is due to two gene pairs. Because there is a 9:4:3 ratio regarding eye color, some gene interaction (epistasis) is indicated.

(c) Symbolism: Croaking: *R_* = utterer; *rr* = mutterer. Eye color: Since the most frequent phenotype is blue eye, let *A_B_* represent the genotypes. For the purple class, a 3/16 group uses the *A_bb* genotypes. The 4/16 class (green) would be the *aaB_* and the *aabb* groups.

(d) The cross involving a blue-eyed, mutterer frog and a purple-eyed, utterer frog would have the genotypes

AABBrr × *AAbbRR*

which would produce an F_1 of *AABbRr,* which would be blue-eyed and utterer. The F_2 will follow a pattern of a 9:3:3:1 ratio because of homozygosity for the *A* locus and heterozygosity for both the *B* and *R* loci.

9/16 *AAB_R_* = blue-eyed, utterer
3/16 *AAB_rr* = blue-eyed, mutterer
3/16 *AAbbR_* = purple-eyed, utterer
1/16 *AAbbrr* = purple-eyed, mutterer

20. A 12:3:1 ratio is obtained, which is a clear sign that epistasis has modified a typical 9:3:3:1 ratio. In this case, cattle in one of the 3/16 classes has the same phenotype as cattle in the 9/16 class. Since the 9/16 class typically takes the genotype of *A_B_,* it seems reasonable to think of the following genotypic classifications:

*A_B_* = solid white (9/16)
*aaB_* = solid white (3/16)
A_bb = black and white spotted (3/16)
aabb = solid black (1/16)

The selection of *bb* as giving the spotted phenotype is arbitrary. One could obtain *AAbb* true-breeding black and white spotted cattle.

22. Symbolism: *A_B_* = black *A_bb* = golden
aabb = golden *aaB_* = brown

The combination of *bb* is epistatic to the *A* locus.

(a) *AAB_* × *aaBB* (other configurations are possible, but each must give all offspring with *A* and *B* dominant alleles)
(b) *AaB_* × *aaBB* (other configurations are possible, but both parents cannot be *Bb*)
(c) *AABb* × *aaBb*
(d) *AABB* × *AAbb*
(e) *AaBb* × *Aabb*
(f) *AaBb* × *aabb*
(g) *aaBb* × *aaBb*
(h) *AaBb* × *AaBb*

Those genotypes that will breed true will be as follows:

black = *AABB*
golden = all genotypes that are *bb*
brown = *aaBB*

24. (a) $C^{ch}C^{ch}$ = chestnut C^cC^c = cremello $C^{ch}C^c$ = palomino

(b) The F_1 resulting from matings between cremello and chestnut horses would be expected to be all palomino. The F_2 would be expected to fall in a 1:2:1 ratio as in the third cross in part (a) above.

26. Cross #1 = (c) Cross #4 = (e)
Cross #2 = (d) Cross #5 = (a)
Cross #3 = (b)
Given that each parental/offspring grouping can only be used once, there are no other combinations.

28. The homozygous dominant type is lethal. Polled is caused by an independently assorting dominant allele, while horned is caused by the recessive allele to polled.

30. Maternal effect genes produce products that are not carried over for more than one generation, as is the case with

organelle and infectious heredity. Crosses that illustrate the transient nature of a maternal effect could include the following:

Female *Aa* × male *aa* ⟶ all offspring of the A phenotype.

Take a female A phenotype from the above cross and conduct the following mating: *aa* × male *Aa*. All offspring may be of the *a* phenotype because all of the offspring will reflect the *genotype* of the mother, not her *phenotype*. This cross illustrates that maternal effects last only one generation. However, depending on particular biochemical/developmental parameters, all crosses may not give these types of patterns.

32. (a, b) The reduced ratio is 12 white, 3 orange, and 1 brown, and in a dihybrid cross (*AaBb* × *AaBb*) the following would occur: 12 white *A_B_* or *aaB_:* 3 orange *A_bb:* 1 brown *aabb*

34. Beatrice, Alice of Hesse, and Alice of Athlone are carriers. There is a 1/2 chance that Princess Irene is a carrier.

Chapter 5. Answers to Now Solve This

5-1. (a) Something is missing from the male-determining system of sex determination at the level of the genes, gene products, or receptors, and so on, and the loss is correlated with CMD1.

(b) The *SOX9* gene, or its product, is probably involved in male development. Perhaps it is activated by *SRY*.

(c) There is probably some evolutionary relationship between the *SOX9* gene and *SRY*. There is considerable evidence that many other genes and pseudogenes are also homologous to *SRY*.

(d) Normal female sexual development does not require the *SOX9* gene or gene product(s).

5-2. Because of X chromosome inactivation in mammals, scientists would be interested in determining whether the nucleus taken from Rainbow (donor) would continue to show such inactivation. Would the inactivated X chromosome retain the property of inactivation? Since X chromosome inactivation is random, CC would have a different patch pattern from her genetic mother based on the random X inactivation alone.

Solutions to Problems and Discussion Questions

2. Your essay should include various aspects of sex chromosomes that contain genes responsible for sex determination. Mention should also be made of those organisms in which autosomes play a role in concert with the sex chromosomes.

6. In humans, the sex is determined by the presence or absence of the *SRY* gene located on the Y chromosome. Therefore, XO individuals are females with ovaries, a uterus, and oviducts, but very few ova. In *Drosophila*, the sex is determined by the balance of female determinants on the X chromosome(s) and male determinants on the autosomes, mediated by the *Sxl* gene. The Y chromosome does not determine sex, but is required for sperm production. Therefore, XO individuals are sterile males.

8. The Y chromosome is male determining in humans, and it is a particular region of the Y chromosome that causes maleness, the sex-determining region (SRY). SRY releases a product called the testis-determining factor (TDF), which causes the undifferentiated gonadal tissue to form testes. Individuals with the 47,XXY complement are males, while 45,XO produces females. In *Drosophila* it is the balance between the number of X chromosomes and the number of haploid sets of autosomes that determines sex. In contrast to humans, XO *Drosophila* are males and the XXY complement is female.

10. (a) female $X^{rw}Y$ × male $X^{+}X^{+}$

F_1:	females:	$X^{+}Y$ (normal)
	males:	$X^{rw}X^{+}$ (normal)
F_2:	females:	$X^{+}Y$ (*normal*)
		$X^{rw}Y$ (reduced wing)
	males:	$X^{rw}X^{+}$ (normal)
		$X^{+}X^{+}$ (normal)

(b) female $X^{rw}X^{rw}$ × male $X^{+}Y$

F_1:	females:	$X^{rw}X^{+}$ (normal)
	males:	$X^{rw}Y$ (reduced wing)
F_2:	females:	$X^{rw}X^{+}$ (*normal*)
		$X^{rw}X^{rw}$ (reduced wing)
	males:	$X^{+}Y$ (normal)
		$X^{rw}Y$ (reduced wing)

12. The zygote develops masculinized characteristics because of the hormones and transcription factors that are produced by the activity of the Y chromosome. Even though the zygote is a female genetically, it has masculinized reproductive organs.

14. Because attached-X chromosomes have a mother-to-daughter inheritance and the father's X is transferred to the son, one would see daughters with the white-eye phenotype and sons with the miniature wing phenotype. In addition, there would be rare (<3%) metafemales (attached-X + X) with wild-type eye color. YY zygotes fail to develop into larvae.

16. A *Barr body* is a differentially staining chromosome seen in some interphase nuclei of mammals with two X chromosomes. There will be one fewer Barr body than the number of X chromosomes. The Barr body is an X chromosome that is considered to be genetically inactive.

18. The *Lyon hypothesis* states that the inactivation of the X chromosome occurs at random early in embryonic development. Such X chromosomes are in some way "marked," such that all clonally related cells have the same X chromosome inactivated.

20. 1/4 of the offspring will be calico.

22. Many organisms have evolved over millions of years under the fine balance of numerous gene products, usually occurring with two copies (identical or similar) of each gene. Many genes required for normal cellular and organismic function in *both* males and females are located on the X chromosome where only one copy occurs. These gene products have nothing to do with sex determination or sex differentiation, but their output must be balanced in some manner.

24. Nondisjunction could have occurred either at meiosis I or meiosis II in the mother, thus giving the XwXwY complement in the offspring.

26. In snapping turtles, sex determination is strongly influenced by temperature such that males are favored in the 26–34°C range. Lizards, on the other hand, appear to have their sex determined by factors other than temperature in the 20–40°C range.

Chapter 6. Answers to Now Solve This

6-1. If the father had hemophilia, it is likely that the Turner syndrome individual inherited the X chromosome from the father and no sex chromosome from the mother. If nondisjunction occurred in the mother, either during meiosis I or meiosis II, an egg with no X chromosome can be the result. See the text for a diagram of primary and secondary nondisjunction.

6-2. The sterility of interspecific hybrids is often caused from a high proportion of univalents in meiosis I. As such, viable gametes are rare and the likelihood of two such gametes "meeting" is remote. Even if partial homology of chromosomes allows some pairing, sterility is usually the rule. The horticulturist may attempt to reverse the sterility by treating the sterile

hybrid with colchicine. Such a treatment, if successful, may double the chromosome number, so each chromosome would now have a homolog with which to pair during meiosis.

6-3. The rare double crossovers within the boundaries of a paracentric inversion heterozygote produce only minor departures from the standard chromosomal arrangement as long as the crossovers involve the same two chromatids. With two-strand double crossovers, the second crossover negates the first. However, three-strand and four-strand double crossovers have consequences that lead to anaphase bridges as well as a high degree of genetically unbalanced gametes.

Solutions to Problems and Discussion Questions

2. Your essay can draw from many examples discussed in the text as examples of deletions, duplications, inversions, translocations, and copy number variations.

4. Some cases of Down syndrome are due to a translocation between chromosomes 14 and 21. The parents are phenotypically normal, yet the translocation leads to trisomy 21 during meiosis.

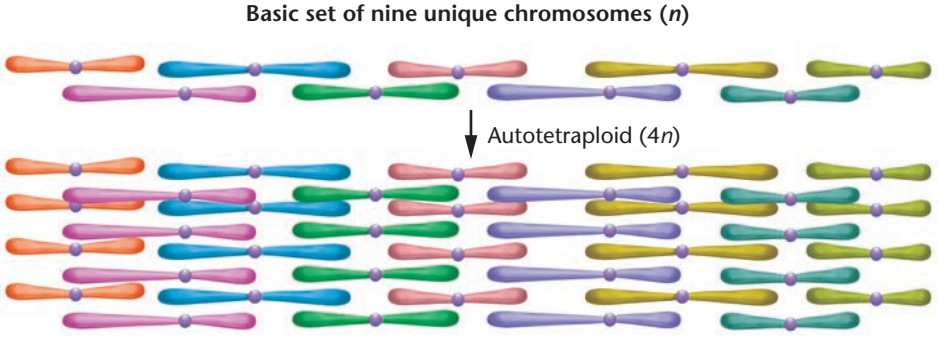
6. Chromosomes can break spontaneously or due to chemicals, and can abnormally fuse with other nonhomologous chromosomes, leading to a loss or rearrangement of the genetic material. Telomere shortening can be another reason why chromosome ends fuse with each other.

8. Due to gene redundancy, polyploid plants have a faster growth. They are, therefore, larger than their diploid relatives and have larger flowers and fruits that have higher economical value. The most significant disadvantage is infertility. The fertility of polyploid plants depends on the ability to generate balanced gametes. Two homologs of each chromosome are required for successful meiosis and fertilization.

10. Organisms with one inverted chromosome and one noninverted homolog are called inversion heterozygotes. Pairing of two such chromosomes is possible only through inversion loops. In case of pairing within the inversion loop, abnormal chromatids consisting of deletions and duplications will be produced.

12. Modern globin genes resulted from a duplication event in an ancestral gene about 500 million years ago. Mutations occurred over time and a chromosomal aberration separated the duplicated genes, leaving the eventual α cluster on chromosome 16 and the eventual β cluster on chromosome 11.

14. Given the basic chromosome set of nine unique chromosomes (a haploid complement), other forms with the "*n* multiples" are forms of autotetraploidy. In the illustration below the *n* basic set is multiplied to various levels as is the autotetraploid in the example.



Individuals organisms with 27 chromosomes are triploids (3*n*) and are more likely to be sterile because there are trivalents at meiosis I that cause a relatively high number of unbalanced gametes to be formed.

16. The cross would be as follows:

$$WWWW \times wwww$$

(assuming that chromosomes pair at meiosis)

F_1: *WWww*

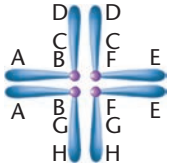
F_2:

	1*WW*	4*Ww*	1*ww*
1*WW*			
4*Ww*	35*W___* and 1*wwww*		
1*ww*			

18. Since two Gl_1 alleles and two ws_3 alleles are present in the triploid, they must have come from the pollen parent. By the wording of the problem, it is implied that the pollen parent contributed an unreduced (2*n*) gamete; however, another explanation, dispermic fertilization, is possible. In this case two Gl_1ws_3 gametes could have fertilized the ovule.

20. **(a)** reciprocal translocation

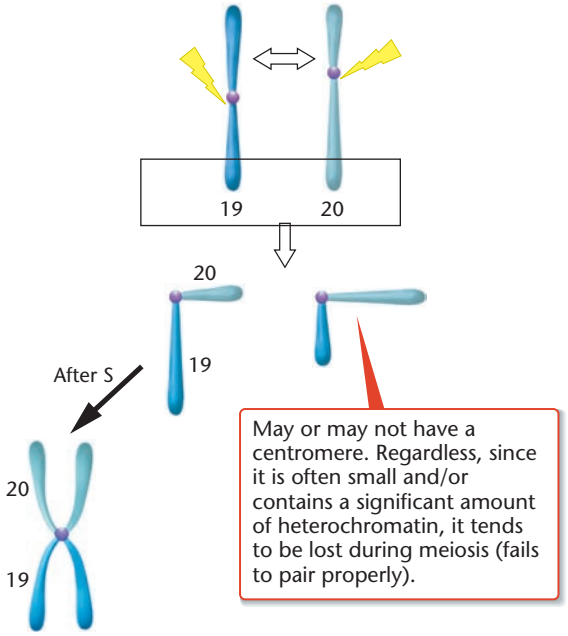
(b)



(c) Notice that all chromosomal segments are present and there is no apparent loss of chromosomal material. However, if the breakpoints for the translocation occurred within genes, then an abnormal phenotype may be the result. In addition, a gene's function is sometimes influenced by its position—its neighboring genes, in other words. If such "position effects" occur, then a different phenotype may result.

22. Instances of Down syndrome are either due to nondisjunction during meiosis in one of the parents, resulting in trisomy 21, or due to a 14/21, D/G translocation in one of the parents (familial Down syndrome). The latter results in a zygote that has 46 chromosomes but three copies of chromosome 21.

24. Below is a description of breakage/reunion events that illustrate a translocation in the relatively small, similarly sized chromosomes 19 (metacentric) and 20 (metacentric/submetacentric). The case described here is shown occurring before S phase duplication. Since the likelihood of such a translocation is fairly small in a general population, inbreeding played a significant role in allowing the translocation to "meet itself."



26. This female will produce meiotic products of the following types:

normal: 18 + 21
translocated: 18/21
translocated plus 21: 18/21 + 21
deficient: 18 only

Note: The 18/21 + 18 gamete is not formed because it would require separation of primarily homologous chromosomes at anaphase I.

Fertilization with a normal 18 + 21 sperm cell will produce the following offspring:

normal: 46 chromosomes
translocation carrier: 45 chromosomes 18/21 + 18 + 21
trisomy 21: 46 chromosomes 18/21 + 21 + 21
monosomic: 45 chromosomes 18 + 18 + 21, lethal

Chapter 7. Answers to Now Solve This

7-1. **(a)** 1/4 *AaBb* 1/4 *Aabb* 1/4 *aaBb* 1/4 *aabb*
(b) 1/4 *AaBb* 1/4 *Aabb* 1/4 *aaBb* 1/4 *aabb*
(c) If the arrangement is *AB/ab* × *ab/ab* then the two types of offspring will be as follows:

1/2 *Ab/ab* 1/2 *aB/ab*

If, however, *A* and *B* are not coupled, then the symbolism would be *Ab/aB* × *aabb*.
The offspring would occur as follows:

1/2 *Ab/ab* 1/2 *aB/ab*

7-2. The most frequent classes are *PZ* and *pz*. These classes represent the parental (noncrossover) groups, which indicates that the original parental arrangement in the testcross was *PZ/pz* × *pz/pz*. Adding the crossover percentages together (6.9 + 7.1) gives 14 percent, which would be the map distance between the two genes.

7-3. Examine the progeny list to see which types are not present. In this case, the double crossover classes are the following: + + *c* and *a b* +
(a, b) Gene *b* is in the middle and the arrangement is as follows.

+*b c*/*a* + +

a − *b* = 7 map units *b* − *c* = 2 map units

(c) The progeny phenotypes that are missing are + + *c* and *a b* +, of which, from 1000 offspring, 1.4 (0.07 × 0.02 × 1000) would be expected. Perhaps by chance or some other unknown selective factor, they were not observed.

Solutions to Problems and Discussion Questions

2. Your essay should include methods of detection through crosses with appropriate, distinguishable markers and that in most cases, the frequency of crossing over is directly related to the distance between genes.

4. With some qualification, especially around the centromeres and telomeres, one can say that crossing over is somewhat randomly distributed over the length of the chromosome. Two loci that are far apart are more likely to have a crossover between them than two loci that are close together.

6. If the probability of one event is 1/X, the probability of two events occurring at the same time will be $1/X^2$.

8. Each cross must be set up in such a way as to reveal crossovers because it is on the basis of crossover frequency that genetic maps are developed. It is necessary that genetic heterogeneity exist so that different arrangements of genes, generated by crossing over, can be distinguished. The organism that is heterozygous must be the sex in which crossing over occurs. Lastly, the cross must be set up so that the phenotypes of the offspring readily reveal their genotypes.

10. The heterozygous parent in the test cross is *RY/ry* × *ry/ry* with the two dominant alleles on one chromosome and the two recessives on the homolog. The map distance would be 10 map units between the *R* and *Y* loci.

12. The map for parts **(a)** and **(b)** is the following:

d.....*b*.....*pr*.....*vg*.....*c*.....*adp*
31 48 54 67 75 83
Map units

The expected map units between *d* and *c* would be 44, between *d* and *vg* 36, and between *d* and *adp* 52. However, because there is a theoretical maximum of 50 map units possible between two loci in any one cross, that distance would be below the 52 determined by simple subtraction.

14. **(a)** P_1: *sc s v*/*sc s v* × + + +/*Y*
F_1: + + +/*sc s v* × *sc s v*/*Y*

(a) The map distances are determined by first writing the proper arrangement and sequence of genes, and then computing the distances between each set of genes.

$$\frac{sc\ \ v\ \ s}{+\ +\ +}$$

sc − *v* = 33 percent (map units)
v − *s* = 10 percent (map units)

(c, d) The coefficient of coincidence = 0.727, which indicates that there were fewer double crossovers than expected; therefore, positive chromosomal interference is present.

16. **(a)** The *short* gene is on chromosome 2 with the *black* gene.
(b) The parental cross is now the following:

Females: *b sh p*/+ + + × Males: *b sh p*/*b sh p*

The new gametes resulting from crossing over in the female would be *b* + and + *sh*. Because the gene *p* is assorting independently, it is not important in this discussion. Fifteen percent of the offspring now contain these recombinant chromatids; therefore, the map distance between the two genes must be 15.

18. The genetic distance is 22 cM (mu).

20. Assign the following symbols, for example:

R = red *r* = yellow
O = oval *o* = long

Progeny A: *Ro/rO* × *rroo* = 10 map units
Progeny B: *RO/ro* × *rroo* = 10 map units

22. **(a)** There would be $2n = 8$ genotypic and phenotypic classes, and they would occur in a 1:1:1:1:1:1:1:1 ratio.
(b) There would be two classes, and they would occur in a 1:1 ratio.
(c) There are 20 map units between the *A* and *B* loci, and locus *C* assorts independently from both the *A* and *B* loci.

24. By having microscopically visible markers on the chromosomes, Creighton and McClintock were able to show microscopically that homologous chromosomal material physically exchanged segments during crossing over.

26. Because sister chromatids are genetically identical (with the exception of rare new mutations), crossing over between sisters provides no increase in genetic variability.

Chapter 8. Answers to Now Solve This

8-1.

Hfr Strain	*Order*
1	*tchro*
2	*hromb*
3	*<<chrom*
4	*mbakt>>*
5	*<<baktc*
Overall:	*tchrombaktc*

8-2. In the first dataset, the transformation of each locus, a^+ and b^+, occurs at a frequency of 0.031 and 0.012, respectively. To determine whether there is linkage, one would determine whether the frequency of double transformants a^+b^+ is greater than that expected by a multiplication of the two independent events. Multiplying 0.031×0.012 gives 0.00037, or approximately 0.04 percent. From this information, one would consider no linkage between these two loci. Notice that this frequency is approximately the same as the frequency in the second experiment, where the loci are transformed independently.

Solutions to Problems and Discussion Questions

2. Your essay should include a description of conjugation, transformation, transduction, and the potential recombination that occurs as a result of these processes.

4. **(a)** The requirement for physical contact between bacterial cells during conjugation was established by placing a filter in a U-tube such that the medium can be exchanged, but the bacteria cannot come in contact. Under this condition, conjugation does not occur.
(b) By treating cells with streptomycin, an antibiotic, it was shown that recombination would not occur if one of the two bacterial strains was inactivated. However, if the other was similarly treated, recombination would occur.
(c) An F^+ bacterium contains a circular, doublestranded, structurally independent DNA molecule. that can direct recombination.

8. The F^+ element can enter the host bacterial chromosome, and upon returning to its independent state, it may pick up a piece of a bacterial chromosome. When transferred to a bacterium with a complete chromosome, a partial diploid, or merozygote, is formed.

10. The phage not only lacks genes for ribosomal construction, but also contains no ribosomes. Upon infection, phage genes are transcribed, and the transcripts are translated using bacterial ribosomes.

12. A single plaque originates from the replicative activity of a single bacteriophage.

14. Assuming the typical introduction of 0.1 ml of the phage suspension to the bacterial solution, since 17 plaques were formed, the initial density of bacteriophage suspension would be 1.7×10^8 phage/ml.

16. Viruses that either lyse the cell or behave as a prophage are temperate phages. Those that only lyse the cell are called virulent phages.

18. Viral recombination occurs when there is a sufficiently high number of infecting viruses so that there is a high likelihood that more than one variant of phage will infect a given bacterium. Under this condition, phage chromosomes can recombine by crossing over.

20. **(a)** Remembering that 0.1 ml is typically used in the plaque assay, the initial concentration of phage per milliliter is greater than 10^5.
(b) Remembering that 0.1 ml is typically used in the plaque assay, the initial concentration of phage per milliliter is around 140×10^6 or 1.4×10^8.
(c) Remembering that 0.1 ml is typically used in the plaque assay, the initial concentration of phage is less than 10^9. Coupling this information with the calculations in part (b) above, it would appear that the initial concentration of phage is around 1×10^8.

Chapter 9. Answers to Now Solve This

9-1. In theory, the general design would be appropriate in that some substance, if labeled, would show up in the progeny of transformed bacteria. However, since the amount of transforming DNA is extremely small compared to the genomic DNA of the recipient bacterium and its progeny, it would be technically difficult to assay for the labeled nucleic acid. In addition, it would be necessary to know that the small stretch of DNA that caused the genetic transformation was actually labeled.

9-2. Guanine = 17.5%, adenine and thymine both = 32.5%.

9-3. Assuming the value of 1.13 is statistically different from 1.00, one can conclude that rubella is a single-stranded RNA virus.

Solutions to Problems and Discussion Questions

2. Your essay should include a description of structural aspects including sugar and base content comparisons. In addition, you should mention complementation aspects, strandedness, flexibility and conformation.

6. Nucleic acids contain large amounts of phosphorus and no sulfur, whereas proteins contain sulfur and no phosphorus. Therefore, the radioisotopes ^{32}P and ^{35}S will selectively label nucleic acids and proteins, respectively. The Hershey and Chase experiment was based on the premise that the substance injected into the bacterium is the substance responsible for producing the progeny phage and therefore must be the hereditary material. The experiment demonstrated that most of the ^{32}P-labeled material (DNA) was injected, while the phage ghosts (protein coats) remained outside the bacterium. Therefore, the nucleic acid must be the genetic material.

8. The early evidence would be considered indirect in that at no time was there an experiment, like transformation in bacteria, in which genetic information in one organism was transferred to another using DNA. Rather, by comparing DNA content in various cell types (sperm and somatic cells) and observing that the *action* and *absorption* spectra of ultraviolet light were cor-related, DNA was considered to be the genetic material. This suggestion was supported by the fact that DNA was shown to be the genetic material in bacteria and some phage. Direct evidence that DNA is the genetic material comes from a variety of observations, including gene transfer that has been facilitated by recombinant DNA techniques.

12.
Uracil: 2,4-dioxypyrimidine
Thymine: 2,4-dioxy-5-methylpyrimidine
Adenine: 6-aminopurine
Guanine: 2-amino-6-oxypurine

16. Because in double-stranded DNA, A = T and G = C (within limits of experimental error), the data presented would have indicated a lack of pairing of these bases in favor of a single-stranded structure or some other nonhydrogen-bonded structure.

Alternatively, from the data it would appear that A = C and T = G, which would negate the chance for typical hydrogen bonding since opposite charge relationships do not exist. Therefore, it is quite unlikely that a tight helical structure would form at all.

20. The nitrogenous bases of nucleic acids (nucleosides, nucleotides, and single- and double-stranded polynucleotides) absorb UV light maximally at wavelengths of 254 to 260 nm. One can often determine the presence and concentration of nucleic acids in a mixture. Since proteins absorb UV light maximally at 280 nm, this is a relatively simple way of dealing with mixtures of biologically important molecules.

UV absorption is greater in single-stranded molecules (hyperchromic shift) than in double-stranded structures. Therefore, by applying denaturing conditions, one can easily determine whether a nucleic acid is in the single- or double-stranded form. In addition, A-T rich DNA denatures more readily than G-C rich DNA. Therefore, one can estimate base content by denaturation kinetics.

22. *A hyperchromic effect* is the increased absorption of UV light as double-stranded DNA (or RNA, for that matter) is converted to single-stranded DNA. As illustrated in the text, the change in absorption is quite significant, with a structure of higher G-C content *melting* at a higher temperature than an A-T rich nucleic acid. If one monitors the UV absorption with a spectrophotometer during the melting process, the hyperchromic shift can be observed. The T_m is the point on the profile (temperature) at which half (50 percent) of the sample is denatured.

24. The reassociation of separate complementary strands of a nucleic acid, either DNA or RNA, is based on hydrogen bonds forming between A-T (or U) and G-C.

26. **(1)** As shown, the extra phosphate is not normally expected.
(2) In the adenine ring, a nitrogen is at position 8 rather than position 9.
(3) The bond from the C-1′ to the sugar should form with the N at position 9 (N-9) of the adenine.
(4) The dinucleotide is a "deoxy" form; therefore, each C-2′ should not have a hydroxyl group. Notice the hydroxyl group at C-2′ on the sugar of the adenylic acid.
(5) At the C-5 position on the thymine residue, there should be a methyl group.
(6) There are too many bonds at the N-3 position on the thymine.
(7) There are too few bonds at the C-5 of thymine.

28. If thymine gets converted to uracil, the A = T base pair will get converted to A = U base pair. Since uracil is unmethylated, it will make the DNA molecule more susceptible to damage. If cytosine gets converted to uracil, the DNA double helix will become unstable as uracil cannot form base pairs with guanine.

30. Without knowing the exact bonding characteristics of hypoxanthine or xanthine, it may be difficult to predict the likelihood of each pairing type. It is likely that both are of the same class (purine or pyrimidine) because the names of the molecules indicate a similarity. In addition, the diameter of the structure is constant, which, under the model to follow, would be expected. In fact, hypoxanthine and xanthine are both purines.

Because there are equal amounts of A, T, and H, one could suggest that they are hydrogen bonded to each other; the same may be said for C, G, and X. Given the molar equivalence of erythrose and phosphate, an alternating sugar-phosphate-sugar backbone as in "earth-type" DNA would be acceptable. A model of a triple helix would be acceptable, since the diameter is constant. Given the chemical similarities to "earth-type" DNA, it is probable that the unique creature's DNA follows the same structural plan.

32. In comparing DNA migration to RNA, even though RNA molecules have the same charge-to-mass ratios, they can exist in a variety of shapes. Complementary intrastrand base pairing can make more compact structures compared to the more relaxed, open conformation. During electrophoresis, compact molecules migrate faster than relaxed, open structures. For electrophoretic size comparisons, RNA molecules must be denatured to eliminate secondary structural variables.

Chapter 10. Answers to Now Solve This

10-1. After one round of replication in the ^{14}N medium, the conservative scheme can be ruled out. After one round of replication in ^{14}N under a dispersive model, the DNA is of intermediate density, just as it is in the semiconservative model. However, in the next round of replication in ^{14}N medium, the density of the DNA is between the intermediate and "light" densities and therefore could be ruled out.

10-2. If the DNA contained parallel strands in the double helix and the polymerase were able to accommodate such parallel strands, there would be continuous synthesis and no Okazaki fragments. Several other possibilities exist. For example, if the DNA existed only as a single strand, the same results would occur.

Solutions to Problems and Discussion Questions

2. Your essay should describe replication as the process of making a daughter nucleic acids from existing ones. Synthesis refers to the precise series of steps, components, and reactions that allow such replication to occur.

4. By labeling the pool of nitrogenous bases of the DNA of *E. coli* with the heavy isotope ^{15}N, it would be possible to "follow" the "old" DNA.

6. Because the semiconservative scheme predicts that *half* of the DNA in each daughter double helix is labeled, it would be difficult to envision a scheme in which three strands are replicated in such a semiconservative manner. It would seem that either the conservative or dispersive scheme would fit more appropriately.

8. Several analytical approaches showed that the products of DNA polymerase I were probably copies of the template DNA. *Base composition* was used initially to compare both templates and products. Within experimental error, those data strongly suggested that the DNA replicated faithfully.

10. The first strain may show an inhibition to replication since the RNase may have destroyed the RNA primer that is necessary for the polymerase to continue with the replication. The second strain may have a mutation in the DNA polymerase that negates the requirement of a free 3′-OH group. An RNA primer would not be necessary in that case.

14. Given a stretch of double-stranded DNA, one could initiate synthesis at a given point and replicate strands either in one direction only (unidirectional) or in both directions (bidirectional). Notice that in the text the synthesis of complementary strands occurs in a *continuous* $5' > 3'$ mode on the leading strand in the direction of the replication fork, and in a *discontinuous* $5' > 3'$ mode on the lagging strand opposite the direction of the replication fork. Such discontinuous replication forms Okazaki fragments.

18. In bacteria, it is the ori sequence. They have a single origin of replication. In yeast, it is autonomously replicating sequence (ARS), which is an AT-rich region where the origin recognition complex (ORC) can bind and initiate replication at several locations. Origins in mammalian cells are unrelated to any specific DNA sequence. There are several origins on each chromosome where replication can start.

20. **(a)** No repair from DNA polymerase I and/or DNA polymerase III.
(b) No DNA ligase activity.
(c) No primase activity.
(d) Only DNA polymerase I activity.
(e) No DNA gyrase activity.

22. If replication is conservative, the first autoradiograms (see metaphase I in the text) would have label distributed on only one side (chromatid) of the metaphase chromosome.

Chapter 11. Answers to Now Solve This

11-1. By having a circular chromosome, no free ends present the problem of linear chromosomes, namely, complete replication of terminal sequences.

11-2. Since eukaryotic chromosomes are "multirepliconic" in that there are multiple replication forks along their lengths, one would expect to see multiple clusters of radioactivity.

11-3. Volume of the nucleus $= 4/3\pi r^3$

$= 4/3 \times 3.14 \times (5 \times 10^3\,\text{nm})^3 = 5.23 \times 10^{11}\,\text{nm}^3$

Volume of the chromosome $= \pi r^2 \times$ length

$= 3.14 \times 5.5\,\text{nm} \times 5.5\,\text{nm} \times (2 \times 10^9\,\text{nm})$
$= 1.9 \times 10^{11}\,\text{nm}^3$

Therefore, the percentage of the volume of the nucleus occupied by the chromatin is

$1.9 \times 10^{11}\,\text{nm}^3 / 5.23 \times 10^{11}\,\text{nm}^3 \times 100$
= about 36.3 percent

Solutions to Problems and Discussion Questions

2. Your essay should include a description of overall chromosomal configuration, such as linearity or circularity, as well as association with chromosomal proteins. In addition, it should describe higher level structures such as condensation in the case of eukaryotic chromosomes.

4. Polytene chromosomes are formed from numerous DNA replications, pairing of homologs, and absence of strand separation or cytoplasmic division. Each chromosome contains about 1000—5000 DNA strands in parallel register. They appear in specific tissues, such as salivary glands, of many dipterans such as *Drosophila*. They appear as comparatively long, wide fibers with sharp light and dark sections (bands) along their length. Such bands (chromomeres) are useful in chromosome identification, etc.

6. Lampbrush chromosomes are homologous pairs of chromosomes held together by chiasmata, with numerous loops of DNA protruding from a central axis of chromomeres. They are located in oocytes in the diplotene stage of the first prophase of meiosis.

8. Digestion of chromatin with endonucleases, such as micrococcal nuclease, gives DNA fragments of approximately 200 base pairs or multiples of such segments. X-ray diffraction data indicated a regular spacing of DNA in chromatin. Regularly spaced bead-like structures (nucleosomes) were identified by electron microscopy.

10. As chromosome condensation occurs, a 300-Å fiber is formed. It appears to be composed of five or six nucleosomes coiled together. Such a structure is called a solenoid. These fibers form a series of loops that further condense into the chromatin fiber and are then coiled into chromosome arms making up each chromatid.

14. SINE = short interspersed elements, a moderately repetitive sequence class; LINE = long interspersed elements. They are called "repetitive" because multiple copies exist.

16. In *E. coli*, the enzymes that control the number of supercoils in the double-stranded circular chromosome are called topoisomerase I and II. Topoisomerase I reduces the number of supercoils in the DNA, and topoisomerase II increases it by binding to the DNA, twisting it, cleaving both the strands, passing the end through the loop it has created, and reforming the phosphodiester bonds. The linking number is the number of turns in the DNA. In order to increase the number of supercoils by five, you would need to reduce the linking number by five.

18. The finding that natural chemical modification of nucleosomal components, as indicated in the question, increases gene activity suggests that changes in the binding of nucleosomes to DNA enable genes to be more accessible to factors that promote gene function. In addition, the finding that heterochromatin, containing fewer genes and more repressed genes, is undermethylated further supports the suggestion that histone modification is functionally related to changes in gene activity.

20. Dividing 3×10^9 base pairs by 10^6 gives an average of 3000 base pairs or 3 kb between *Alu* sequences.

Chapter 12. Answers to Now Solve This

12-1. (a) GGG = 3/4 × 3/4 × 3/4 = 27/64
GGC = 3/4 × 3/4 × 1/4 = 9/64
GCG = 3/4 × 1/4 × 3/4 = 9/64
CGG = 1/4 × 3/4 × 3/4 = 9/64

CCG = 1/4 × 1/4 × 3/4 = 3/64
CGC = 1/4 × 3/4 × 1/4 = 3/64
GCC = 3/4 × 1/4 × 1/4 = 3/64
CCC = 1/4 × 1/4 × 1/4 = 1/64

(b) Glycine: GGG and one G_2C (adds up to 36/64)
Alanine: one G_2C and one C_2G (adds up to 12/64)
Arginine: one G_2C and one C_2G (adds up to 12/64)
Proline: one C_2G and CCC (adds up to 4/64)

(c) Glycine: GGG, GGC
Alanine: CGG, GCC, CGC, GCG
Arginine: GCG, GCC, CGC, CGG
Proline: CCC,CCG

12-2. Because of a triplet code, a trinucleotide sequence will, once initiated, remain in the same reading frame and produce the same code all along the sequence regardless of the initiation site. If a tetranucleotide is used, such as ACGUACGUACGU . . . :

Codons:	ACG	UAC	GUA	CGU	ACG
Amino acids:	thr	tyr	val	arg	thr
	CGU	ACG	UAC	GUA	CGU
	arg	thr	tyr	val	arg
	GUA	CGU	ACG	UAC	GUA
	val	arg	thr	tyr	val
	UAC	GUA	CGU	ACG	UAC
	tyr	val	arg	thr	tyr

12-3. Apply complementary bases, substituting U for T:
(a) Sequence 1: 3′-GAAAAAACGGUA-5′
Sequence 2: 3′-UGUAGUUAUUGA-5′
Sequence 3: 3′-AUGUUCCCAAGA-5′
(b) Sequence 1: met-ala-lys-lys
Sequence 2: ser-tyr-[ter]
Sequence 3: arg-thr-leu-val
(c) Apply complementary bases: 3′-GAAAAAACGGTA-5′

Solutions to Problems and Discussion Questions

2. Your essay should include a description of the nature and structure of the genetic code, the enzymes and logistics of transcription, and the chemical nature of polymerization.

4. This sequence can be read as three possible repeating triplets—UUC, UCU, and CUU—depending on the initiation point. Hence, three different polypeptide homopolymers are produced, containing either phenylalanine (phe), serine (ser), or leucine (leu).

6. Given that AGG = arg, information from the AG copolymer indicates that AGA also codes for arg and that GAG must therefore code for glu. Coupling this information with that of the AAG copolymer, one can see that GAA must also code for glu and AAG must code for lys.

8. List the substitutions, then from the code table apply the codons to the original amino acids. Select codons that provide single base changes.

Original		Substitutions
threonine	⟵	alanine
_AC (U, C, A, or G)		_GC (U, C, A, or G)
glycine	⟵	serine
_GG (U or C)		_AG (U or C)
isoleucine	⟵	valine
_AU (U, C, or A)		_GU (U, C, or A)

10. The enzyme generally functions in the degradation of RNA; however, in an *in vitro* environment, with high concentrations of the ribonucleoside diphosphates, the direction of the reaction can be forced toward polymerization. *In vivo*, the concentration of ribonucleoside diphosphates is low and the degradative process is favored.

12. Applying the coding dictionary, the following sequences are "decoded":

Sequence 1: met-pro-asp-tyr-ser-(term)
Sequence 2: met-pro-asp-(term)

The 12th base (a uracil) is deleted from sequence #1, thereby causing a frameshift mutation that introduced a terminating triplet UAA.

14. GCU, GCC, GCA, and GCG—all code for alanine. Therefore, six single-base substitutions will result in an amino acid substitution at position 180. They are ACU, UCU, CCU, GAU, GUU, and GGU.

16. Leu: UUA, UUG, CUU, CUC, CUA, CUG. Hence, there is a preponderance of these codons.

Ala: GCU, GCC, GCA, GCG
Tyr: UAU, UAC

20. In an overlapping code, two amino acids will be affected; in a nonoverlapping code, one amino acid will be affected.

22.

Proline:	C_3 and one of the C_2A triplets
Histidine:	one of the C_2A triplets
Threonine:	one C_2A triplet and one A_2C triplet
Glutamine:	one of the A_2C triplets
Asparagine:	one of the A_2C triplets
Lysine:	A_3

24. **(a, b)** Alternative splicing occurs when pre-mRNAs are spliced in more than one way to yield various combinations of exons in the final mRNA product. Upon translation of a group of alternatively spliced mRNAs, a series of related proteins, called isoforms, are produced. It is likely that alternative splicing evolved to provide a variety of functionally related proteins in a particular tissue from one original source. Some tissues might be more prone to develop alternative splicing if they depend on a number of related protein functions.

Chapter 13. Answers to Now Solve This

13-1. One can conclude that the tRNA and not the amino acid is involved in recognition of the codon.

13-2. With the codes for valine being GUU, GUC, GUA, and GUG, single base changes from glutamic acid's GAA and GAG can cause the glu>>>val switch. The same can be said for lysine with its AAG codon. The normal glutamic acid is a negatively charged amino acid, whereas valine carries no net charge and lysine is positively charged. Given these significant charge changes, one would predict some, if not considerable, influence on protein structure and function.

Solutions to Problems and Discussion Questions

2. Your essay should include a general description of translation in which a functional ribosome, in conjunction with mRNA, orders amino acids and forms peptide bonds between them.

4. Amino acids are specifically and individually attached to the 3′ end of tRNAs, which possess a three-base sequence (the anticodon) that can base-pair with three bases of mRNA (codons). Messenger RNA contains a copy of the triplet codes, which are stored in DNA. The sequences of bases in mRNA interact, three at a time, with the anticodons of tRNAs. Enzymes involved in transcription include the following: RNA polymerase (*E. coli*) and RNA polymerase I, II, III (eukaryotes). Those involved in translation include the following: aminoacyl tRNA synthetases, peptidyl transferase, and GTPdependent release factors.

6. (a) The final output of a given gene may be influenced by the stability of an mRNA, and the stability of an mRNA is determined in part by its base content and sequence. (b) Differential splicing of mRNA (actually mRNA precursors) can influence how much of a given product will be made from a gene.

8. The different regions present in a tRNA molecule are the anticodon arm and loop, acceptor arm, amino acid binding site, D arm, TΨC arm, and variable loop.

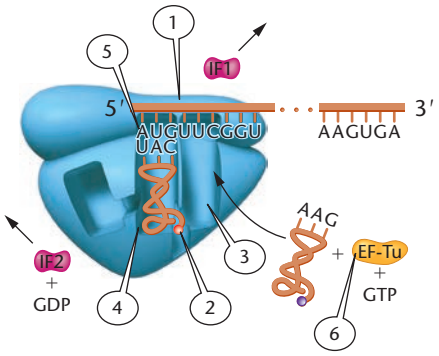
10. The quaternary level results from the associations of individual polypeptide chains.

14. Having the precise intragenic location of mutations as well as the ability to isolate the products, especially mutant products, allows scientists to compare the locations of lesions within genes. Mutations occurring nearer the initiation site in a gene will produce proteins with defects near the N-terminus. In this problem, the lesions cause chain termination; therefore, the nearer the mutations to the 5′ end of the mRNA, the shorter the polypeptide product. Relating the position of the mutation with the length of the product establishes the colinear relationship.

18. All of the substitutions involve one base change.

20. Because cross (a) is essentially a monohybrid cross, there would be no difference in the results if crossing over occurred (or did not occur) between the *a* and *b* loci.

22.



Chapter 14. Answers to Now Solve This

14-1. The phenotypic influence of any base change is dependent on a number of factors including, its location in coding or non-coding regions, its potential in dominance or recessiveness, and its interaction with other base sequences in the genome. If a base change is located in a non-coding region, there may be no influence on the phenotype, however, some non-coding regions in a traditional sense, may influence other genes and/or gene products.

14-2. If a gene is incompletely penetrant, it may be present in a population and only express itself under certain conditions. It is unlikely that the gene for hemophilia behaved in this manner. If a gene's expression is suppressed by another mutation in an individual, it is possible that offspring may inherit a given gene and not inherit its suppressor. Such offspring would have hemophila. Since all genetic variations must arise at some point, it is possible that the mutation in the Queen Victoria family was new, arising in a cell of the father. Lastly, given that the mother was heterozygous by chance, no other individuals in her family were unlucky enough to receive the mutant gene.

14-3. Any agent that inhibits DNA replication, either directly or indirectly, through mutation and/or DNA crosslinking, will suppress the cell cycle and may be useful in cancer therapy. Since guanine alkylation often leads to mismatched bases, they can often be repaired by a variety of mismatched repair mechanisms. However, DNA crosslinking can be repaired by recombinational mechanisms; thus, for such agents to be successful in cancer therapy, suppressors of DNA repair systems are often used in conjunction with certain cancer drugs.

14-4. Ethylmethane sulfonate (EMS) alkylates the keto groups at the 6^{th} position of guanine and at the 4^{th} position of thymine. In each case, base-pairing affinities are altered and transition mutations result. Altered bases are not readily repaired and once the transition to normal bases occurs through replication, such mutations avoid repair altogether.

Solutions to Problems and Discussion Questions

2. Your essay should include a brief description of the genomic differences between diploid and haploid organisms and, with the exception of phenomena such as cell death, disease, and cancer, mutational circumstances are attributable to both groups of organisms.

4. Since a somatic mutation first appears in a single cell, it is highly unlikely that the organism will be sufficiently altered to respond to a screen because none of the other cells in that organism will have the same mutation. That's not to say that somatic mutations can't influence the organism. Cancer cells generally originate from a single altered cell and can have a profound influence on the fate of an organism.

6. Diploid organisms have homologous chromosomes, so the wild type gene may compensate for the mutated gene. Haploid organisms have a single set of chromosomes that can be mutated easily with an observable phenotype.

8. Tautomeric shifts can result in mutations by allowing hydrogen bonding of normally noncomplementary bases so that incorrect nucleotide bases may be added during DNA replication.

10. Frameshift mutations are likely to change more than one amino acid in a protein product because as the reading frame is shifted, a different set of codons is generated. In addition, there is the possibility that a nonsense triplet could be introduced, thus causing premature chain termination.

12. When DNA is damaged, mutations are likely. In many cases, such mutations are deleterious to the health of the organism. Several mechanisms have evolved to reduce the impact of such mutations; cell-cycle arrest to quarantine a cell line or allow DNA repair and programmed cell death (apoptosis). If damaged DNA cannot be repaired through cell-cycle arrest, programmed cell death is often activated to rid the cell population of mutant cell lines.

14. The polymerase would encounter cytosines more frequently, which would be the nucleotide to be misincorporated more frequently.

16. *Xeroderma pigmentosum* is a form of human skin cancer caused by perhaps several rare autosomal genes, which interfere with the repair of damaged DNA. Since cancer is caused by mutations in several types of genes, interfering with DNA repair can enhance the occurrence of these types of mutations.

18. Mutations in *MutH*, *MutL*, and *MutS* in *E. coli* can adversely affect DNA mismatch repair. The equivalent genes in humans are *hMSH2* and *hMLH1*.

20. Replication slippage is a process that generates small deletions and insertions during DNA replication. While it can occur anywhere in the genome, it is most prevalent in regions already containing repeated sequences. Thus, repeated sequences are hypermutable.

22. Unscheduled DNA synthesis represents DNA repair. Complementation groups:

XP1	*XP4*	*XP5*
XP2		*XP6*
XP3		*XP7*

The groupings (complementation groups) indicate that there are at least three "genes" that form products necessary for unscheduled DNA synthesis. All of the cell lines that are in the same complementation group are defective in the same product.

26. It is probable that the IS occupied or interrupted normal function of a controlling region related to the *galactose* genes, which are in an operon with one controlling upstream element.

28. First, while less likely, one might suggest that transposons, for one reason or another, are more likely to insert in noncoding regions of the genome. One might also suggest that they are more stable in such regions. Second, and more likely, it is possible that transposons insert rather randomly and that selection eliminates those that have interrupted coding regions of the genome.

Chapter 15. Answers to Now Solve This

15-1. (a) It is likely that either premature termination of translation will occur (from the introduction of a nonsense triplet in a reading frame) or the normal chain termination will be ignored. Regardless, a mutant condition for the *Z* gene will be likely. If such a cell is placed on a lactose medium, it will be incapable of growth because β-galactosidase is not available.

(b) If the deletion occurs early in the *A* gene, one might expect impaired function of the *A* gene product, but it will not influence the use of lactose as a carbon source.

15-2. (a) With no lactose and no glucose, the operon is off because the *lac* repressor is bound to the operator and although CAP is bound to its binding site, it will not override the action of the repressor.

(b) With lactose added to the medium, the *lac* repressor is inactivated and the operon is transcribing the structural genes. With no glucose, the CAP is bound to its binding site, thus enhancing transcription.

(c) With no lactose present in the medium, the *lac* repressor is bound to the operator region, and since glucose inhibits adenyl cyclase, the CAP protein will not interact with its binding site. The operon is therefore "off."

(d) With lactose present, the *lac* repressor is inactivated; however, since glucose is also present, CAP will not interact with its binding site. Under this condition transcription is severely diminished and the operon can be considered to be "off."

15-3. Should hypermethylation occur in one of many DNA repair genes, the frequency of mutation would increase because the DNA repair system is compromised. The resulting increase in mutations might occur in tumor suppressor genes or proto-oncogenes.

15-4. General transcription factors associate with a promoter to stimulate transcription of a specific gene. Some *trans*-acting elements, when bound to enhancers, interact with coactivators to enhance transcription by forming an enhanceosome that stimulates transcription initiation. Transcription can be repressed when certain proteins bind to silencer DNA elements and generate repressive chromatin structures. The same molecule may bind to a different chromosomal regulatory site (enhancer or silencer), depending on the molecular environment of a given tissue type.

Solutions to Problems and Discussion Questions

2. Your essay should include a description of the evolutionary advantages of the efficient response to environmental resources and challenges (antibiotics, for example) when such resources are present. Having related functions in operons provides for coordinated responses.

4. See that under *negative* control, the regulatory molecule interferes with transcription, whereas in *positive* control, the regulatory molecule stimulates transcription. In an *inducible system*, the repressor that normally interacts with the operator to inhibit transcription is inactivated by an *inducer*, thus permitting transcription. In a *repressible system*, a normally inactive repressor is *activated* by a *corepressor*, thus enabling it (the activated repressor) to bind to the operator to inhibit transcription.

6. $I^+O^+Z^+$ = **inducible** because a repressor protein can interact with the operator to turn off transcription.

$I^-O^+Z^+$ = **constitutive** because the repressor gene is mutant; therefore, no repressor protein is available.

$I^+O^cZ^+$ = **constitutive** because even though a repressor protein is made, it cannot bind with the mutant operator.

$I^-O^+Z^+/F'I^+$ = **inducible** because even though there is one mutant repressor gene, the other I^+ gene, on the F factor, produces a normal repressor protein that is diffusible and capable of interacting with the operon to repress transcription.

$I^+O^cZ^+/F'O^c$ = **constitutive** because there is a constitutive operator (O^c) next to a normal *Z* gene. Constitutive synthesis of β-galactosidase will occur.

$I^sO^+Z^+$ = **repressed** because the product of the I^s gene is *insensitive* to the inducer lactose and thus cannot be inactivated. The repressor will continually interact with the operator and shut off transcription regardless of the presence or absence of lactose.

$I^sO^+Z^+/F'I^+$ = **repressed** because, as in the previous case, the product of the I^s gene is insensitive to the inducer lactose and thus cannot be inactivated. The repressor will continually interact with the operator and shut off transcription regardless of the presence or absence of lactose. The fact that there is a normal I^+ gene is of no consequence because once a repressor from I^s binds to an operator, the presence of normal repressor molecules will make no difference.

8. The mutations described are consistent with the structure of the lac repressor. The N-terminal portion of the repressor is involved in DNA binding, while the C-terminal portion is more involved in association with lactose and its analogs.

10. Generally, cooperative binding occurs when the final outcome is greater than the simple sum of its parts. In the case of transcription factors, each factor has little impact on transcription; however, when all components are present, a cooperative interaction (binding) occurs and a functional complex is made.

12. (a) Because activated CAP is a component of the cooperative binding of RNA polymerase to the *lac* promoter, absence of a functional *crp* would compromise the positive control exhibited by CAP.

(b) Without a CAP binding site there would be a reduction in the inducibility of the *lac* operon.

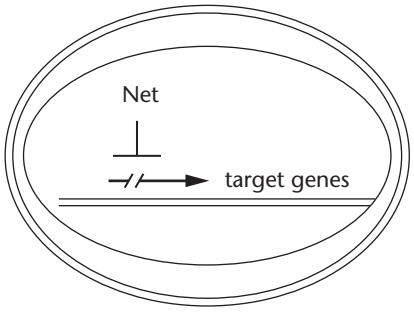
14. *X* is a repressor, *Y* is a protein that binds *X* and removes repression, and *A* induces the expression of *S* by binding *Y*.

16. Oil stimulates the production of a protein, which turns on (positive control) genes to metabolize oil. The different results in strains #2 and #4 suggest a *cis*-acting system. Because the operon by itself (when mutant as in strain #3) gives constitutive synthesis of the structural genes, a *cis*-acting system is also supported. The *cis*-acting element is most likely part of the operon.

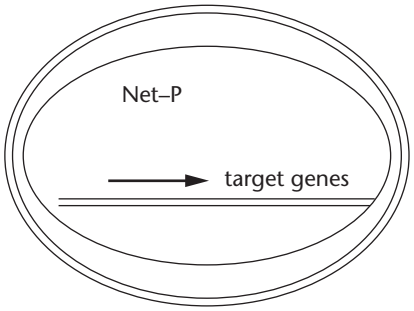
20. Transcription factors are transcription regulatory proteins that can increase or reduce the levels of transcription initiation. They bind promoters, enhancers, and silencers.

22. Generally, one determines the influence of various regulatory elements by removing necessary elements or adding extra elements. In addition, examining the outcome of mutations within such elements often provides insight as to function.

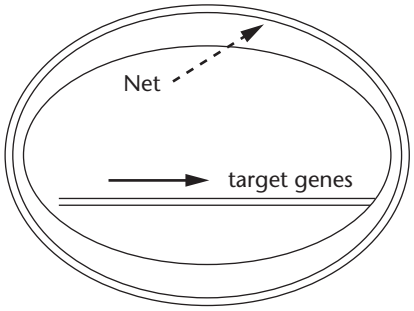
24.



Neutral Conditions



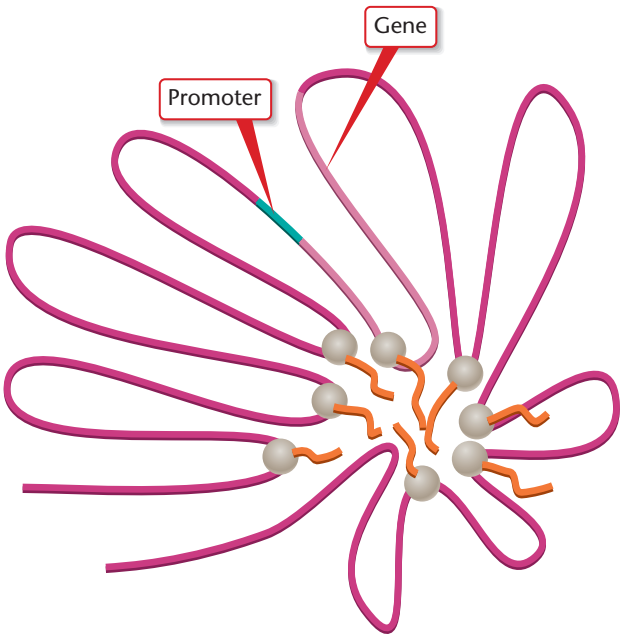
Phosphorylated Net



UV and Heat Shock

Sketches modified from Ducret et al. 1999 *Molecular and Cellular Biology* 19:7076–7087.

26. Following is a sketch of several RNA polymerase molecules (filled circles) in what might be a transcription factory. In this diagram there are eight RNA molecules shown being transcribed. Nascent transcripts are shown extending from the RNA polymerase molecules. For simplicity, only one promoter is shown and one structural gene is shown.



Chapter 16. Answers to Now Solve This

16-1. Being able to distinguish leukemic cells from healthy cells allows one to not only target therapy to specific cell populations, but it also allows for the quantification of responses to therapy. Because such cells produce a hybrid protein, it may be possible to develop a therapy, perhaps an immunotherapy, based on the uniqueness of the BCR/ABL protein.

16-2. *p53* is a tumor suppressor gene that protects cells from multiplying with damaged DNA. It is present in its mutant state in more than 50 percent of all tumors. Since the immediate control of a critical and universal cell cycle checkpoint is mediated by *p53*, mutation will influence a wide range of cell types. *p53*'s action is not limited to specific cell types.

16-3. Even if a major "cancer-causing" gene is transmitted, other genes, often new mutations, are usually necessary in order to drive a cell towards tumor formation. Full expression of the cancer phenotype is likely to be the result of interplay among a variety of genes and therefore show variable penetrance and expressivity.

Solutions to Problems and Discussion Questions

2. Your essay should describe the general influence of genetics in cancer. Since a variety of factors alter gene output and such output controls the cell cycle, it is likely that such factors could cause cancer.

4. Cancer cells do not require growth factors and proliferative signals to enter the mitotic cycle as opposed to normal cells. Negative regulatory signals that stop normal cells from proliferating are also not effective in cancer cells. One of the determinants of entry into the cell cycle is the G1 restriction point, controlled by the retinoblastoma protein, pRB. pRB is a tumor suppressor that is commonly mutated in certain cancers, thereby a block is lifted at the G1 entry. This causes growth-factor independency in cancer cells.

6. pRB-bound E2F transcription factors prevent the transcription of S-phase-specific early response genes (including cyclin D). pRB is released from E2Fs only in its hyperphosphorylated state, directing the cell-cycle entry. pRB is a major tumor suppressor.

8. Apoptosis is a natural process involved in morphogenesis and a protective mechanism against cancer formation. During apoptosis, nuclear DNA becomes fragmented, cellular structures are disrupted, and the cells are dissolved. Caspases are involved in the initiation and progress of apoptosis.

10. The nonphosphorylated form of pRB binds to transcription factors such as E2F, causing inactivation and suppression of the cell cycle. Phosphorylation of pRB activates the cell cycle by releasing transcription factors (E2F) to advance the cell cycle. With the phosphorylation site inactivated in the PSM-RB form, phosphorylation cannot occur, thereby leaving the cell cycle in a suppressed state.

12. Various kinases can be activated by breaks in DNA. One kinase, called ATM, and/or a kinase called Chk2 phosphorylates BRCA1 and p53. The activated p53 arrests replication during the S phase to facilitate DNA repair. The activated BRCA1 protein, in conjunction with BRCA2, mRAD51, and other nuclear proteins, is involved in repairing the DNA.

14. In the mutant state *oncogenes* induce or maintain uncontrolled cell division; that is, there is a gain of function. Generally, this gain of function takes the form of increased or abnormally continuous gene output. On the other hand, loss of function is generally attributed to mutations in tumor-suppressor genes, which function to halt passage through the cell cycle. When such genes are mutant, they have lost their capacity to halt the cell cycle. Such mutations are generally recessive.

16. It is less expensive, in terms of both human suffering and money, to seek preventive measures for as many diseases as possible. However, having gained some understanding of the mechanisms of disease, in this case cancer, it must also be stated that no matter what preventive measures are taken, it will be impossible to completely eliminate disease from the human population. It is extremely important, however, that we increase efforts to educate and protect the human population from as many hazardous environmental agents as possible.

18. The *p53* protein initiates several different responses to DNA damage including cell-cycle arrest followed by DNA repair and apoptosis if DNA cannot be repaired. Mutations in *p53* make cells unable to achieve this. As a result, they move unchecked through the cell cycle, regardless of the condition of the DNA. Therefore, cells lacking *p53* have high mutation rates, and accumulate those types of mutations that lead to cancer.

20. DNA lesions brought about by natural radiation (X rays, ultraviolet light), dietary substances, and substances in the external environment can lead to cancer. In addition, normal metabolism creates oxidative end products that can damage DNA, proteins, and lipids.

22. No, she will still have the general population risk of about 10 percent. In addition, it is possible that genetic tests will not detect all breast cancer mutations.

24. All cancer cells can proliferate to form tumors. However, if cells in the tumor also have the ability to break loose, enter the bloodstream, invade other tissues, and form secondary tumors (metastases), they become malignant.

26. As with many forms of cancer, a single gene alteration is not the only requirement. The authors (Bose et al.) state "but only infrequently do the cells acquire the additional changes necessary to produce leukemia in humans." Some studies indicate that variations (often deletions) in the region of the breakpoints may influence expression of CML.

28. (a, b) Even though there are changes in the *BRCA1* gene, they do not always have physiological consequences.

Such neutral polymorphisms make screening difficult in that one cannot always be certain that a mutation will cause problems for the patient.

(c) The polymorphism in *PM*2 is probably a silent mutation because the third base of the codon is involved.

(d) The polymorphism in *PM*3 is probably a neutral missense mutation because the first base is involved. However, because there is some first codon position degeneracy, it is possible for the mutation to be silent.

Chapter 17. Answers to Now Solve This

17-1. (a) Bacteria that have been transformed with the recombinant plasmid will be resistant to tetracycline, and therefore tetracycline should be added to the medium.

(b) Colonies that grow on a tetracycline medium only should contain the insert. Those bacteria that do not grow on the ampicillin medium probably contain the *Drosophila* DNA insert.

(c) Resistance to both antibiotics by a transformed bacterium could be explained in several ways. First, if cleavage with the *Pst*I was incomplete, then no change in biological properties of the uncut plasmids would be expected. Also, it is possible that the cut ends of the plasmid were ligated together in the original form with no insert.

17-2. Using the human nucleotide sequence, one can produce a probe to screen the library of the African okapi. Second, one can use the amino acid sequence and the genetic code to generate a complementary DNA probe for screening of the library. The probe is used, through hybridization, to identify the DNA that is complementary to the probe and can allow one to identify the library clone containing the DNA of interest. Cells with the desired clone are then picked from the original plate and the plasmid is isolated from the cells.

Solutions to Problems and Discussion Questions

2. Your essay should include an appreciation for the relative ease in which sections of DNA can be inserted into various vectors and the amplification and isolation of such DNA. You should also include the possibilities of modifying recombinant molecules.

4. Even though the human gene coding for insulin contains a number of introns, a cDNA generated from insulin mRNA is free of introns. Plasmids containing insulin genes (from cDNA) are free of introns, so no processing issue surfaces.

6. *Eco*RI should be used for the restriction site GAATTC. The complementary sequence is CTTAAG.

8. Plasmids were the first to be used as cloning vectors, and they are still routinely used to clone relatively small fragments of DNA. Because of their small size, they are relatively easy to separate from the host bacterial chromosome, and they have relatively few restriction sites. They can be engineered fairly easily (i.e., polylinkers and reporter genes added). BACs are artificial bacterial chromosomes that can be engineered for certain qualities such as carrying relatively large inserts.

YACs (yeast artificial chromosomes) contain telomeres, an origin of replication, and a centromere and are extensively used to clone DNA in yeast. With selectable markers (TRP1 and URA3) and a cluster of restriction sites, DNA inserts ranging from 100 kb to 1000 kb can be cloned and inserted into yeast. Since yeast, being a eukaryote, undergoes many of the typical RNA and protein processing steps of other, more complex eukaryotes, the advantages are numerous when working with eukaryotic genes.

10. No. The tumor-inducing plasmid (Ti) that is used to produce genetically modified plants is specific for the bacterium *Agrobacterium tumifaciens,* which causes tumors in many plant species. There is no danger that this tumor-inducing plasmid will cause tumors in humans.

12. The total number of molecules after 15 cycles would be 16,384, or $(2)^{14}$.

14. A cDNA library provides DNAs from RNA transcripts and is, therefore, useful in identifying what are likely to be functional DNAs. If one desires an examination of noncoding as well as coding regions, a genomic library would be more useful.

16. Assuming that one has knowledge of the amino acid sequence of the protein product or the nucleotide sequence of the target nucleic acid, a degenerate set of DNA strands can be prepared for cloning into an appropriate vector or amplified by PCR. A variety of labeling techniques can then be used, through hybridization, to identify complementary base sequences contained in the genomic library. One must know at least a portion of the amino acid sequence of the protein product or its nucleic acid sequence in order for the procedure to be applied. Some problems can occur through degeneracy in the genetic code (not allowing construction of an appropriate probe), the possible existence of pseudogenes in the library (hybridizations with inappropriate related fragments in the library), and variability of DNA sequences in the library due to introns (causing poor or background hybridization).

18. *Taq* polymerase is from a bacterium called *Thermus aquaticus,* which typically lives in hot springs. It is heat stable like some other enzymes used in PCR that are isolated from thermal vents in the ocean floor.

20. A knockout animal has a piece of DNA missing, whereas a transgenic animal usually has a piece of DNA added.

22. Until the host organism contains the knockout gene in the homozygous state in its sex cells, the knockout gene can not be faithfully transmitted at high frequency.

Chapter 18. Answers to Now Solve This

18-1. (a) To annotate a gene, one identifies gene-regulatory sequences found upstream of genes (promoters, enhancers, and silencers), downstream elements (termination sequences), and in-frame triplet nucleotides that are part of the coding region of the gene. In addition, 5′ and 3′ splice sites that are used to distinguish exons from introns as well as polyadenylation sites are also used in annotation.

(b) Similarity to other annotated sequences often provides insight as to a sequence's function and may serve to substantiate a particular genetic assignment. Direct sequencing of cDNAs from various tissues and developmental stages aids in verification.

(c) Taking an average of 20,000 for the estimated number of genes in the human genome and computing the percentage represented by 3141 gives 15.7 percent. It appears as if chromosome 1 is gene rich.

18-2. Since structural and chemical factors determine the function of a protein, it is likely to have several proteins share a considerable amino acid sequence identity, but not be functionally identical. Since the *in vivo* function of such a protein is determined by secondary and tertiary structures, as well as local surface chemistries in active or functional sites, the nonidentical sequences may have considerable influence on function. Note that the query matches to different site positions within the target proteins. A number of other factors suggesting different functions include associations with other molecules (cytoplasmic, membrane, or extracellular), chemical nature and position of binding domains, posttranslational modification, signal sequences, and so on.

18-3. Because blood is relatively easy to obtain in a pure state, its components can be analyzed without fear of tissue-site contamination. Second, blood is intimately exposed to virtually all cells of the body and may therefore carry chemical markers to certain abnormal cells it represents. Theoretically, it is an ideal probe into the human body. However, when blood is removed from the body, its proteome changes, and those

changes are dependent on a number of environmental factors. Thus, what might be a valid diagnostic for one condition might not be so for other conditions. In addition, the serum proteome is subject to change depending on the genetic, physiologic, and environmental state of the patient. Age and sex are additional variables that must be considered. Validation of a plasma proteome for a particular cancer would be strengthened by demonstrating that the stage of development of the cancer correlates with a commensurate change in the proteome in a relatively large, statistically significant pool of patients. Second, the types of changes in the proteome should be reproducible and, at least until complexities are clarified, involve tumorigenic proteins. It would be helpful to have comparisons with archived samples of each individual at a disease-free time.

Solutions to Problems and Discussion Questions

2. Your essay should include a description of traditional recombinant DNA technology involving cutting and splicing genes, as well as modern methods of synthesizing genes of interest, PCR amplification, microarray analysis, etc.

4. Whole-genome shotgun sequencing involves randomly cutting the genome into numerous smaller segments. Overlapping sequences are used to identify segments that were once contiguous, eventually producing the entire sequence. Difficulties in alignment often occur in repetitive regions of the genome. Map-based sequencing relies on known landmarks (genes, nucleotide polymorphisms, etc.) to orient the alignment of cloned fragments that have been sequenced. Compared to whole-genome sequencing, the map-based approach is somewhat cumbersome and time consuming. Whole-genome sequencing has become the most common method for assembling genomes, with map-based cloning being used to resolve the problems often encountered during whole-genome sequencing.

6. The main goals of the Human Genome Project are to establish, categorize, and analyze functions for human genes. As stated in the text:

> To analyze genetic variations between humans, including the identification of single-nucleotide polymorphisms (SNPs)
>
> To map and sequence the genomes of several model organisms used in experimental genetics, including *E. coli*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *M. musculus* (the mouse)
>
> To develop new sequencing technologies, such as high-throughput computer-automated sequencers, to facilitate genome analysis
>
> To disseminate genome information, both among scientists and the general public

8. One initial approach to annotating a sequence is to compare the newly sequenced genomic DNA to the known sequences already stored in various databases. The National Center for Biotechnology Information (NCBI) provides access to BLAST (Basic Local Alignment Search Tool) software, which directs searches through databanks of DNA and protein sequences. A segment of DNA can be compared to sequences in major databases such as GenBank to identify matches that align in whole or in part. One might seek similarities to a sequence on chromosome 11 in a mouse and find that or similar sequences in a number of taxa. BLAST will compute a similarity score or identity value to indicate the degree to which two sequences are similar. BLAST is one of many sequence alignment algorithms (RNA-RNA, protein-protein, etc.) that may sacrifice sensitivity for speed.

10. Because many repetitive regions of the genome are not directly involved in production of a phenotype, they tend to be isolated from selection and show considerable variation in redundancy. Length variation in such repeats is unique among individuals (except for identical twins) and, with various detection methods, provides the basis for DNA fingerprinting. Single-nucleotide polymorphisms also occur frequently in the genome and can be used to distinguish individuals.

12. The Personal Genome Project (PGP) provides individual sequences of diploid genomes, and results of such projects indicate that the HGP may underestimate genome variation by as much as fivefold. Genome variation between individuals may be 0.5 percent rather than the 0.1 percent estimated from the HGP. Since the PGP provides sequence information on individuals, fundamental questions about human diversity and evolution may be more answerable.

14. A number of new subdisciplines of molecular biology will provide the infrastructure for major advances in our understanding of living systems. The following terms identify specific areas within that infrastructure:

> proteomics—proteins in a cell or tissue
> metabolomics—enzymatic pathways
> glycomics—carbohydrates of a cell or tissue
> toxicogenomics—toxic chemicals
> metagenomics—environmental issues
> pharmacogenomics—customized medicine
> transcriptomics—expressed genes

Many other "-omics" are likely in the future.

16. Most microarrays, known also as gene chips, consist of a glass slide that is coated, using a robotic system, with single-stranded DNA molecules. Some microarrays are coated with single-stranded sequences of expressed sequenced tags or DNA sequences that are complementary to gene transcripts. A single microarray can have as many as 20,000 different spots of DNA, each containing a unique sequence. Researchers use microarrays to compare patterns of gene expression in tissues under different conditions or to compare gene-expression patterns in normal and diseased tissues. In addition, microarrays can be used to identify pathogens. Microarray databases allow investigators to compare any given pattern to others worldwide.

18. In general, one would expect certain factors (such as heat or salt) to favor evolution to increase protein stability: distribution of ionic interactions on the surface, density of hydrophobic residues and interactions, and number of hydrogen and disulfide bonds. As seen from examining the codon table, a high GC ratio would favor the amino acids Ala, Gly, Pro, Arg, and Trp and minimize the use of Ile, Phe, Lys, Asn, and Tyr. How codon bias influences actual protein stability is not yet understood. Most genomic sequences change by relatively gradual responses to mild selection over long periods of time. They strongly resemble patterns of common descent; that is, they are conserved. Although the same can be said for organisms adapted to extreme environments, extraordinary physiological demands may dictate unexpected sequence bias.

20. (a, b) With exon sequencing, one may get lucky and find an issue within a gene that has relevance. Given the multitude of genetic variations among individuals, this might be similar to finding a needle in a haystack. Many significant genomic functions are regulated outside the exon pool. Introns and mitochondria are highly variable among individuals, but may have some relevance in some health-related conditions.

Chapter 19. Answers to Now Solve This

19-1. Antigens are usually quite large molecules, and in the process of digestion, they are sometimes broken down into smaller molecules, thus becoming ineffective in stimulating the immune system. Some individuals are allergic to the food they eat, testifying to the fact that all antigens are not completely degraded or modified by digestion. In some cases, ingested antigens do indeed stimulate the immune

system (e.g., oral polio vaccine) and provide a route for immunization. Localized (intestinal) immunity can sometimes be stimulated by oral introduction of antigens, and in some cases, this can offer immunity to ingested pathogens.

19-2. It will hybridize by base complementation to the normal DNA sequence.

Solutions to Problems and Discussion Questions

2. Your essay should include a description of genomic applications that relate to agriculture, health and welfare, scientific exploration and appreciation of the earth's flora and fauna, etc. In addition, areas of patent protection, personal privacy, and potential agricultural and environmental hazards should be addressed.

4. From a purely scientific viewpoint, there will be no added danger to consuming cow's milk from cloned animals. However, some individuals may have an aversion to organismic cloning, and supporting such activities through consumption of products of cloned organisms may be viewed negatively on moral grounds. It is likely that public sentiment will pressure for labeling of "cloned products" on the grounds that consumers should be able to make an informed choice as to the origin of such products.

6. The Venter team compared a number of genomes each with a small number of genes and identified 256 genes that may represent the minimum number of genes for life. The team also used transposon-based mutations to determine the number of genes essential for life. Finally, it synthesized short DNA segments and assembled them into a synthetic genome that possessed characteristics of living systems. Recipient bacteria having different characteristics indicated that genome conversion had occurred.

8. Since both mutations occur in the CF gene, children who possess both alleles will suffer from CF. With both parents heterozygous, each child born will have a 25 percent chance of developing CF.

10. Using restriction enzyme analysis to detect point mutations in humans is a tedious trial-and-error process. Given the size of the human genome in terms of base sequences and the relatively low number of unique restriction enzymes, the likelihood of matching a specific point mutation, separate from other normal sequence variations, to a desired gene is low.

12. A GWAS analyzes SNPs, specific differences in genes, CNVs, or changes in the epigenome, such as methylation patterns in particular regions of a chromosome. By determining which SNPs, CNVs, or epigenome changes co-occur in individuals with the disease, scientists can calculate the disease risk associated with each variation.

14. In the case of haplo-insufficient mutations, gene therapy holds promise; however, in "gain-of-function" mutations, in all probability the mutant gene's activity or product must be compromised. Addition of a normal gene probably will not help unless it can compete out the mutant gene product.

16. Certainly, information provided to physicians and patients about genetic testing is a strong point in favor of wide distribution. It would probably be helpful for companies involved in genetic testing to also participate by providing information peculiar to their operations. It would be necessary, however, that any individual results from tests would be held in strict confidence. It would be helpful if pooled statistical data would be available to the public in terms of frequencies of false positives and negatives, as well as population and/or geographical distributions.

18. It is a personal decision to have one's genome sequenced, but in doing so one must be armed with information as to the expected variability of the genome and the possibility of false positives. A publicly available genome might lead to employment bias, changes in personal relationships, and so on.

20. Raw genomic information is difficult to interpret. 23andMe now provides only ancestry information and has stopped giving health-related data. In November 2013, the FDA warned 23andMe to stop marketing its genomic service. However, while not reversing its 2013 decision, 23andMe has recently (February, 2015) received authorization to market a specific personal test for Bloom syndrome.

22. (a) Since a gene is a product of the natural world, it does not conform to section 101 of U.S. patent laws, which govern patentable matter.

(b) Since both the direct-to-consumer test for the *BRCA1* and *BRCA2* genes and Venter's "first-ever human-made life form" are original in their process or development, they should be patentable. However, the *BRCA1* and *BRCA2* genes are works of nature, and the genes themselves should not be patentable.

Chapter 20. Answers to Now Solve This

20-1. It is possible that your screen was more inclusive, that is, it identified more subtle alterations than the screen of others. You may have identified several different mutations (multiple alleles) in some of the same genes.

20-2. Because in *ftz/ftz* embryos, the engrailed product is absent and in *en/en* embryos *ftz* expression is normal, one can conclude that the *ftz* gene product regulates, either directly or indirectly, *en*. Because the *ftz* gene is expressed normally in *en/en* embryos, the product of the *engrailed* gene does not regulate expression of *ftz*.

20-3. Since *her-1*$^{-}$ mutations cause males to develop into hermaphrodites, and *tra-1*$^{-}$ mutations cause hermaphrodites to develop into males, one may hypothesize that the *her-1*$^{+}$ gene produces a product that suppresses hermaphrodite development, while the *tra-1*$^{+}$ gene product is needed for hermaphrodite development.

Solutions to Problems and Discussion Questions

2. Your essay should describe the overall development in both plants and animals as being dependent on families of genes that are controlled by regulatory elements. Describe elements that evolved in plants and animals given their independent evolution.

4. The syncytial blasterm is formed as nuclei migrate to the egg's outer margin or cortex, where additional divisions take place. Plasma membranes organize around each of the nuclei at the cortex, thus creating the cellular blastoderm.

6. Maternal genes regulate the expression of first three groups of zygotic genes—gap, pair-rule, and segment polarity genes. Mutations in zygotic genes have embryo-lethal phenotypes.

8. Because the polar cytoplasm contains information to form germ cells, one would expect such a transplantation procedure to generate germ cells in the anterior region.

10. This experiment will include designing suitably tagged antibodies that will bind the protein and react with a substrate to produce a colored or chromogenic product that can be visualized using microscopy or imaging techniques.

12. A dominant gain-of-function mutation is one that changes the specificity or expression pattern of a gene or gene product. The "gain-of-function" *Antp* mutation causes the wild type *Antennapedia* gene to be expressed in the eye-antenna disc and mutant flies have legs on the head in place of antenna.

14. Given the information in the problem, it is likely that this gene normally controls the expression of *BX-C* genes in all body segments. The wild type product of *esc* stored in the egg may be required to interpret the information correctly stored in the egg cortex.

16. The *Polycomb* gene family induces changes in chromatin that influence *Hox* gene expression. A gene in *Arabidopsis* has significant homology to the *Polycomb* gene family and also works by altering chromatin structure. The cross reactivity is thus related to the Polycomb product's effect on chromatin. Such

parallel functions indicate that mechanisms of regulation are conserved over vast evolutionary distances.

Chapter 21. Answers to Now Solve This

21-1. (a) Since 1/256 of the F_2 plants are 20 cm and 1/256 are 40 cm, there must be 4 gene pairs involved in determining flower size.

(b) Since there are nine size classes, one can conduct the following backcross: *AaBbCcDd* × *AABBCCDD*. The frequency distribution in the backcross would be

1/16 = 40 cm	4/16 = 32.5 cm
4/16 = 37.5 cm	1/16 = 30 cm
6/16 = 35 cm	

21-2. (a) Taking the sum of the values and dividing by the number in the sample gives the following means:

mean sheep fiber length = 7.7 cm
mean fleece weight = 6.4 kg

The variance for each is

variance sheep fiber length = 6.097
variance fleece weight = 3.12

The standard deviation is the square root of the variance:

sheep fiber length = 2.469
fleece weight = 1.766

(b, c) The covariance for the two traits is 30.36/7, or 4.34, while the correlation coefficient is +0.998.

(d) There is a very high correlation between fleece weight and fiber length and it is likely that this correlation is not by chance. Even though correlation does not mean cause-and-effect, it would seem logical that as you increased fiber length, you would also increase fleece weight. It is probably safe to say that the increase in fleece weight is directly related to an increase in fiber length.

21-3. The role of genetics and the role of the environment can be studied by comparing the expression of traits in monozygotic and dizygotic twins. The higher concordance value for monozygotic twins as compared with the value for dizygotic twins indicates a significant genetic component for a given trait. Notice that for traits including blood type, eye color, and mental retardation, there is a fairly significant difference between MZ and DZ groups. However, for measles, the difference is not as significant, indicating a greater role of the environment. Hair color has a significant genetic component as do idiopathic epilepsy, schizophrenia, diabetes, allergies, cleft lip, and club foot. The genetic component to mammary cancer is present but minimal according to these data.

Solutions to Problems and Discussion Questions

2. Your essay should include a description of various ratios typical of Mendelian genetics as compared with the more blending, continuously varying expressions of neo-Mendelian modes of inheritance. It should contrast discontinuous inheritance and continuous patterns.

4. (a) There are two alleles at each locus for a total of four alleles.

(b, c) We can say that each gene (additive allele) provides an equal unit amount to the phenotype and the colors differ from each other in multiples of that unit amount. The number of additive alleles needed to produce each phenotype is given below.

1/16 = dark red	= *AABB*
4/16 = medium–dark red	= *2AABb, 2AaBB*
6/16 = medium red	= *AAbb, 4AaBb, aaBB*
4/16 = light red	= *2aaBb, 2Aabb*
1/16 = white	= *aabb*

(d) F_1 = all light red
F_2 = 1/4 medium red 2/4 light red 1/4 white

6. (a, b) There are four gene pairs involved.

(c) Since there is a difference of 24 cm between the extremes, 24 cm/8 = 3 cm for each increment (each of the additive alleles).

(d) A typical F_1 cross that produces a "typical" F_2 distribution would be where all gene pairs are heterozygous (*AaBbCcDd*), independently assorting, and additive. There are many possible sets of parents that would give an F_1 of this type. The limitation is that each parent has genotypes that give a height of 24 cm as stated in the problem. Because the parents are inbred, it is expected that they are fully homozygous. An example is

AABBccdd × *aabbCCDD*

(e) Since the *aabbccdd* genotype gives a height of 12 cm and each uppercase allele adds 3 cm to the height, there are many possibilities for an 18 cm plant:

AAbbccdd, *AaBbccdd*, *aaBbCcdd*, etc.

Any plant with seven uppercase letters will be 33 cm tall:

AABBCCDd, *AABBCcDD*, *AABbCCDD*, for example.

8. For height, notice that average differences between MZ twins reared together (1.7 cm) and those MZ twins reared apart (1.8 cm) are similar (meaning little environmental influence) and considerably less than differences of DZ twins (4.4 cm) or sibs (4.5 cm) reared together. These data indicate that genetics plays a major role in determining height.

However, for weight, notice that MZ twins reared together have a much smaller (1.9 kg) difference than MZ twins reared apart, indicating that the environment has a considerable impact on weight. By comparing the weight differences of MZ twins reared apart with DZ twins and sibs reared together one can conclude that the environment has almost as much an influence on weight as genetics.

For ridge count, the differences between MZ twins reared together and those reared apart are small. For the data in the table, it would appear that ridge count and height have the highest heritability values.

10. Monozygotic twins are derived from the splitting of a single fertilized egg and are therefore of identical genetic makeup. When such twins are raised in the same versus different settings, an estimate of relative hereditary and environmental influences can often be made.

12. (a) Using the following equations, H^2 and h^2 can be calculated as follows.

For back fat: Broad-sense heritability = H^2 = 12.2/30.6 = .398
Narrow-sense heritability = h^2 = 8.44/30.6 = .276

For body length: Broad-sense heritability = H^2 = 26.4/52.4 = .504
Narrow-sense heritability = h^2 = 11.7/52.4 = .223

(b) Of the two traits, selection for back fat would produce more response.

14. (a) For vitamin A: $h_A^2 = V_A/V_P = V_A/(V_E + V_A + V_D) = 0.097$
For cholesterol: $h_A^2 = 0.223$

(b) Cholesterol content should be influenced to a greater extent by selection.

16. Using the equation $h^2 = (M2 - M)/(M1 - M)$, $0.25 = (M2 - 20)/(24 - 20) = 21\,g$

18. Using the equation $h^2 = (M2 - M)/(M1 - M)$, $0.2 = (M2 - 52)/(61 - 52)$, $M2 = (0.2 \times 9) + 52 = 53.8$"

20. (a) There are two ways to answer this section, a hard way and an easy way. The hard way would to take a big sheet of paper, make the cross (*AaBbCcDdEeFf* × *AaBbCcDdEeFf*), collect the genotypes, and calculate the ratios.

This method would be very laborious and error-prone. The easy way would be to re-read the material on the binomial expansion and note the pattern preceding

each expression. Notice that all numbers other than the 1's are equal to the sum of the two numbers directly above them. By enlarging the numbers to include six gene pairs, you can arrive at the thirteen classes and their frequencies

3" = 1 4" = 12 5" = 66
6" = 220 7" = 495 8" = 792
9" = 924 10" = 792 11" = 495
12" = 220 13" = 66 14" = 12
15" = 1

(b) 3" = 1 4" = 6 5" = 15
6" = 20 7" = 15 8" = 6
9" = 1

22. The level of blood sugar varies considerably from individual to individual, day to day, and hour to hour, and on a population level, it displays continuous variation. However the diagnosis of Type II diabetes is set by relatively fixed criteria. A fasting blood sugar level of 126 mg/dL or higher, repeated on different days, is diagnostic of diabetes. A casual (non-fasting) blood sugar level of 200 mg/dL or higher is suggestive of diabetes. In either case, while the level of blood sugar is influenced by a variety of factors (polygenic and environmental), the actual diagnosis of the disease leads one to be classified as diabetic or not diabetic. Since there are only two phenotypic classes (or three if one included the prediabetic state), diabetes is referred to as a threshold trait.

Chapter 22. Answers to Now Solve This

22-1. Because the alleles follow a dominant/recessive mode, one can use the equation $\sqrt{q^2}$ to calculate *q*, from which all other aspects of the answer depend. The frequency of *aa* types is determined by dividing the number of nontasters (37) by the total number of individuals (125).

$$q^2 = 37/125 = 0.296$$
$$q = 0.544$$
$$p = 1 - q$$
$$p = 0.456$$

The frequencies of the genotypes are determined by applying the formula $p^2 + 2pq + q^2$ as follows:

Frequency of $AA = p^2 = (0.456)^2 = 0.208$ or 20.8%
Frequency of $Aa = 2pq = 2(0.456)(0.544) = 0.496$ or 49.6%
Frequency of $aa = q^2 = (0.544)^2 = 0.296$ or 29.6%

22-2. (a) For the *CCR5* analysis, first determine *p* and *q*. Since one has the frequencies of all the genotypes, one can add 0.6 and 0.351/2 to provide *p* (= .7755); *q* will be 0.049 and .351/2 = 0.2245

The equilibrium values will be as follows

Frequency of $l/l = p^2 = (.7755)^2 = .6014$ or 60.14%
Frequency of $l/\Delta 32 = 2pq = 2(.7755)(.2245)$
$= .3482$ or 34.82%
Frequency of $\Delta 32/\Delta 32 = q^2 = (.2245)^2$
$= .0504$ or 5.04%

Comparing these equilibrium values with the observed values strongly suggests that the observed values are drawn from a population in Hardy-Weinberg equilibrium.

(b) For the *AS* (sickle-cell) analysis, first determine *p* and *q*. Since one has the frequencies of all the genotypes, one can add .756 and .242/2 to provide *p* (= .877); *q* will be 1 − .877 or .123.

The equilibrium values will be as follows:

Frequency of $AA = p^2 = (.877)^2 = .7691$ or 76.91%
Frequency of $AS = 2pq = 2(.877)(.123)$
$= .2157$ or 21.57%
Frequency of $SS = q^2 = (.123)^2 = .0151$ or 1.51%

Comparing these equilibrium values with the observed values suggests that the observed values may be drawn from a population that is not in equilibrium. Notice that there are more heterozygotes than predicted, and fewer *SS* types in the population. Since data are given in percentages, χ^2 values can not be computed.

22-3. Given that the recessive allele *a* is present in the homozygous state (q^2) at a frequency of 0.0001, the value of *q* is 0.01 and $p = 0.99$.
(a) *q* is 0.01.
(b) $p = 1 - q$ or 0.99
(c) $2pq = 2(0.01)(0.99) = 0.0198$ (or about 1/50)
(d) $2pq \times 2pq$
$= 0.0198 \times 0.0198 = 0.000392$ (or about 1/255)

22-4. The probability that the woman (with no family history of CF) is heterozygous is $2pq$ or 2(1/50)(49/50). The probability that the man is heterozygous is 2/3. The probability that a child with CF will be produced by two heterozygotes is 1/4. Therefore, the overall probability of the couple producing a CF child is $98/2500 \times 2/3 \times 1/4 = 0.00653$, or about 1/153.

Solutions to Problems and Discussion Questions

2. Your essay should include a discussion of the original sources of variation coming from mutation and that migration can cause gene frequencies to change in a population if the immigrants have different gene frequencies compared to the host population. You should also describe selection as resulting from the biased passage of gametes and offspring to the next generation.

4. There must be evidence that gene flow does not occur among the groups being called different species. Classifications above the species level (genus, family, etc.) are not based on such empirical data. Indeed, classification above the species level is somewhat arbitrary and based on traditions that extend far beyond DNA sequence information. In addition, recall that DNA sequence divergence is not always directly proportional to morphological, behavioral, or ecological divergence. While the genus classifications provided in this problem seem to be invalid, other factors, well beyond simple DNA sequence comparison, must be considered in classification practices.

6. Calculate *p* and *q*, then apply the equation $p^2 + 2pq + q^2$ to determine genotypic frequencies in the next generation.

p = frequency of $A = 0.2 + 0.3 = 0.5$
$q = 1 - p = 0.5$
Frequency of $AA = p^2 = 0.25$ or 25%
Frequency of $Aa = 2pq = 0.5$ or 50%
Frequency of $aa = q^2 = 0.25$ or 25%

The initial population was not in equilibrium; however, after one generation of mating under the Hardy–Weinberg conditions the population is in equilibrium and will continue to be so (and not change) until one or more of the Hardy–Weinberg conditions is not met. Note that *equilibrium* does not necessarily mean *p* and *q* equal 0.5.

8. In order for the Hardy–Weinberg equations to apply, the population must be in Hardy–Weinberg equilibrium.

10. Yes, it is in equilibrium. The distribution will be the same in the next generation.

12. $M = 0.6; N = 0.4$

14. 18% of the population will be heterozygous.

18. The approximate similarity of mutation rates among genes and lineages should provide more credible estimates of divergence times of species and allow for broader interpretations of sequence comparisons. It also provides for increased understanding of the mutational processes that govern evolution among mammalian genomes. For instance, if the rate

of mutation is fairly constant among lineages or cells that have a more rapid turnover, it indicates that replication-related errors do not make a significant contribution to mutation rates.

22. Because of degeneracy in the code, there are some nucleotide substitutions, especially in the third base, that do not change amino acids. In addition, if there is no change in the overall charge of the protein, it is likely that electrophoresis will not separate the variants. If a positively charged amino acid is replaced by an amino acid of like charge, then the overall charge on the protein is unchanged. The same may be said for other negatively charged and neutral amino acid substitutions.

24. In general, speciation involves the gradual accumulation of genetic changes to a point where reproductive isolation occurs. Depending on environmental or geographic conditions, genetic changes may occur slowly or rapidly. They can involve point mutations or chromosomal changes.

26. **(a,b)** The pattern of genetic distances through time indicates that from the present to about 25,000 years ago, modern humans and Cro-Magnons show an approximately constant number of differences. Conversely, there is an abrupt increase in genetic distance seen in comparing modern humans and Cro-Magnons with Neanderthals. The results indicate a clear discontinuity among modern humans, Cro-Magnons, and Neanderthals with respect to genetic variation in the mitochondrial DNAs sampled. Assuming that the sampling and analytical techniques used to generate the data are valid, it appears that Neanderthals made little, if any, genetic contributions to the Cro-Magnon or modern European gene pool. It could be argued that the absence of Neanderthal mtDNA lineages in living humans is a consequence of random drift or lineage extinction since the disappearance of Neanderthals. However, the examination of ancient Cro-Magnon mtDNA shows no evidence of a historical relationship and suggests that Neanderthals were not genetically related to the ancestors of modern humans.

Special Topic 1

Review Question Answers

2. In general, periodic methylation occurs at CpG-rich regions and promoter sequences. When a gene is imprinted by methylation, it remains transcriptionally silent. In a mammalian embryo, imprinting may silence only the paternal set of chromosomes, for example.

4. Reversible histone modifications influence the structure of chromatin by altering the accessibility of nucleosomes to the transcriptional machinery. These chromatin alterations "open" or "close" genes for transcription.

6. Imprinting usually involves certain genes, restricted in number, that are altered by passage through meiosis. A maternally derived imprint or a paternally derived imprint may occur. Imprinted alleles are transcriptionally silent in all cells of the organism, whereas epigenetic modifications (methylation) can be reactivated by environmental signals.

8. In addition to functioning in cellular signaling, microRNAs play a significant role in the developing embryo. MiRNAs are involved with RNA silencing through RISCs that act as repressors of gene expression. They do so by making mRNAs less likely to be translated.

Discussion Question Answers

2. While data are scant, some studies have shown that children born after *in vitro* fertilization are at risk for low to very low birth weight that may have resulted from abnormal imprinting. There also appears to be an increased risk of the ART child having Beckwith-Wiedemann syndrome. Given these data, it would seem reasonable that such information should be provided to prospective parents of an ART child. Each couple would need to reach a decision based on available science and their own value and belief sets.

4. Plant miRNAs are known to downregulate gene expression and some foods are the source of miRNAs that circulate in body fluids of humans. Given this information, it has been suggested that as yet poorly understood environmental factors may play a significant role in the regulation of gene function in humans. At this point it might be premature to design a dietary regimen based on such a frail understanding of the role of plant miRNAs in humans.

Special Topic 2

Review Question Answers

2. Since RNA can both serve in information storage and transfer and catalyze reactions, it has been hypothesized that RNA was the precursor to molecular life-like events. In addition, RNAs are components of many primitive yet biologically significant reactions.

4. DNA methylation provides a defense to the integration of foreign DNA into the bacterial chromosome, whereas *CRISPR* loci transcribe crRNAs that guide nucleases to invading complementary DNAs in order to destroy them.

6. Through a series of transcriptive and Dicer-related activities, siRNAs are formed that are complementary to centromeric DNA. A RITS silencing complex forms that leads to methylation, thus triggering heterochromatin formation. At this point, the evolution of such a complex process of heterchromatin formation is not well understood.

8. Different regions within cells are often associated with specialized functions. Some mRNAs are specific for products that are exported while some are destined for intracellular functions. To achieve various cellular functions, localization is required.

Discussion Question Answers

2. Negative or positive regulation depends on whether the ribosome binding site is masked or available. When repression occurs, the RBS is masked by sRNA. When positive regulation occurs, sRNA pairing unmasks the RBS.

4. In bacteria and Archaea, foreign DNA can be inserted into *CRISPR* loci in the genome, which brings about transcription of crRNAs that guide nucleases to invading complementary DNAs. In addition, foreign DNA can be digested by restriction endonucleases. One form of eukaryotic genome protection involves piRNAs that are pivotal in silencing transposons, mobile DNA sequences that change position. Associated with Piwi proteins, certain proteins (such as RNA-induced silencing complexes [RISC]) target transposon-derived RNAs and their complementary sequences. This process represses transposon transcription by promoting DNA methylation of transposon DNA. A broadly functioning protective mechanism involves siRNA in association with RISC and Dicer, an RNase III protein.

6. Even though a particular species of mRNA may be fairly uniformly distributed throughout a cell, it does not follow that it is uniformly translated. It is likely that different domains reside in cells that contain different translational signals. If an mRNA finds itself in a particular molecular environment, it may be destined for translation, whereas that same mRNA in another part of a cell may not have the environmental stimulation necessary for translation.

Special Topic 3

Review Question Answers

2. With the development of the polymerase chain reaction, trace samples of DNA can be used, commonly in forensic applications. STRs are like VNTRs, but the repeat portion is shorter, between two and nine base pairs, repeated from 7 to 40 times. A core set of STR loci, about 13, is most often used in forensic applications.

4. Since males typically contain a Y chromosome (exceptions include transgender and mosaic individuals), gender separation of a mixed tissue sample is easily achieved by Y chromosome profiling. In addition, STR profiling is possible for over 200 loci; however, because of the relative stability of DNA in the Y chromosome, it is difficult to differentiate between DNA from fathers and sons or male siblings.

6. Like Y chromosome DNA, mtDNA is relatively stable because it undergoes very little, if any, recombination. Since there is a high copy number of mitochondria in cells, it is especially useful in situations where samples are small, old, or degraded, which is often the case in catastrophes.

8. The Combined DNA Index System (CODIS) is a collection of DNA databases and analytical tools of both state and federal governments, maintained by the FBI. DNA profiles are collected from convicted offenders, forensic investigations, and in some states, those suspected of crimes as well as from unidentified human remains and missing persons (in cases where DNA is available).

10. The prosecutor's fallacy attempts to equate guilt with a numerical probability produced by a single piece of evidence. Just because a match occurs between a crime scene and a suspect, it does not mean that the suspect is guilty. Human error, contamination, or evidence tampering all contribute to the complexities of interpreting DNA profiling data.

Discussion Question Answers

2. To gain information as to laws and regulations in various states, one could navigate to "Welcome to the DNA Laws Database" within the National Conference of State Legislatures website. There, one can select a particular state for its laws and regulations regarding DNA collection and profiling. In general, one will see that most states contain descriptions of the following topics:
(a) various DNA databases used
(b) methods of DNA collection
(c) post-conviction DNA collection of felons
(d) oversight and advisory committees
(e) convicted offender statutes

4. Somatic mosaicism and chimerism involve a mixture of cell types, the origin of which may involve a variety of embryonic events, some of which are understood. Since a single individual may contain a mixed population of cells, a DNA sample taken from one tissue site may not match a DNA sample taken from another site. This can lead to a conflicted set of results when it comes to matching a DNA sample to a sample of DNA from a crime scene. Taking DNA samples from various sites on an individual may be useful in mitigating such confusion. In addition, in STR DNA profiling, mosaicism may present itself at the electrophoresis/analysis stage by additional peaks or peak height imbalances.

Special Topic 4

Review Question Answers

2. Herceptin is used in the treatment of breast cancer that targets the epidermal growth factor receptor 2 (*HER-2*) gene located on chromosome 17. Overexpression of this gene occurs in about 25 percent of invasive breast cancer cases. Herceptin is a monoclonal antibody that binds specifically to inhibit the HER-2 receptor.

4. Cytochrome P450 is composed of a family of enzymes that are encoded by 57 different genes. Certain variants of cytochrome P450 metabolize drugs slowly and can lead to harmful accumulations of a drug. Other variants cause drugs to be eliminated quickly, which can lead to drug ineffectiveness. A pivotal gene, *CYP2D6*, influences the metabolism of approximately 25 percent of all drugs, while *VKORC1* influences the response to warfarin, an anticoagulant drug.

6. Recently, large-scale genomic sequencing has shown that each tumor is genetically unique. With such information, it is often possible to provide a personalized diagnosis and possibly apply personalized treatments. One such example is the use of Herceptin for the treatment of breast cancer; another is the use of Erbitux and Vectibix to inhibit epidermal growth factor receptors that are commonly expressed in cancer cells.

8. Using the search function in the PharmGKB database one can find a number of references that discuss the variants of *CYP2D6* and tamoxifen. For example, according to Hertz et al. (Hertz, D. et al. 2012. *The Oncologist.* 17(5): 2011-0418), tamoxifen efficacy is dependent on the highly polymorphic cytochrome P450 gene (*CYP2D6*). Depending on a particular variant genotype, tamoxifen treatment outcome is highly inconsistent. The entire Hertz et al. paper is available through the PharmGKB database and provides a complete and detailed description of the interactions of *CYP2D6* variants and tamoxifen.

Discussion Question Answers

2. There are several bridges that must be crossed before one can claim universal use and acceptance. First, it will be necessary to close the gap between data collection and interpretation of complex interactions. Second, personalized medicines are dependent on the development of effective therapies that have few side effects and a reasonable cost. Finally, given the complexity of living systems, there will likely be diseases for which therapies will be difficult to develop. In addition, hopefully, incentives will be sufficient for entities to develop therapies for rare, financially less-rewarding diseases.

4. At present, genetic discrimination does exist; however, recent developments in health care laws seek to minimize such discrimination by medical insurance companies. It remains to be seen whether genetic discrimination in the workplace continues.

Special Topic 5

Review Question Answers

2. Genetic engineering allows genetic material to be transferred within and between species and to alter expression levels of genes. A transgenic organism is one that involves the transfer of genetic material between different species, whereas the term cisgenic is sometimes used in cases where gene transfers occur within a species.

4. Herbicide-tolerant plants make up approximately 70 percent of all GM plants, the majority of which confer tolerance to the herbicide glyphosate. Glyphosate interferes with the enzyme 5-enolpyruvylshikimate-3-phosphate synthetase, which is present in all plants and is required for the synthesis of aromatic amino acids phenylalanine, tyrosine, and tryptophan.

6. The first iteration of Golden Rice involved the introduction of phytoene synthetase originating from the daffodil plant and carotene desaturase from a bacterium engineered into the rice plant. Resulting rice grains were yellow in color due to the production of beta-carotene. Later versions of Golden

Rice 2 involved the introduction of similar genes from maize thus leading to a much higher production of beta-carotene. At present, Golden Rice 2 is being tested in preparation for use in Bangladesh and the Philippines.

8. The biolistic method of gene introduction achieves DNA transfer by coating the transforming DNA in a heavy metal to form particles that are fired at high speed into plant cells using a gene gun. The introduced DNA may migrate into the cell nucleus and integrate into a plant chromosome.

10. This GM plant is resistant to the herbicide glyphosate, a broad-spectrum herbicide, because glyphosate interferes with the enzyme 5-enolpyruvylshikimate-3-phosphate synthetase, which is necessary for the plant to synthesize the aromatic amino acids phenylalanine, tyrosine, and tryptophan. The *epsps* gene was cloned from *Agrobacterium* strain CP4 and introduced into soybeans using biolistic bombardment.

Discussion Question Answers

2. There are many positions taken and bills filed in various states to address the question of GM food labeling. Generally, many feel a "right to know" would allow consumers to make educated choices about the food they consume. They would consider it an advantage to be able to judge the safety of a given food if they had information about the possibility that it contains GM components. Others wonder about the usefulness of a GM label if there is little information provided as to how the food has been modified. Of what value would it be to know that food was genetically modified if the science and specifics about the modifications were not included? How much background knowledge would be needed by the consumer to be able to interpret such information?

Special Topic 6

Review Question Answers

2. In *ex vivo* gene therapy, a potential genetic correction takes place in cells that have been removed from the patient. *In vivo* gene therapy treats cells of the body through the introduction of DNA into the patient.

4. In many cases, therapeutic DNA hitches a ride with genetically engineered viruses, such as retrovirus or adenovirus vectors. Nonviral delivery methods may use chemical assistance to cross cell membranes, nanoparticles, or cell fusion with artificial vesicles.

6. White blood cells, T cells in this case, were used because they are key players in the mounting of an immune response, which Ashanti was incapable of developing. A normal copy of the *ADA* gene was engineered into a retroviral vector, which then infected many of her T cells. Those cells that expressed the *ADA* gene were then injected into Ashanti's bloodstream, and some of them populated her bone marrow. At the time of Ashanti's treatment, targeted gene therapy was not possible, so integration of the *ADA* gene into Ashanti's genome probably did not replace her defective gene.

8. To some extent, targeted gene therapy is designed to alleviate one of the major pitfalls of gene therapy, random DNA integration. In addition, recent research holds promise for approaches of targeted removal and even the silencing of defective genes. DNA editing makes use of nucleases and zinc-finger arrangements to remove defective genes from the genome.

10. One method of gene inhibition follows from the use of RNA interference (RNAi) whereby double-stranded RNA molecules are delivered into cells, and the enzyme Dicer cleaves them into relatively short pieces of RNA (siRNA). siRNA can form a complex with enzymes that target mRNA. Another approach to silence genes involves the use of antisense RNA in which RNA is introduced that is complementary to a strand of mRNA, thus blocking its translation.

Discussion Question Answers

2. Generally, gene therapy is an accepted procedure, given appropriate conditions, for the relief of genetic disease states. Since it is a fairly expensive medical approach, considerable debate attends its use. It remains to be seen whether insurance companies will embrace what might be considered experimental treatments. Use of gene therapy to enhance the competitive status of individuals (genetic enhancement or gene doping) is presently viewed as cheating by most organizations and the public. It is unlikely that germ-line therapy will be viewed favorably by the public or scientific communities; however, this and other issues mentioned here will be the subject of considerable future debate.

Glossary

abortive transduction An event in which transducing DNA fails to be incorporated into the recipient chromosome.

accession number An identifying number or code assigned to a nucleotide or amino acid sequence for entry and cataloging in a database.

acentric chromosome Chromosome or chromosome fragment with no centromere.

acridine dyes A class of organic compounds that bind to DNA and intercalate into the double-stranded structure, producing local disruptions of base pairing. These disruptions result in nucleotide additions or deletions in the next round of replication.

acrocentric chromosome Chromosome with the centromere located very close to one end. Human chromosomes 13, 14, 15, 21, and 22 are acrocentric.

additive variance Genetic variance attributed to the substitution of one allele for another at a given locus. This variance can be used to predict the rate of response to phenotypic selection in quantitative traits.

allele One of the possible alternative forms of a gene, often distinguished from other alleles by phenotypic effects.

allele-specific oligonucleotide (ASO) Synthetic nucleotides, usually 15–20 bp in length, that under carefully controlled conditions will hybridize only to a perfectly matching complementary sequence.

allopatric speciation Process of speciation associated with geographic isolation.

allopolyploid Polyploid condition formed by the union of two or more distinct chromosome sets with a subsequent doubling of chromosome number.

allotetraploid An allopolyploid containing two genomes derived from different species.

allozyme An allelic form of a protein that can be distinguished from other forms by electrophoresis.

alternative splicing Generation of different protein molecules from the same pre-mRNA by incorporation of a different set and order of exons into the mRNA product.

***Alu* sequence** A DNA sequence of approximately 300 bp found interspersed within the genomes of primates that is cleaved by the restriction enzyme *Alu* I. In humans, 300,000–600,000 copies are dispersed throughout the genome and constitute some 3–6 percent of the genome. See *short interspersed elements*.

Ames test A bacterial assay developed by Bruce Ames to detect mutagenic compounds; it assesses reversion to histidine independence in the bacterium *Salmonella typhimurium*.

aminoacyl tRNA A covalently linked combination of an amino acid and a tRNA molecule. Also referred to as a charged tRNA.

amniocentesis A procedure in which fluid and fetal cells are withdrawn from the amniotic layer surrounding the fetus; used for genetic testing of the fetus.

aneuploidy A condition in which the chromosome number is not an exact multiple of the haploid set.

annotation Analysis of genomic nucleotide sequence data to identify the protein-coding genes, the nonprotein-coding genes, and the regulatory sequences and function(s) of each gene.

anticodon In a tRNA molecule, the nucleotide triplet that binds to its complementary codon triplet in an mRNA molecule.

antiparallel A term describing molecules in parallel alignment but running in opposite directions. Most commonly used to describe the opposite orientations of the two strands of a DNA molecule.

antisense oligonucleotide A short, single-stranded DNA or RNA molecule complementary to a specific sequence.

antisense RNA An RNA molecule (synthesized *in vivo* or *in vitro*) with a ribonucleotide sequence that is complementary to part of an mRNA molecule.

apoptosis A genetically controlled program of cell death, activated as part of normal development or as a result of cell damage.

Argonaute A family of proteins that are found within the RNA-induced silencing complex (RISC) and have endonuclease activity associated with the destruction of target mRNAs.

artificial selection See *selection*.

ascospore A meiotic spore produced in certain fungi.

attached-X chromosome Two conjoined X chromosomes that share a single centromere and thus migrate together during cell division.

attenuator A nucleotide sequence between the promoter and the structural gene of some bacterial operons that regulates the transit of RNA polymerase, reducing transcription of the neighboring structural gene.

autogamy A process of self-fertilization resulting in homozygosis.

autonomously replicating sequences (ARS) Origins of replication, about 100 nucleotides in length, found in yeast chromosomes.

autopolyploidy Polyploid condition resulting from the duplication of one diploid set of chromosomes.

autoradiography Production of a photographic image by radioactive decay. Used to localize radioactively labeled compounds within cells and tissues or to identify radioactive probes in various blotting techniques. See *Southern blotting*.

autosomes Chromosomes other than the sex chromosomes. In humans, there are 22 pairs of autosomes.

autotetraploid An autopolyploid condition composed of four copies of the same genome.

auxotroph A mutant microorganism or cell line that requires the addition of a nutritional substance for growth. Wild-type strains can synthesize this substance, and do not require it added for growth.

backcross A cross between an F_1 heterozygote and one of the P_1 parents (or an organism with a genotype identical to one of the parents).

bacteriophage A virus that infects bacteria, using it as the host for reproduction (also, *phage*).

balanced lethals Recessive, nonallelic lethal genes, each carried on different homologous chromosomes. When organisms carrying balanced lethal genes are interbred, only organisms with genotypes identical to the parents (heterozygotes) survive.

balanced translocation carrier An individual with a chromosomal translocation in which there has been an exchange of genetic information with no associated extra or missing genetic material.

balancer chromosome A chromosome containing one or more inversions that suppress crossing over with its homolog and which carries a dominant marker that is usually lethal when homozygous.

Barr body Densely staining DNA-positive mass seen in the somatic nuclei of mammalian females. Discovered by Murray Barr, this body represents an inactivated X chromosome.

base analog A purine or pyrimidine base that differs structurally from one normally used in biological systems but whose chemical behavior is the same.

base substitution A single base change in a DNA molecule that produces a mutation.

bidirectional replication A mechanism of DNA replication in which two replication forks move in opposite directions from a common origin.

binary switch gene A gene that acts to program a cell to follow one of a number of possible developmental pathways.

bioinformatics A field that focuses on the design and use of software and computational methods for the storage, analysis, and management of biological information such as nucleotide or amino acid sequences.

biometry The application of statistics and statistical methods to biological problems.

biotechnology Commercial and/or industrial processes that utilize biological organisms or products.

bivalents Synapsed homologous chromosomes in the first prophase of meiosis.

BLAST (Basic Local Alignment Search Tool) A software application for comparing sequence

data (DNA, RNA, protein) to search for sequence similarities.

broad heritability That proportion of total phenotypic variance in a population that can be attributed to genotypic variance.

CAAT box A highly conserved DNA sequence found in the untranslated promoter region of eukaryotic genes. This sequence is recognized by transcription factors.

cancer stem cells Tumor-forming cells in a cancer that can give rise to all the cell types in a particular form of cancer. These cells have the properties of normal stem cells: self-renewal and ability to differentiate into multiple cell types.

capillary electrophoresis A collection of analytical methods that separates large and small charged molecules in a capillary tube by their size to charge ratio.

carrier An individual heterozygous for a recessive trait.

cDNA (complementary DNA) DNA synthesized from an RNA template by the enzyme reverse transcriptase.

cell cycle The sequence of growth phases of an individual cell; divided into G1 (gap 1), S (DNA synthesis), G2 (gap 2), and M (mitosis). Cells that temporarily or permanently withdraw from the cell cycle are said to enter the G0 stage.

CEN The DNA region of centromeres critical to their function. In yeasts, fragments of chromosomal DNA, about 120 bp in length, that when inserted into plasmids confer the ability to segregate during mitosis.

centimorgan (cM) A unit of distance between genes on chromosomes representing 1 percent crossing over between two genes. Equivalent to 1 map unit (m.u.).

central dogma The concept that genetic information flow progresses from DNA to RNA to proteins. Although exceptions are known, this idea is central to an understanding of gene function.

centriole A cytoplasmic organelle composed of nine groups of microtubules, generally arranged in triplets. Centrioles function in the generation of cilia and flagella and serve as foci for the spindles in cell division.

centromere The specialized heterochromatic chromosomal region at which sister chromatids remain attached after replication, and the site to which spindle fibers attach to the chromosome during cell division. Location of the centromere determines the shape of the chromosome during the anaphase portion of cell division. Also known as the primary constriction.

centrosome Region of the cytoplasm containing a pair of centrioles.

chaperone A protein that regulates the folding of a polypeptide into a functional three-dimensional shape.

chiasma (pl., chiasmata) The crossed strands of nonsister chromatids seen in diplotene of the first meiotic division. Regarded as the cytological evidence for exchange of chromosomal material, or crossing over.

chi-square (χ^2) analysis Statistical test to determine whether or not an observed set of data is equivalent to a theoretical expectation.

chorionic villus sampling (CVS) A technique of prenatal diagnosis in which chorionic fetal cells are retrieved intravaginally or transabdominally and used to detect cytogenetic and biochemical defects in the embryo.

chromatid One of the longitudinal subunits of a replicated chromosome.

chromatin The complex of DNA, RNA, histones, and nonhistone proteins that make up uncoiled chromosomes, characteristic of the eukaryotic interphase nucleus.

chromatin immunoprecipitation (ChIP) An analytical method used to identify DNA-binding proteins that bind to DNA sequences of interest.

chromatin remodeling A process in which the structure of chromatin is chemically altered by a protein complex, resulting in changes in the transcriptional state of genes within the altered region.

chromomere A coiled, beadlike region of a chromosome, most easily visualized during cell division.

chromosomal aberration Any duplication, deletion, or rearrangement of the otherwise diploid chromosomal content of an organism. Sometimes referred to as a chromosomal mutation.

chromosome In prokaryotes, a DNA molecule containing the organism's genome; in eukaryotes, a DNA molecule complexed with proteins and RNA to form a threadlike structure containing genetic information that is visible during mitosis and meiosis.

chromosome banding Technique for the differential staining of mitotic chromosomes to produce a characteristic banding pattern.

chromosome map A diagram showing the location of genes on chromosomes.

chromosome puff A localized uncoiling and swelling in a polytene chromosome, usually regarded as a sign of active transcription.

chromosome theory of inheritance The idea put forward independently by Walter Sutton and Theodore Boveri that chromosomes are the carriers of genes and the basis for the Mendelian mechanisms of segregation and independent assortment.

***cis*-acting sequence** A DNA sequence that regulates the expression of a gene located on the same chromosome. This contrasts with a trans-acting element where regulation is under the control of a sequence on the homologous chromosome.

***cis* configuration** The arrangement of two genes (or two mutant sites within a gene) on the same homolog, such as

$$\frac{a^1 \quad a^2}{+ \quad +}$$

***cis–trans* test** A genetic test to determine whether two mutations are located within the same cistron (or gene).

cline A gradient of genotype or phenotype distributed over a geographic range.

clone Identical molecules, cells, or organisms derived from a single ancestor by asexual or parasexual methods.

CODIS (Combined DNA Index System) A standardized set of 13 short tandem repeat (STR) DNA sequences used by law enforcement and government agencies in preparing DNA profiles.

codominance Condition in which the phenotypic effects of a gene's alleles are fully and simultaneously expressed in the heterozygote.

codon A triplet of ribonucleotides that specifies a particular amino acid or a start or stop signal in the genetic code.

coefficient of coincidence A ratio of the observed number of double crossovers divided by the expected number of such crossovers.

coefficient of inbreeding The probability that two alleles present in a zygote are descended from a common ancestor.

coefficient of selection (*s*) A measurement of the reproductive disadvantage of a given genotype in a population.

cohesin A protein complex that holds sister chromatids together during mitosis and meiosis and facilitates attachments of spindle fibers to kinetochores.

colchicine An alkaloid compound that inhibits spindle formation during cell division used during the preparation of karyotypes.

colinearity The linear relationship between the nucleotide sequence in a gene (or the RNA transcribed from it) and the order of amino acids in the polypeptide chain specified by the gene.

competence In bacteria, the transient state or condition during which the cell can bind and internalize exogenous DNA molecules, making transformation possible.

complementarity Chemical affinity between nitrogenous bases of nucleic acid strands as a result of hydrogen bonding. Responsible for the base pairing between the strands of the DNA double helix and between DNA and RNA strands.

complementation test A genetic test to determine whether two mutations occur within the same gene (or cistron). If two mutations are present in a cell at the same time and produce a wild-type phenotype (i.e., they complement each other), they are often nonallelic. If a mutant phenotype is produced, the mutations are noncomplementing and are often allelic.

complete linkage A condition in which two genes are located so close to each other that no recombination occurs between them.

complex trait A trait whose phenotype is determined by the interaction of multiple genes and environmental factors.

concordance Pairs or groups of individuals with identical phenotypes. In twin studies, a condition in which both twins exhibit or fail to exhibit a trait under investigation.

conditional mutation A mutation expressed only under a certain condition; that is, a wild-type phenotype is expressed under certain (permissive) conditions and a mutant phenotype under other (restrictive) conditions.

conjugation Temporary fusion of two single-celled organisms for the sexual transfer of genetic material.

consanguineous Related by a common ancestor within the previous few generations.

consensus sequence The sequence of nucleotides in DNA or amino acids in proteins most often present in a particular gene or protein under study in a group of organisms.

contig A continuous DNA sequence reconstructed from overlapping DNA sequences derived by cloning or sequence analysis.

continuous variation Phenotype variation in which quantitative traits range from one phenotypic extreme to another in an overlapping or continuous fashion.

copy number variation (CNV) DNA segments larger than 1 kb that are repeated a variable number of times in the genome.

cosmid A vector designed to allow cloning of large segments of foreign DNA composed of the *cos* sites of phage λ inserted into a plasmid.

CpG island A short region of regulatory DNA found upstream of genes that contain unmethylated stretches of sequence with a high frequency of C and G nucleotides.

CRISPR Regions of the prokaryotic genome that contain Clusters of Regularly Interspaced Short Palindromic Repeats. Following insertion of foreign DNA into CRISPR loci, transcription from these regions produces CRISPR RNAs, which guide nucleases to invading complementary DNAs to destroy them.

CRISPR/Cas The adaptive immunity mechanism present in many prokaryotes, which utilizes CRISPR RNAs to guide Cas nucleases to invading complementary DNAs to destroy them. The CRISPR/Cas mechanism has also been exploited to introduce specific mutations in many types of eukaryotes.

crossing over The exchange of chromosomal material (parts of chromosomal arms) between homologous chromosomes by breakage and reunion. The exchange of material between nonsister chromatids during meiosis is the basis of genetic recombination.

crRNA biogenesis One step in the CRISPR/Cas mechanism in which RNAs are transcribed and processed by Cas proteins.

***C* value** The haploid amount of DNA present in a genome.

***C* value paradox** The apparent paradox that there is no relationship between the size of the genome and the evolutionary complexity of species.

cytogenetics A branch of biology in which the techniques of both cytology and genetics are used in genetic investigations.

cytokinesis The division or separation of the cytoplasm during mitosis or meiosis.

dalton (Da) A unit of mass equal to that of the hydrogen atom, which is 1.67×10^{-24} gram. A unit used in designating molecular weights.

degenerate code The representation of a given amino acid by more than one codon.

deletion A chromosomal mutation, also referred to as a deficiency, involving the loss of chromosomal material.

deme A local interbreeding population.

de novo Newly arising; synthesized from less complex precursors.

density gradient centrifugation A method of separating macromolecular mixtures by the use of centrifugal force and solutions of varying density.

deoxyribonuclease (DNase) A class of enzymes that breaks down DNA into oligonucleotide fragments by introducing single-stranded or double-stranded breaks into the double helix.

deoxyribonucleic acid (DNA) A macromolecule usually consisting of nucleotide polymers comprising antiparallel chains in which the sugar residues are deoxyribose and which are held together by hydrogen bonds. The primary carrier of genetic information.

determination Establishment of a specific pattern of gene activity and developmental fate for a given cell, usually prior to any manifestation of the cell's future phenotype.

dicentric chromosome A chromosome having two centromeres, which can be pulled in opposite directions during anaphase of cell division.

Dicer An enzyme (a ribonuclease) that cleaves double-stranded RNA (dsRNA) and pre-microRNAs (miRNAs) to form small interfering RNA (siRNA) molecules about 20 to 25 nucleotides long that serve as guide molecules for the degradation of mRNA molecules with sequences complementary to the siRNA.

dideoxynucleotide A nucleotide containing a deoxyribose sugar lacking a 3′ hydroxyl group. It stops further chain elongation when incorporated into a growing polynucleotide and is used in the Sanger method of DNA sequencing.

differentiation The complex process of change by which cells and tissues attain their adult structure and functional capacity.

dihybrid cross A genetic cross involving two characters in which the parents possess different forms of each character (e.g., yellow, round × green, wrinkled peas).

diploid (2*n*) A condition in which each chromosome exists in pairs; having two of each chromosome.

directional selection A selective force that changes the frequency of an allele in a given direction, either toward fixation (frequency of 100%) or toward elimination (frequency of 0%).

discontinuous replication of DNA The synthesis of DNA in discontinuous fragments on the lagging strand during replication. The fragments, known as Okazaki fragments, are subsequently joined by DNA ligase to form a continuous strand.

discontinuous variation Pattern of variation for a trait whose phenotypes fall into two or more distinct classes.

discordance In twin studies, a situation where one twin expresses a trait but the other does not.

disjunction The separation of chromosomes during the anaphase stage of cell division.

disruptive selection Simultaneous selection for phenotypic extremes in a population, usually resulting in the production of two phenotypically discontinuous strains.

dizygotic twins Twins produced from separate fertilization events; two ova fertilized independently. Also known as fraternal twins.

DNA fingerprinting A molecular method for identifying an individual member of a population or species using restriction enzyme digestion followed by Southern blot hybridization with minisatellite probes.

DNA footprinting A technique for identifying a DNA sequence that binds to a particular protein.

DNA microarray An ordered arrangement of DNA sequences or oligonucleotides on a substrate (often glass) used in quantitative assays of DNA–DNA or DNA–RNA binding to measure profiles of gene expression.

DNA profiling A method for identification of individuals that uses variations in the length of short tandem repeating DNA sequences (STRs) that are widely distributed in the genome.

dominant negative mutation A mutation whose gene product acts in opposition to the normal gene product, usually by binding to it to form dimers.

dosage compensation A genetic mechanism that equalizes the levels of expression of genes at loci on the X chromosome.

double crossover Two separate events of chromosome breakage and exchange occurring within the same tetrad during meiosis.

double helix The model for DNA structure proposed by James Watson and Francis Crick, in which two antiparallel hydrogen-bonded polynucleotide chains are wound into a right-handed helical configuration 2 nm in diameter, with 10 base pairs per full turn.

driver mutation A mutation in a cancer cell that contributes to tumor progression.

Drosha A nuclear enzyme with nuclease activity that is involved in the maturation of microRNAs, which removes 5′ and 3′ non-self-complementary regions of a primary miRNA to produce a pre-mRNA.

duplication A chromosomal aberration in which a segment of the chromosome is repeated.

dyad The products of tetrad separation or disjunction at meiotic prophase I. Each dyad consists of two sister chromatids joined at the centromere.

electrophoresis A technique that separates a mixture of molecules by their differential migration through a stationary medium (such as a gel) under the influence of an electrical field.

electroporation A technique that uses an electric pulse to move polar molecules across the plasma membrane into the cell.

ELSI (Ethical, Legal, Social Implications) A program established by the National Human Genome Research Institute in 1990 as part of the Human Genome Project to sponsor research on the ethical, legal, and social implications of genomic research and its impact on individuals and social institutions.

embryonic stem cells (ESC) Cells derived from the inner cell mass of early blastocyst mammalian embryos. These cells are pluripotent, meaning they can differentiate into any of the embryonic or adult cell types characteristic of the organism.

ENCODE (Encyclopedia of DNA Elements) An international effort to identify and analyze all functional DNA elements of the human genome that are involved in the regulation of gene expression.

endogenous siRNAs (endo-siRNAs) Short interfering RNAs that are derived from endogenous sources such as bi-directional transcription of repetitive sequences (centromeres and transposons) that are processed by Dicer.

enhancer A DNA sequence that enhances transcription and the expression of structural genes, often acting over a distance of thousands of base pairs located upstream, downstream, or internal to the gene they affect.

epigenesis The idea that an organism or organ arises through the sequential appearance and development of new structures, in contrast to preformationism, which holds that development is the result of the assembly of structures already present in the egg.

epigenetics The study of modifications in an organism's pattern of gene expression or phenotypic expression that are not attributable to alterations in the nucleotide sequence (mutations) of the organism's DNA.

epimutation The abnormal repression or activation of a gene caused by errors in epigenetic mechanisms of gene regulation.

epistasis Nonreciprocal interaction between nonallelic genes such that one gene influences or interferes with the expression of another gene, leading to a specific phenotype.

equational division A division stage where the number of centromeres is not reduced by half.

euchromatin Chromatin or chromosomal regions that are lightly staining and are relatively uncoiled during the interphase portion of the cell cycle. Euchromatic regions contain most of the structural genes.

eugenics A movement advocating the improvement of the human species by selective breeding. Positive eugenics refers to the promotion of breeding between people thought to possess favorable genes, and negative eugenics refers to the discouragement of breeding among those thought to have undesirable traits.

euphenics Medical or genetic intervention to reduce the impact of defective genotypes.

euploid Polyploid with a chromosome number that is an exact multiple of a basic chromosome set.

evolution Descent with modification. The emergence of new kinds of plants and animals from preexisting types.

excision repair Removal of damaged DNA segments followed by repair synthesis with the correct nucleotide sequence.

exon The DNA segments of a gene that contain the sequences that, through transcription and translation, are eventually represented in the amino acid sequence of the final polypeptide product.

expressed sequence tag (EST) All or part of the nucleotide sequence of a cDNA clone. ESTs are used as markers in the construction of genetic maps.

expression vector Plasmids or phages carrying promoter regions designed to cause expression of inserted DNA sequences.

expressivity The degree to which a phenotype for a given trait is expressed.

extracellular RNAs (exRNAs) Various types of RNAs (such as mRNAs and microRNAs) that are secreted in association with proteins or in vesicles for protection and that serve to signal other cells.

extranuclear inheritance Transmission of traits by genetic information contained in cytoplasmic organelles such as mitochondria and chloroplasts. Sometimes called *extrachromosomal inheritance*.

F^- cell A bacterial cell that does not contain a fertility factor and that acts as a recipient in bacterial conjugation.

F^+ cell A bacterial cell that contains a fertility factor and that acts as a donor in bacterial conjugation.

F factor An episomal plasmid in bacterial cells that confers the ability to act as a donor in conjugation.

F′ factor A fertility factor that contains a portion of the bacterial chromosome.

F_1 generation First filial generation; the progeny resulting from the first cross in a series.

F_2 generation Second filial generation; the progeny resulting from a cross of the F_1 generation.

F pilus On bacterial cells possessing an F factor, a filament-like projection that plays a role in conjugation.

familial trait A trait transmitted through and expressed by members of a family. Often used to describe a trait whose precise mode of inheritance is not clear.

fate map A diagram of an embryo showing the location of cells whose developmental fate is known.

fetal cell sorting A noninvasive method of prenatal diagnosis that recovers and tests fetal cells from the maternal circulation.

filial generations See *F_1, F_2 generations*.

fitness A measure of the relative survival and reproductive success of a given individual or genotype.

fluctuation test A statistical test demonstrating that bacterial mutations arise spontaneously, in contrast to being induced by selective agents.

fluorescence *in situ* hybridization (FISH) A method of *in situ* hybridization that utilizes probes labeled with a fluorescent tag, causing the site of hybridization to fluoresce when viewed using ultraviolet light.

flush–crash cycle A period of rapid population growth followed by a drastic reduction in population size.

folded-fiber model A model of eukaryotic chromosome organization in which each sister chromatid consists of a single chromatin fiber wound like a tightly coiled skein of yarn.

forensic science The use of laboratory scientific methods to obtain data used in criminal and civil law cases.

forward genetics The classical approach used to identify a gene controlling a phenotypic trait in the absence of knowledge of the gene's location in the genome or its DNA sequence. An approach contrasted with *reverse genetics*.

founder effect The establishment of a population by a small number of individuals whose genotypes carry only a fraction of the alleles in the parental population.

fragile site A heritable gap, or nonstaining region, of a chromosome that can be induced to generate chromosome breaks.

frameshift mutation A mutational event leading to the insertion of one or more base pairs in a gene, shifting the codon reading frame in all codons that follow the mutational site.

functional genomics The study of gene function based on the resulting RNAs or proteins they encode.

G1 checkpoint A point in the G1 phase of the cell cycle when a cell either becomes committed to initiating DNA synthesis and continuing the cycle or withdraws into the G0 resting stage.

G0 A nondividing but metabolically active state that cells may enter from the G1 phase of the cell cycle.

gain-of-function mutation A mutation that produces a phenotype different from that of the normal allele and from any loss-of-function alleles.

gamete A specialized reproductive cell with a haploid number of chromosomes.

gap genes Genes expressed in contiguous domains along the anterior–posterior axis of the *Drosophila* embryo that regulate the process of segmentation in each domain.

GC box In eukaryotes, a region in a promoter containing a 5′-GGGCGG-3′ sequence, which is a binding site for transcriptional regulatory proteins.

GenBank An international, open-source database of publicly available DNA sequences.

gene The fundamental physical unit of heredity, whose existence can be confirmed by allelic variants and which occupies a specific chromosomal locus. A DNA sequence coding for a single polypeptide.

gene amplification The process by which gene sequences are selected and differentially replicated either extrachromosomally or intrachromosomally.

gene conversion The process of nonreciprocal recombination by which one allele in a heterozygote is converted into the corresponding allele.

gene duplication An event leading to the production of a tandem repeat of a gene sequence during replication.

gene family A number of closely related genes derived from a common ancestral gene by duplication and sequence divergence over evolutionary time.

gene flow The gradual exchange of genes between two populations; brought about by the dispersal of gametes or the migration of individuals.

gene interaction Production of novel phenotypes by the interaction of alleles of different genes.

gene knockout A gene in an organism that is inactivated for the purpose of studying gene function. Sometimes called gene targeting.

gene pool The total of all alleles possessed by the reproductive members of a population.

gene-regulatory networks Genes and DNA sequences that interact with each other and with cell signaling systems to coordinate the expression of gene sets that control the formation of body structures.

gene targeting A transgenic technique used to create and introduce a specifically altered gene into an organism. Gene targeting often involves the induction of a specific mutation in a cloned gene that is then introduced into the genome of a gamete involved in fertilization. The organism produced is bred to produce adults homozygous for the mutation, for example, the creation of a *gene knockout*.

gene therapy The delivery of therapeutic sequences (DNA or RNA) to treat or correct genetic disease conditions.

genetically modified organism (GMO) A plant or animal whose genome carries a gene transferred from another species by recombinant DNA technology that is expressed to produce a gene product.

genetic anticipation The phenomenon in which the severity of symptoms in genetic disorders increases from generation to generation and the age of onset decreases from generation to generation. It is caused by the expansion of trinucleotide repeats within or near a gene and was first observed in myotonic dystrophy.

genetic background The impact of the collective genome of an organism on the expression of a gene under investigation.

genetic code The deoxynucleotide triplets that encode the 20 amino acids or specify termination of translation.

genetic drift Random variation in allele frequency from generation to generation, most often observed in small populations.

genetic engineering The technique of altering the genetic constitution of cells or individuals by the selective removal, insertion, or modification of individual genes or gene sets.

genetic equilibrium A condition in which allele frequencies in a population are neither increasing nor decreasing.

genetic erosion The loss of genetic diversity from a population or a species.

genetic fine structure analysis Intragenic recombinational analysis that provides intragenic mapping information at the level of individual nucleotides.

genetic load Average number of recessive lethal genes carried in the heterozygous condition by an individual in a population.

genetic polymorphism The stable coexistence of two or more distinct genotypes for a given trait in a population. When the frequencies of two alleles for such a trait are in equilibrium, the condition is called a balanced polymorphism.

genetics The branch of biology concerned with study of inherited variation. More specifically, the study of the origin, transmission, expression, and evolution of genetic information.

genome The set of hereditary information encoded in the DNA of an organism, including both the protein-coding and non–protein-coding sequences.

genome-wide association studies (GWAS) Analysis of genetic variation across an entire genome, searching for linkage (associations) between variations in DNA sequences and a genome region encoding a specific phenotype.

genomic imprinting The process by which the expression of an allele depends on whether it has been inherited from a male or a female parent. Also referred to as parental imprinting.

genomic library A collection of clones that contains all the DNA sequences of an organism's genome.

genomics A subdiscipline of the field of genetics generated by the union of classical and molecular biology with the goal of sequencing and understanding genes, gene interaction, genetic elements, as well as the structure and evolution of genomes.

genotype The allelic or genetic constitution of an organism; often, the allelic composition of one or a limited number of genes under investigation.

germ line An embryonic cell lineage that forms the reproductive cells (eggs and sperm).

germ plasm Hereditary material transmitted from generation to generation.

Goldberg–Hogness box A short nucleotide sequence 20–30 bp upstream from the initiation site of eukaryotic genes to which RNA polymerase II binds. The consensus sequence is TATAAAA. Also known as a TATA box.

gynandromorphy An individual composed of cells with both male and female genotypes.

haploid (*n*) A cell or an organism having one member of each pair of homologous chromosomes. Also referred to as the gametic chromosome number.

haploinsufficiency In a diploid organism, a condition in which an individual possesses only one functional copy of a gene with the other inactivated by mutation. The amount of protein produced by the single copy is insufficient to produce a normal phenotype, thus leading to an abnormal phenotype. In humans, this condition is present in many autosomal dominant disorders.

haplotype A set of alleles from closely linked loci carried by an individual inherited as a unit.

HapMap Project An international effort by geneticists to identify haplotypes (closely linked genetic markers on a single chromosome) shared by certain individuals as a way of facilitating efforts to identify, map, and isolate genes associated with disease or disease susceptibility.

Hardy–Weinberg law The principle that genotype frequencies will remain in equilibrium in an infinitely large, randomly mating population in the absence of mutation, migration, and selection.

helix–turn–helix (HTH) motif In DNA-binding proteins, the structure of a region in which a turn of four amino acids holds two a helices at right angles to each other.

hemizygous Having a gene present in a single dose in an otherwise diploid cell. Usually applied to genes on the X chromosome in heterogametic males.

heritability A relative measure of the degree to which observed phenotypic differences for a trait are genetic.

heterochromatin The heavily staining, late-replicating regions of chromosomes that are prematurely condensed in interphase.

heteroduplex A double-stranded nucleic acid molecule in which each polynucleotide chain has a different origin. It may be produced as an intermediate in a recombinational event or by the *in vitro* reannealing of single-stranded, complementary molecules.

heterogametic sex The sex that produces gametes containing unlike sex chromosomes. In mammals, the male is the heterogametic sex.

heterokaryon A somatic cell containing nuclei from two different sources.

heterozygote An individual with different alleles at one or more loci. Such individuals will produce unlike gametes and therefore will not breed true.

Hfr Strains of bacteria exhibiting a high frequency of recombination. These strains have a chromosomally integrated F factor that is able to mobilize and transfer part of the chromosome to a recipient F^- cell.

high-throughput DNA sequencing A collection of DNA sequencing methods that outperform the standard (Sanger) method of DNA sequencing by a factor of 100–1000 and reduce sequencing costs by more than 99 percent. Also called *next generation sequencing*.

histone methyltransferase An enzyme that catalyzes the addition of methyl groups to the histone proteins and thus modifies chromatin condensation.

Holliday structure In DNA recombination, an intermediate seen in transmission electron microscope images as an X-shaped structure showing four single-stranded DNA regions.

homeobox A sequence of about 180 nucleotides that encodes a sequence of 60 amino acids called a *homeodomain*, which is part of a DNA-binding protein that acts as a transcription factor.

homeotic mutation A mutation that causes a tissue normally determined to form a specific organ or body part to alter its pathway of differentiation and form another structure.

homogametic sex The sex that produces gametes that do not differ with respect to sex-chromosome content; in mammals, the female is homogametic.

homologous chromosomes Chromosomes that synapse or pair during meiosis and that are identical with respect to their genetic loci and centromere placement.

homozygote An individual with identical alleles for a gene or genes of interest. These individuals will produce identical gametes (with respect to the gene or genes in question) and will therefore breed true.

horizontal gene transfer The nonreproductive transfer of genetic information from an organism to another, across species and higher taxa (even domains). This mode is contrasted with vertical gene transfer, which is the transfer of genetic information from parent to offspring. In some species of bacteria and archaea, up to 5 percent of the genome may have originally been acquired through horizontal gene transfer.

hot spots Genome regions where mutations are observed with a high frequency. These include a predisposition toward single-nucleotide substitutions or unequal crossing over.

Human Genome Project (HGP) An international effort to determine the sequence of the human genome, to identify all genes in the genome, and to map all genes to specific chromosomes, among other goals.

human immunodeficiency virus (HIV) An RNA-containing retrovirus associated with the onset and progression of acquired immunodeficiency syndrome (AIDS).

Human Microbiome Project An effort to sequence complete genomes for an estimated 600 to 1000 microorganisms (bacteria, viruses and yeast) that live on and inside humans.

hybrid An individual produced by crossing parents from two different genetic strains.

hybrid vigor The general superiority of a hybrid over a purebred.

imprinting See genomic imprinting

inborn error of metabolism A genetically controlled biochemical disorder; usually an enzyme defect that produces a clinical syndrome.

inbreeding depression A decrease in viability, vigor, or growth in progeny after several generations of inbreeding.

incomplete dominance Expressing a heterozygous phenotype that is distinct from the phenotype of either homozygous parent. Also called *partial dominance*.

independent assortment The independent behavior of each pair of homologous chromosomes during their segregation in meiosis I. The random distribution of maternal and paternal homologs into gametes.

inducible enzyme system An enzyme system under the control of an inducer, a regulatory molecule that acts to block a repressor and allow transcription.

initiation codon The nucleotide triplet AUG that in an mRNA molecule codes for incorporation of the amino acid methionine as the first amino acid in a polypeptide chain.

interference (*I*) A measure of the degree to which one crossover affects the incidence of another crossover in an adjacent region of the same chromatid. Negative interference increases the chance of another crossover; positive interference reduces the probability of a second crossover event.

interphase In the cell cycle, the interval between divisions.

intron Any segment of DNA that lies between coding regions in a gene. Introns are transcribed but are spliced out of the RNA product and are not represented in the polypeptide encoded by the gene. Also known as an intervening sequence.

inversion A chromosomal aberration in which a chromosomal segment has been reversed.

in vitro Literally, *in glass*; outside the living organism; occurring in an artificial environment.

***in vitro* evolution** A stepwise process of small changes to a nucleic acid that mimics natural selection, but that occurs outside of living cells.

in vivo Literally, *in the living*; occurring within the living body of an organism.

isotopes Alternate forms of atoms with identical chemical properties that have the same atomic number but differ in the number of neutrons (and thus their mass) contained in the nucleus.

isozyme Any of two or more distinct forms of an enzyme with identical or nearly identical chemical properties but differ in some property such as net electrical charge, pH optima, number and type of subunits, or substrate concentration.

karyokinesis The process of nuclear division.

karyotype The chromosome complement of a cell or an individual. Often used to refer to the arrangement of metaphase chromosomes in a sequence according to length and centromere position.

kinetochore A fibrous structure with a size of about 400 nm, located within the centromere. It is the site of microtubule attachment during cell division.

Kozak sequence A short nucleotide sequence adjacent to the initiation codon that is recognized as the translational start site in eukaryotic mRNA.

lagging strand During DNA replication, the strand synthesized in a discontinuous fashion, in the direction opposite of the replication fork.

lampbrush chromosomes Meiotic chromosomes characterized by extended lateral loops. Although most intensively studied in amphibians, these structures occur in meiotic cells of organisms ranging from insects to humans.

lariat structure A structure formed during pre-mRNA processing by formation of a 5′ to 3′ bond in an introns, leading to removal of that intron from an mRNA molecule.

leader sequence That portion of an mRNA molecule from the 5′ end to the initiating codon, often containing regulatory or ribosome binding sites.

leading strand During DNA replication, the strand synthesized continuously in the direction of the replication fork.

lethal gene A gene whose expression results in premature death of the organism at some stage of its life cycle.

leucine zipper In DNA-binding proteins, a structural motif characterized by a stretch in which every seventh amino acid residue is leucine, with adjacent regions containing positively charged amino acids. Leucine zippers on two polypeptides may interact to form a dimer that binds to DNA.

locus (pl., loci) The site or place on a chromosome where a particular gene is located.

long interspersed elements (LINEs) Long, repetitive sequences found interspersed in the genomes of higher organisms.

long noncoding RNAs (lncRNAs) RNAs that are longer than 200 nucleotides and do not encode for polypeptides. lncRNAs have various functions including epigenetic modifications of DNA and regulation of the activity of transcription factors.

long terminal repeat (LTR) A sequence of several hundred base pairs found at both ends of a retroviral DNA.

loss-of-function mutation Mutations that produce alleles that encode proteins with reduced or no function.

Lyon hypothesis The proposal describing the random inactivation of the maternal or paternal X chromosome in somatic cells of mammalian females early in development.

lysis The disintegration of a cell brought about by the rupture of its membrane.

lysogenic bacterium A bacterial cell carrying the DNA of a temperate bacteriophage integrated into its chromosome.

lysogeny The process by which the DNA of an infecting phage becomes repressed and integrated into the chromosome of the bacterial cell it infects.

map unit A measure of the genetic distance between two genes, corresponding to a recombination frequency of 1 percent. See *centimorgan (cM)*.

maternal effect Phenotypic effects in offspring attributable to genetic information transmitted through the oocyte derived from the maternal genome.

maternal inheritance The transmission of traits strictly through the maternal parent, usually due to DNA found in the cytoplasmic organelles, the mitochondria, or chloroplasts.

meiosis The process of cell division in gametogenesis or sporogenesis during which the diploid number of chromosomes is reduced to the haploid number.

melting profile (T_m) The temperature at which a population of double-stranded nucleic acid molecules is half-dissociated into single strands.

merozygote A partially diploid bacterial cell containing, in addition to its own chromosome, a chromosome fragment introduced into the cell by transformation, transduction, or conjugation.

messenger RNA (mRNA) An RNA molecule transcribed from DNA and translated into the amino acid sequence of a polypeptide.

metacentric chromosome A chromosome that has a centrally located centromere and therefore chromosome arms of equal lengths.

metafemale In *Drosophila*, a poorly developed female of low viability with a ratio of X chromosomes to sets of autosomes that exceeds 1.0.

metagenomics The study of DNA recovered from organisms collected from the environment as opposed to those grown as laboratory cultures. Often used for estimating the diversity of organisms in an environmental sample.

metamale In *Drosophila*, a poorly developed male of low viability with a ratio of X chromosomes to sets of autosomes that is below 0.5.

metastasis The process by which cancer cells spread from the primary tumor and establish malignant tumors in other parts of the body.

methylation Enzymatic transfer of methyl groups from S-adenosylmethionine to biological molecules, including phospholipids, proteins, RNA, and DNA. Methylation of DNA is associated with the regulation of gene expression and with epigenetic phenomena such as imprinting.

microRNA Single-stranded RNA molecules approximately 20–23 nucleotides in length that regulate gene expression by participating in the degradation of mRNA.

microsatellite A short, highly polymorphic DNA sequence of 1–4 base pairs, widely distributed in the genome, that is used as a molecular marker in a variety of methods. Also called a *simple sequence repeat* (*SSR*).

minimal medium A medium containing only the essential nutrients needed to support the growth and reproduction of wild-type strains of an organism. Usually comprised of inorganic components that include a carbon and nitrogen source.

minisatellite Series of short tandem repeat sequences (STRs) 10–100 nucleotides in length that occur frequently throughout the genome of eukaryotes. Because the number of repeats at each locus is variable, the loci are known as variable number tandem repeats (VNTRs). Used in DNA fingerprinting and DNA profiles.

mismatch repair A form of excision repair of DNA in which the repair mechanism is able to distinguish between the strand with the error and the strand that is correct.

missense mutation A mutation that changes a codon to that of another amino acid and thus results in an amino acid substitution in the translated protein.

mitosis A form of cell division producing two progeny cells identical genetically to the parental cell—that is, the production of two cells from one, each having the same chromosome complement as the parent cell.

model genetic organism An experimental organism conducive to efficiently conducted research whose genetics is intensively studied on the premise that the findings can be applied to other organisms.

molecular clock In evolutionary studies, a method that counts the number of differences in DNA or protein sequences as a way of measuring the time elapsed since two species diverged from a common ancestor.

monohybrid cross A genetic cross involving only one character (e.g., $AA \times aa$).

monophyletic group A taxon (group of organisms) consisting of an ancestor and all its descendants.

monosomic An aneuploid condition in which one member of a chromosome pair is missing; having a chromosome number of $2n-1$.

monozygotic twins Twins produced from a single fertilization event; the first division of the zygote produces two cells, each of which develops into an embryo. Also known as *identical twins*.

multigene family A set of genes descended from a common ancestral gene usually by duplication and subsequent sequence divergence. The globin genes are an example of a multigene family.

multiple alleles The presence of three or more alleles of the same gene in a population of organisms.

mutagen Any agent that causes an increase in the spontaneous rate of mutation.

mutation The process that produces an alteration in DNA or chromosome structure; in genes, the source of new alleles.

mutation rate The frequency with which mutations take place at a given locus or in a population.

natural selection Differential reproduction among members of a species owing to variable fitness conferred by genotypic differences.

network map Computer-generated representation of interacting genes, proteins, and other molecules based on experimental data or proposed interactions.

neutral mutation A mutation with no immediate adaptive significance or phenotypic effect.

noncoding RNAs (ncRNAs) RNAs that do not encode for polypeptides.

noncrossover gamete A gamete whose chromosomes have undergone no genetic recombination.

nondisjunction A cell division error in which homologous chromosomes or the sister chromatids fail to separate and migrate to opposite poles; responsible for defects such as monosomy and trisomy.

noninvasive prenatal genetic diagnosis (NIPGD) A noninvasive method of fetal genotyping that uses a maternal blood sample to analyze thousands of fetal loci using fetal DNA fragments present in the maternal blood.

nonsense codons The nucleotide triplets (UGA, UAG, and UAA) in an mRNA molecule that signal the termination of translation.

nonsense mutation A mutation that changes a codon specifying an amino acid into a termination codon, leading to premature termination during translation of mRNA.

Northern blotting An analytic technique in which RNA molecules are separated by electrophoresis and transferred by capillary action to a nylon or nitrocellulose membrane. Specific RNA molecules can then be identified by hybridization to a labeled nucleic acid probe.

nuclease An enzyme that breaks bonds in nucleic acid molecules.

nucleoid The DNA-containing region within the cytoplasm in prokaryotic cells.

nucleolar organizer region (NOR) A chromosomal region containing the genes for rRNA; most often found in physical association with the *nucleolus*.

nucleolus The nuclear site of ribosome biosynthesis and assembly; usually associated with or formed in association with the DNA comprising the *nucleolar organizer region*.

nucleoside In nucleic acid chemical nomenclature, a purine or pyrimidine base covalently linked to a ribose or deoxyribose sugar molecule.

nucleosome In eukaryotes, a nuclear complex consisting of four pairs of histone molecules wrapped by two turns of a DNA molecule. The major structure associated with the organization of chromatin in the nucleus.

nucleotide In nucleic acid chemical nomenclature, a nucleoside covalently linked to one or more phosphate groups. Nucleotides containing a single phosphate linked to the 5′ carbon of the ribose or deoxyribose are the building blocks of nucleic acids.

nucleus The membrane-bound cytoplasmic organelle of eukaryotic cells that contains the chromosomes and nucleolus.

null allele A mutant allele that produces no functional gene product. Usually inherited as a recessive trait.

null hypothesis (H0) Used in statistical tests, the hypothesis that there is no real difference between the observed and expected datasets. Statistical methods such as chi-square analysis are used to test the probability associated with this hypothesis.

Okazaki fragment The short, discontinuous strands of DNA produced on the lagging strand during DNA synthesis.

oligonucleotide A linear sequence of about 10–20 nucleotides connected by 5′-3′ phosphodiester bonds.

oncogene A gene whose activity promotes uncontrolled proliferation in eukaryotic cells. Usually a mutant gene derived from a *proto-oncogene*.

Online Mendelian Inheritance in Man (OMIM) A database listing all known genetic disorders and disorders with genetic components. It also contains a listing of all known human genes and links genes to genetic disorders.

open reading frame (ORF) A nucleotide sequence organized as triplets that encodes the amino acid sequence of a polypeptide, including an initiation codon and a termination codon.

operator region In bacterial DNA, a region that interacts with a specific repressor protein to regulate the expression of an adjacent gene or gene set.

operon A genetic unit consisting of one or more structural genes encoding polypeptides, and an adjacent operator gene that regulates the transcriptional activity of the structural gene or genes.

outbreeding depression Reduction in fitness in the offspring produced by mating genetically diverse parents. It is thought to result from a lowered adaptation to local environmental conditions.

overlapping code A hypothetical genetic code in which any given triplet is shared by more than one adjacent codon.

pair-rule genes Genes expressed as stripes around the blastoderm embryo during development of the *Drosophila* embryo.

paleogenomics The recovery, sequencing, and analysis of genes and genomes from fossils of extinct species.

palindrome In genetics, a double-stranded DNA segment where each strand's base sequence is identical when read 5′ to 3′. For example:

5′-GAATTC-3′
3′-CTTAAG-5′

Palindromic sequences are noteworthy as recognition and cleavage sites for restriction endonucleases.

paracentric inversion A chromosomal inversion that does not include the region containing the centromere.

pedigree In human genetics, a diagram showing the ancestral relationships and transmission of genetic traits over several generations in a family.

P element In *Drosophila*, a transposable DNA element responsible for hybrid dysgenesis.

penetrance The frequency, expressed as a percentage, with which individuals of a given genotype manifest at least some degree of a specific mutant phenotype associated with a trait.

pericentric inversion A chromosomal inversion that involves both arms of the chromosome and thus the centromere.

pharmacogenomics The study of how genetic variation influences the action of pharmaceutical drugs in individuals.

phenotype The overt appearance of a genetically controlled trait.

Philadelphia chromosome The product of a reciprocal translocation in humans that contains the short arm of chromosome 9, carrying the *C-ABL* oncogene, and the long arm of chromosome 22, carrying the *BCR* gene.

phosphodiester bond In nucleic acids, the system of covalent bonds by which a phosphate group links adjacent nucleotides, extending from the 5′ carbon of one pentose sugar (ribose or deoxyribose) to the 3′ carbon of the pentose sugar in the neighboring nucleotide. Phosphodiester bonds create the backbone of nucleic acid molecules.

photoreactivation repair Light-induced repair of damage caused by exposure to ultraviolet light. Associated with an intracellular enzyme system.

phyletic evolution The gradual transformation of one species into another over time; so-called vertical evolution.

pilus A filamentlike projection from the surface of a bacterial cell. Often associated with cells possessing F factors.

piRNAs PIWI-interacting RNAs. Short dsRNA sequences associated with PIWI proteins. Participate in transcriptional and/or post transcriptional mechanisms to silence transposons and repetitive sequences in germ cells.

plaque On an otherwise opaque bacterial lawn, a clear area caused by the growth and reproduction of a single bacteriophage.

plasmid An extrachromosomal, circular DNA molecule that replicates independently of the host chromosome.

pleiotropy Condition in which a single mutation causes multiple phenotypic effects.

ploidy A term referring to the basic chromosome set or to multiples of that set.

point mutation A mutation that can be mapped to a single locus. At the molecular level, a mutation that results in the substitution of one nucleotide for another. Also called a *gene mutation*.

polar body Produced in females at either the first or second meiotic division of gametogenesis, a discarded cell that contains one of the nuclei of the division process, but almost no cytoplasm as a result of an unequal cytokinesis.

polycistronic mRNA A messenger RNA molecule that encodes the amino acid sequence of two or more polypeptide chains in adjacent structural genes.

polygenic inheritance The transmission of a phenotypic trait whose expression depends on the additive effect of a number of genes.

polylinker A segment of DNA that has been engineered to contain multiple sites for restriction enzyme digestion. Polylinkers are usually found in engineered vectors such as plasmids.

polymerase chain reaction (PCR) A method for amplifying DNA segments that depends on repeated cycles of denaturation, primer annealing, and DNA polymerase–directed DNA synthesis.

polymerases Enzymes that catalyze the formation of DNA and RNA from deoxynucleotides and ribonucleotides, respectively.

polymorphism The existence of two or more discontinuous, segregating phenotypes in a population.

polynucleotide A linear sequence of 20 or more nucleotides, joined by 5′-3′ phosphodiester bonds.

polypeptide A molecule composed of amino acids linked together by covalent peptide bonds. This term is used to denote the amino acid chain before it folds into its functional three-dimensional protein configuration.

polyploid A cell or individual having more than two haploid sets of chromosomes.

polysome A structure composed of two or more ribosomes associated with an mRNA and associated tRNAs engaged in translation. Also called a *polyribosome*.

polytene chromosome Literally, a many-stranded chromosome; one that has undergone numerous rounds of DNA replication without separation of the replicated strands, which remain in exact parallel register. The result is a giant chromosome with aligned chromomeres displaying a characteristic banding pattern, most often studied in *Drosophila* larval salivary gland cells.

population A local group of actually or potentially interbreeding individuals belonging to the same species.

population bottleneck A drastic reduction in population size and consequent loss of genetic diversity, followed by an increase in population size. The rebuilt population has a gene pool with reduced diversity caused by genetic drift.

positional cloning The identification and subsequent cloning of a gene in the absence of knowledge of its polypeptide product or function. The process uses cosegregation of mutant phenotypes with DNA markers to identify the chromosome containing the gene; the position of the gene is identified establishing linkage with additional markers.

position effect Change in expression of a gene associated with a change in the gene's location within the genome.

posttranslational modification The processing or modification of the translated polypeptide chain by enzymatic cleavage,

or the addition of phosphate groups, carbohydrate chains, or lipids.

posttranscriptional modification Changes made to pre-mRNA molecules during conversion to mature mRNA. These include the addition of a methylated cap at the 5′ end and a poly-A tail at the 3′ end, excision of introns, and exon splicing.

postzygotic isolation mechanism A barrier that prevents or reduces inbreeding by acting after fertilization to produce nonviable, sterile hybrids or hybrids of lowered fitness.

preadaptive mutation A mutational event that later becomes of adaptive significance.

preimplantation genetic diagnosis (PGD) The removal and genetic analysis of unfertilized oocytes, polar bodies, or single cells from an early embryo (3–5 days old).

prezygotic isolation mechanism A barrier that reduces inbreeding by preventing courtship, mating, or fertilization.

Pribnow box In prokaryotic genes, a 6-bp sequence to which the sigma (σ) subunit of RNA polymerase binds, upstream from the beginning of transcription. The consensus sequence for this box is TATAAT.

primary miRNAs (pri-miRNAs) The product of transcription of a microRNA gene, which has a 5′ methylated cap, a 3′ polyadenylated tail, and a hairpin structure due to self-complementary sequences.

primary sex ratio Ratio of males to females at fertilization, often expressed in decimal form (e.g., 1.06).

primer In nucleic acids, a short length of RNA or single-stranded DNA required for initiating synthesis directed by polymerases.

prion An infectious pathogenic agent devoid of nucleic acid and composed of a protein, PrP, with a molecular weight of 27,000–30,000 Da. Prions are known to cause scrapie, a degenerative neurological disease in sheep; bovine spongiform encephalopathy (BSE, or mad cow disease) in cattle; and similar diseases in humans, including kuru and Creutzfeldt–Jakob disease.

proband An individual who is the focus of a genetic study leading to the construction of a pedigree tracking the inheritance of a genetically determined trait of interest. Formerly known as a *propositus*.

probe A macromolecule such as DNA or RNA that has been labeled and can be detected by an assay such as autoradiography or fluorescence microscopy. Probes are used to identify target molecules, genes, or gene products.

product law In statistics, the probability that two independent events occurring simultaneously is equal to the product of their individual probabilities.

promoter element An upstream regulatory region of a gene to which RNA polymerase binds prior to the initiation of transcription.

proofreading A molecular mechanism for scanning and correcting errors in replication, transcription, or translation.

prophage A bacteriophage genome integrated into a bacterial chromosome that is replicated along with the bacterial chromosome. Bacterial cells carrying prophages are said to be *lysogenic* and to be capable of entering the *lytic cycle*, whereby the phage is replicated.

propositus (female, proposita) See *proband*.

protein domain Amino acid sequences with specific conformations and functions that are structurally and functionally distinct from other regions on the same protein.

proteome The entire set of proteins expressed by a cell, tissue, or organism at a given time. The study of the proteome is referred to as proteomics.

proto-oncogene A gene that functions to initiate, facilitate, or maintain cell growth and division. Proto-oncogenes can be converted to *oncogenes* by mutation.

protoplast A bacterial or plant cell with the cell wall removed. Sometimes called a *spheroplast*.

prototroph A strain (usually of a microorganism) that is capable of growth on a defined, minimal medium. Wild-type strains are usually regarded as prototrophs and contrasted with *auxotrophs*.

pseudoalleles Genes that behave as alleles to one another by complementation but can be separated from one another by recombination.

pseudoautosomal region A region on the human Y chromosome that is also represented on the X chromosome. Genes found in this region of the Y chromosome have a pattern of inheritance that is indistinguishable from genes on autosomes.

pseudodominance The expression of a recessive allele on one homolog owing to the deletion of the dominant allele on the other homolog.

pseudogene A nonfunctional gene with sequence homology to a known structural gene present elsewhere in the genome. It differs from the functional version by insertions or deletions and by the presence of flanking direct-repeat sequences of 10–20 nucleotides.

punctuated equilibrium A pattern in the fossil record of long periods of species stability, punctuated with brief periods of species divergence.

pyrosequencing A high-throughput method of DNA sequencing that determines the sequence of a single-stranded DNA molecule by synthesis of a complementary strand. During synthesis, the sequence is determined by the chemiluminescent detection of pyrophosphate release that accompanies nucleotide incorporation into a newly synthesized strand of DNA.

quantitative real-time PCR (qPCR) A variation of PCR (polymerase chain reaction) that uses fluorescent probes to quantitate the amount of DNA or RNA product present after each round of amplification.

quantitative trait loci (QTLs) Two or more genes that act on a single polygenic trait in a quantitative way.

quantum speciation Formation of a new species within a single or a few generations by a combination of selection and drift.

quorum sensing A mechanism used to regulate gene expression in bacteria, in response to changes in cellular population density.

rad A unit of absorbed dose of radiation with an energy equal to 100 ergs per gram of irradiated tissue.

radioactive isotope An unstable isotope with an altered number of neutrons that emits ionizing radiation during decay as it is transformed to a stable atomic configuration.

random amplified polymorphic DNA (RAPD) A PCR method that uses random primers about 10 nucleotides in length to amplify unknown DNA sequences.

reading frame A linear sequence of codons in a nucleic acid.

reannealing Formation of double-stranded DNA molecules from denatured single strands.

recessive An allele whose potential genetic expression is overridden in the heterozygous condition by a dominant allele.

reciprocal translocation A chromosomal aberration in which nonhomologous chromosomes exchange parts.

recombinant DNA technology A collection of methods used to create DNA molecules by *in vitro* ligation of DNA from two different organisms, and the replication and recovery of such recombinant DNA molecules.

recombination The process that leads to the formation of new allele combinations on chromosomes.

reductional division The chromosome division that halves the number of centromeres and thus reduces the chromosome number by half in the daughter cells. The first division of meiosis is a reductional division

rem Radiation equivalent in humans; the dosage of radiation that will cause the same biological effect as one roentgen of X rays.

renaturation The process by which a denatured protein or nucleic acid returns to its normal three-dimensional structure.

repetitive DNA sequence A DNA sequence present in many copies in the haploid genome.

replication fork The Y-shaped region of a chromosome associated with the site of DNA replication.

replicon The unit of DNA replication, beginning with DNA sequences necessary for the initiation of DNA replication. In bacteria, the entire chromosome is a replicon.

replisome The complex of proteins, including DNA polymerase, that assembles at the bacterial replication fork to synthesize DNA.

repressible enzyme system An enzyme or group of enzymes whose synthesis is regulated by the intracellular concentration of certain metabolites.

repressor A protein that binds to a regulatory sequence adjacent to a gene and blocks transcription of the gene.

reproductive isolation Absence of interbreeding between populations, subspecies, or species.

resistance transfer factor (RTF) A component of R plasmids that confers the ability to transfer the R plasmid between bacterial cells by conjugation.

restriction endonuclease A bacterial nuclease that recognizes specific nucleotide sequences in a DNA molecule, often a *palindrome*, and cleaves or nicks the DNA at those sites.

restriction fragment length polymorphism (RFLP) Variation in the length of DNA fragments generated by restriction endonucleases. These variations are caused by mutations that create or abolish cutting sites for restriction enzymes. RFLPs are inherited in a codominant fashion and are extremely useful as genetic markers.

restriction map A map of restriction enzyme cutting sites in a sequence of DNA.

restriction site A DNA sequence, often palindromic, recognized by a restriction endonuclease. The enzyme binds to the restriction site and cleaves the DNA at that site.

retrotransposon Mobile genetic elements that are major components of many eukaryotic genomes; these elements are copied by means of an RNA intermediate and can be inserted at a distant chromosomal site.

retrovirus A type of virus that uses RNA as its genetic material and employs the enzyme reverse transcriptase during its life cycle.

reverse genetics An experimental approach used to discover gene function after the gene has been identified, cloned, and sequenced.

reversion A mutation that restores the wild-type phenotype.

R factor (R plasmid) A bacterial plasmid that carries antibiotic resistance genes. Most R plasmids have two components: an r-determinant, which carries the antibiotic resistance genes, and the resistance transfer factor (RTF).

Rh factor An antigenic system first described in the rhesus monkey.

ribonucleic acid (RNA) A nucleic acid similar to DNA but characterized by the pentose sugar ribose, the pyrimidine uracil, and the single-stranded nature of the polynucleotide chain. Several forms are recognized, including ribosomal RNA, messenger RNA, transfer RNA, and a variety of small regulatory RNA molecules.

ribonucleoprotein (RNP) particles Complexes of RNA-binding proteins that regulate associated mRNAs.

ribose The five-carbon sugar associated with ribonucleosides and ribonucleotides associated with RNA.

ribosomal RNA (rRNA) The RNA molecules that are the structural components of the ribosomal subunits. In prokaryotes, these are the 16*S*, 23*S*, and 5*S* molecules; in eukaryotes, they are the 18*S*, 28*S*, and 5*S* molecules.

ribosome A ribonucleoprotein organelle consisting of two subunits, each containing RNA and protein molecules. Ribosomes are the site of translation of mRNA codons into the amino acid sequence of a polypeptide chain.

ribozymes RNAs that catalyze specific biochemical reactions.

RNA-binding proteins (RBPs) Proteins that bind to RNAs and have various activities such as influencing RNA translation, splicing, stability, degradation, and localization.

RNA-directed RNA polymerase (RdRP) An enzyme that uses an RNA molecule as a template to synthesize the formation of a complementary RNA strand.

RNA-induced gene silencing A mechanism by which small noncoding RNAs and the RNA-induced silencing complex (RISC) negatively regulate transcription or negatively regulate mRNAs post transcriptionally.

RNA-induced silencing complex (RISC) A protein complex containing an Argonaute family protein with endonuclease activity. siRNAs and miRNAs guide RISC to complementary mRNAs to cleave them.

RNA-induced transcriptional silencing (RITS) A mechanism by which small noncoding RNAs direct a protein complex to complementary DNA sequences to methylate nearby histones and thus silence transcription from this locus in the genome.

RNA interference (RNAi) Inhibition of gene expression in which a protein complex (RNA-induced silencing complex, or RISC), containing a partially complementary RNA strand binds to an mRNA, leading to degradation or reduced translation of the mRNA.

RNA World Hypothesis A hypothesis describing life initially based on RNA serving informational (genomes) and catalytic (ribozymes) roles that predates current life forms based on DNA genomes and protein catalysts.

Robertsonian translocation A chromosomal aberration created by breaks in the short arms of two acrocentric chromosomes followed by fusion of the long arms of these chromosomes at the centromere. Also called *centric fusion*.

roentgen (R) A unit of measure of radiation emission, corresponding to the amount that generates 2.083×10^9 ion pairs in one cubic centimeter of air at 0°C and an atmospheric pressure of 760 mm of mercury.

Sanger sequencing DNA sequencing by synthesis of DNA chains that are randomly terminated by incorporation of a nucleotide analog (dideoxynucleotides) followed by sequence determination by analysis of resulting fragment lengths in each reaction.

satellite DNA DNA that forms a minor band when genomic DNA is centrifuged in a cesium salt gradient. This DNA usually consists of short repetitive sequences.

secondary sex ratio The ratio of males to females at birth, usually expressed in decimal form (e.g., 1.05).

segment polarity genes Genes that regulate the spatial pattern of differentiation within each segment of the developing *Drosophila* embryo.

segregation The separation of maternal and paternal homologs of each homologous chromosome pair into gametes during meiosis.

selection Changes in the frequency of alleles and genotypes in populations as a result of differential reproduction.

selection coefficient (s) A quantitative measure of the relative fitness of one genotype compared with another. Same as *coefficient of selection*.

semiconservative replication A mode of DNA replication in which a double-stranded molecule replicates in such a way that the daughter molecules are each composed of one parental (old) and one newly synthesized strand.

semisterility A condition in which a percentage of all zygotes are inviable.

sex chromosome A chromosome, such as the X or Y in humans, which is involved in sex determination.

sexduction Transmission of chromosomal genes from a donor bacterium to a recipient cell by means of the F factor.

sex-influenced inheritance Phenotypic expression conditioned by the sex of the individual. A heterozygote may express one phenotype in one sex and an alternate phenotype in the other sex (e.g., pattern baldness in humans).

sex-limited inheritance A trait that is expressed in only one sex even though the trait may not be X-linked.

Shine–Dalgarno sequence The nucleotides AGGAGG that serve as a ribosome-binding site in the leader sequence of prokaryotic genes. The 16*S* RNA of the small ribosomal subunit contains a complementary sequence to which the mRNA binds.

short interspersed elements (SINEs) Repetitive sequences found in the genomes of higher organisms. The 300-bp *Alu* sequence is a SINE.

shotgun cloning The cloning of random fragments of genomic DNA into a vector (a plasmid or phage), usually to produce a library from which clones can be selected for use, as in sequencing.

sibling species Species that are morphologically similar but reproductively isolated from one another.

sigma (σ) factor In RNA polymerase, a polypeptide subunit that recognizes the DNA binding site for the initiation of transcription.

single-nucleotide polymorphism (SNP) A variation in a single nucleotide pair in DNA, usually detected during genomic analysis. Present in at least 1 percent of a population, a SNP is useful as a genetic marker.

single-stranded binding proteins (SSBs) In DNA replication, proteins that bind to and stabilize the single-stranded regions of DNA that result from the action of unwinding proteins.

siRNAs Small (or short) interfering RNAs. Short 20- to 25-nucleotide-long double-stranded RNA sequences with 2 nucleotide

3′ overhangs processed by Dicer that participate in transcriptional and/or post transcriptional mechanisms of gene regulation.

sister chromatid exchange (SCE) A crossing over event in meiotic or mitotic cells involving the reciprocal exchange of chromosomal material between sister chromatids joined by a common centromere. Such exchanges can be detected cytologically after BrdU incorporation into the replicating chromosomes.

site-directed mutagenesis A process that uses a synthetic oligonucleotide containing a mutant base or sequence as a primer for inducing a mutation at a specific site in a cloned gene.

small noncoding RNAs (sncRNAs) A general term used to describe short RNAs that do not encode for polypeptides and associate with the RNA-induced silencing complex (RISC) to regulate transcription or to regulate mRNAs post-transcriptionally.

small nuclear RNA (snRNA) Abundant species of small RNA molecules ranging in size from 90 to 400 nucleotides that in association with proteins form RNP particles known as snRNPs or *snurps*. Located in the nucleoplasm, snRNAs have been implicated in the processing of pre-mRNA and may have a range of cleavage and ligation functions.

solenoid structure A feature of eukaryotic chromatin conformation generated by nucleosome supercoiling.

somatic mutation A nonheritable mutation occurring in a somatic cell.

SOS response The induction of enzymes for repairing damaged DNA in *Escherichia coli*. The response involves activation of an enzyme that cleaves a repressor, activating a series of genes involved in DNA repair.

Southern blotting A technique developed by Edwin Southern in which DNA fragments produced by restriction enzyme digestion are separated by electrophoresis and transferred by capillary action to a nylon or nitrocellulose membrane. Specific DNA fragments can be identified by hybridization to a complementary radioactively labeled nucleic acid probe using the technique of *autoradiography*.

spacer DNA DNA sequences found between genes. Usually, these are repetitive DNA segments.

species A group of actually or potentially interbreeding individuals that is reproductively isolated from other such groups.

spectral karyotype A display of all the chromosomes in an organism as a karyotype with each chromosome stained in a different color.

spliceosome The nuclear macromolecule complex within which splicing reactions occur to remove introns from pre-mRNAs.

spontaneous mutation A mutation that arises in the absence of an external force; one that is not induced.

spore A unicellular body or cell encased in a protective coat. Produced by some bacteria, plants, and invertebrates, spores are capable of surviving in unfavorable environmental conditions and give rise to a new individual upon germination. In plants, spores are the haploid products of meiosis.

Src A protein kinase that phosphorylates many target proteins to regulate their activity such as regulating the activity of the RNA-binding protein ZBP1.

sRNA Small noncoding RNA in prokaryotes that regulate gene expression by binding to mRNAs to influence translation or by binding to proteins and modifying their function.

SRY The sex-determining region of the Y chromosome, found near the chromosome's pseudoautosomal boundary. Accumulated evidence indicates that this gene's product is the testis-determining factor (TDF).

stabilizing selection Preferential reproduction of individuals with genotypes close to the mean for the population. A selective elimination of genotypes at both extremes.

standard deviation (*s*) A quantitative measure of the amount of variation in a sample of measurements from a population calculated as the square root of the variance.

stone age genomics Sequence analysis of ancient DNA samples.

strain A group of organisms with common ancestry with physiological or morphological characteristics of interest for genetic analysis or domestication.

STR sequences Short tandem repeats 2–9 base pairs long that are found within minisatellites. These sequences are used to prepare DNA profiles in forensics, paternity identification, and other applications.

structural gene A gene that encodes the amino acid sequence of a polypeptide chain.

sublethal gene A mutation causing lowered viability, with death before maturity in less than 50 percent of the individuals carrying the gene.

submetacentric chromosome A chromosome with the centromere placed so that one arm of the chromosome is slightly longer than the other.

sum law In statistics, the probability of one of two mutually exclusive outcomes occurring, where that outcome can be achieved by two or more events, being equal to the sum of their individual probabilities.

suppressor mutation A mutation that acts to completely or partially restore the function lost by a mutation at another site.

sympatric speciation Speciation occurring in populations that inhabit, at least in part, the same geographic range.

synapsis The pairing of homologous chromosomes at meiosis.

synaptonemal complex (SC) A submicroscopic structure consisting of a tripartite nucleoprotein ribbon that forms between the paired homologous chromosomes of the first meiotic division.

syndrome A group of characteristics or symptoms associated with a disease or an abnormality. An affected individual may express a number of these characteristics but not necessarily all of them.

synthetic biology A field of research that combines science and engineering to construct novel biological-based systems and/or organisms that do not exist in nature.

systems biology A field that identifies and analyzes gene and protein networks to gain an understanding of intracellular regulation of metabolism, intra- and intercellular communication, and complex interactions within, between, and among cells.

TATA box See *Goldberg–Hogness box*.

tautomeric shift A reversible isomerization in a molecule, brought about by a shift in the location of a hydrogen atom. In nucleic acids, tautomeric shifts in the bases of nucleotides can cause changes in other bases during replication and can act as a source of mutations.

telocentric chromosome A chromosome in which the centromere is located at its very end.

telomerase The enzyme that adds short, tandemly repeated DNA sequences to the ends of eukaryotic chromosomes.

telomere The heterochromatic terminal region of a chromosome.

telomere repeat-containing RNA (TERRA) Large noncoding RNA molecules transcribed from telomere repeats that are an integral part of telomeric heterochromatin.

temperate phage A bacteriophage that can become a prophage, integrating its DNA into the chromosome of the host bacterial cell and making the latter lysogenic.

temperature-sensitive mutation A conditional mutation that produces a mutant phenotype at one temperature range and a wild-type phenotype at another.

template The single-stranded DNA or RNA molecule that specifies the sequence of a complementary nucleotide strand synthesized by DNA or RNA polymerase.

testcross A cross between an individual whose genotype at one or more loci may be unknown and an individual who is homozygous recessive for the gene or genes in question.

tetrad The four chromatids that make up paired homologs in the prophase of the first meiotic division. In eukaryotes with a predominant haploid stage (some algae and fungi), a tetrad also denotes the four haploid cells produced by a single meiotic division.

tetrad analysis A method that analyzes gene linkage and recombination in organisms with a predominant haploid phase in their life cycle.

tetranucleotide hypothesis An early theory of DNA structure proposing that the molecule was composed of repeating units, each consisting of the four nucleotides represented by adenine, thymine, cytosine, and guanine.

third-generation sequencing (TGS) Technologies based on high-throughput methods that sequence a single-stranded DNA molecule.

thymine dimer In a polynucleotide strand, a lesion consisting of two adjacent thymine bases that become joined by a covalent bond. Usually caused by exposure to ultraviolet light, this lesion inhibits DNA replication.

totipotent The capacity of a cell or an embryo part to differentiate into all cell types characteristic of an adult. This capacity is usually progressively restricted during development. Used interchangeably with *pluripotent*.

trait Any detectable phenotypic variation of a particular inherited character.

***trans*-acting element** A gene product (usually a diffusible protein or an RNA molecule) that acts to regulate the expression of a target gene.

***trans* configuration** An arrangement in which two mutant alleles are on opposite homologs, such as

$$\frac{a^1 \quad +}{+ \quad a^2}$$

transcription Transfer of genetic information from DNA by the synthesis of a complementary RNA molecule using a DNA template.

transcriptome The set of mRNA molecules present in a cell at any given time.

transdetermination Change in developmental fate of a cell or group of cells.

transduction Virally mediated bacterial recombination. Also used to describe the transfer of eukaryotic genes mediated by a retrovirus.

transfer RNA (tRNA) A small ribonucleic acid molecule that "adapts" a triplet codon to its corresponding amino acid during translation.

transformation Heritable change in a cell or an organism brought about by exogenous DNA. Known to occur naturally and also used in *recombinant DNA* studies.

transgenic organism An organism whose genome has been modified by the introduction of external DNA sequences into the germ line (sometimes called knock-in organism).

transition mutation A mutational event in which one purine is replaced by another or one pyrimidine is replaced by another.

translation The derivation of the amino acid sequence of a polypeptide from the base sequence of an mRNA molecule in association with a ribosome and tRNAs.

translocation A chromosomal mutation associated with the reciprocal or nonreciprocal transfer of a chromosomal segment from one chromosome to another. Also denotes the movement of mRNA through the ribosome during translation.

transmission genetics The field of genetics concerned with heredity and the mechanisms by which genes are transferred from parent to offspring.

transposable element A DNA segment that moves to other sites in the genome, essentially independent of sequence homology. Usually, such elements are flanked at each end by short inverted repeats of 20–40 base pairs.

transversion mutation A mutational event in which a purine is replaced by a pyrimidine or a pyrimidine is replaced by a purine.

trinucleotide repeat A tandemly repeated cluster of three nucleotides (such as CTG) within or near a gene. Certain diseases (myotonic dystrophy, Huntington disease) are caused by expansion in copy number of such repeats.

triploidy The condition in which a cell or an organism possesses three haploid sets of chromosomes.

trisomy The condition in which a cell or an organism possesses two copies of each chromosome except for one, which is present in three copies (designated $2n + 1$).

tumor-suppressor gene A gene that encodes a product that normally functions to suppress cell division. Mutations in tumor-suppressor genes activate cell division and cause tumor formation.

two-dimensional gel electrophoresis (2DGE): Method for separating polypeptides in two dimensions, first by size (molecular weight) and second by electrical charge (isoelectric point).

unequal crossing over A crossover between two improperly aligned homologs, producing one homolog with three copies of a region and the other with one copy of that region.

variable number tandem repeats (VNTRs) Short, repeated DNA sequences (of 2–20 nucleotides) present as tandem repeats between two restriction enzyme sites. Variation in the number of repeats creates DNA fragments of differing lengths following restriction enzyme digestion. Used in early versions of *DNA fingerprinting*.

variance (s^2) A statistical measure of the variation of values from a central value, calculated as the square of the standard deviation.

variegation Patches of differing phenotypes, such as color, in a tissue.

vector In recombinant DNA, an agent such as a phage or plasmid into which a foreign DNA segment will be inserted and used to transform host cells.

vertical gene transfer The transfer of genetic information from parents to offspring generation after generation.

virulent phage A bacteriophage that infects, replicates within, and lyses bacterial cells, releasing new phage particles.

Western blotting An analytical technique in which proteins are separated by gel electrophoresis and transferred by capillary action to a nylon membrane or nitrocellulose sheet. A specific protein can be identified through hybridization to a labeled antibody.

wild type The most commonly observed phenotype or genotype, designated as the norm or standard.

wobble hypothesis An idea proposed by Francis Crick, stating that the third base in an anticodon can align in several ways to allow it to recognize more than one base in the codons of mRNA.

W, Z chromosomes The sex chromosomes in species where the female is the heterogametic sex (WZ).

X chromosome The sex chromosome present in species where the female is the homogametic sex (XX).

X chromosome inactivation In mammalian females, the random cessation of transcriptional activity of the maternally or paternally derived X chromosome. This event, which occurs early in development, is a mechanism of dosage compensation, and all progeny cells inactivate the same X chromosome.

XIST A locus in the X-chromosome inactivation center that controls inactivation of the X chromosome in mammalian females.

X-linkage The pattern of inheritance resulting from genes located on the X chromosome.

X-ray crystallography A technique for determining the three-dimensional structure of molecules by analyzing X-ray diffraction patterns produced by bombarding crystals of the molecule under study with X-rays.

YAC Yeast artificial chromosome. A cloning vector constructed using chromosomal components including telomeres (from a ciliate), and centromeres, origin of replication, and marker genes from yeast. YACs are used to clone long stretches of eukaryotic DNA.

Y chromosome The sex chromosome in species where the male is heterogametic (XY).

Z-DNA An alternative "zig-zag" structure of DNA in which the two antiparallel polynucleotide chains form a left-handed double helix. Implicated in regulation of gene expression.

zinc finger A class of DNA-binding domains seen in proteins. They have a characteristic pattern of cysteine and histidine residues that complex with zinc ions, throwing intermediate amino acid residues into a series of loops or fingers.

zip code A sequence found in some mRNAs that serves as a binding site for RNA-binding proteins that influence mRNA localization within the cell and localized translation.

zip code binding protein 1 (ZBP1) An RNA-binding protein that binds to actin mRNAs and regulates where they are translated within the cell.

zygote The diploid cell produced by the fusion of haploid gametic nuclei.

Credits

PHOTOS

Chapter 1 01-COa, Sinclair Stammers/Science Source; 01-COb, Mark Smith/Science Source; 01-COc, Alberto Salguero; F01-02, Biophoto Association/Science Source; F01-03, Photo Researchers/Science Source; F01-07, Eye of Science/Science Source; F01-08, Roslin Institute; F01-09a, Pearson Education; F01-09b, Hermann Eisenbeiss/Science Source; F01-10a, Jeremy Burgess/Science Source; F01-10b, David Mc Carthy/Science Source

Chapter 2 02-CO, Dr. Andrew S. Bajer, University of Oregon;F02-02, CNRI/SPL/Science Source; F02-04, Dr. David Ward, Yale University; F02-12a, Biophoto Associates/Science Source; F02-12b, Andrew Syred/Science Source; F02-12c, Science Source

Chapter 3 03-CO National Library of Medicine

Chapter 4 04-CO, Juniors Bildarchiv/GmbH/Stock Photo/Alamy; P78, Dr. Ralph Somes; J. James Bitgood, University of Wisconsin, Animal Sciences Department; Dr. Ralph Somes; J. James, Bitgood, University of Wisconsin, Animal Sciences Department; P82, Shout It Out Design/Shutterstock; Zuzule/Shutterstock; Julia Remezova/Shutterstock; F04-01a, John Kaprielian/Science Source; F04-03b, Jackson Laboratory; F04-03c, Jackson Laboratory; F04-08, RoJo Images/Shutterstock; F04-12a, Prisma Bildagentur/AG/Stock Photo/Alamy; F04-13, Hans Rienhard/Photoshot Holdings Ltd.; F04-14, Debra P. Hershkowitz/Photoshot Holdings Ltd.; F04-15a, Tanya Wolff; F04-15b, Tanya Wolff; F04-15c, Tanya Wolff; F04-16a, Jane Burton/Photoshot Holdings Ltd.; F04-16b, Dr. William S. Klug; F04-17b, Aida Ricciardiello/Shutterstock; F04-18a, Dr. Ronald A. Butow, Department of Molecular Biology and Oncology, University of Texas Southwestern Medical Center; F04-18b, Dr. Ronald A. Butow, Department of Molecular Biology and Oncology, University of Texas Southwestern Medical Center

Chapter 5 05-CO, Wessex Reg/Genetics Centre/Wellcome Images; P92, Texas A&M University/AP Images; F05-02a, Catherine G. Palmer; F05-8a Maria Gallegos; F05-04a, Michael Abbey/Science Source;F05-06a, Sari ONeal/Shutterstock; F05-06b, William S. Klug; F05-02b, Catherine G. Palmer; F05-04b, Michael Abbey/Science Source

Chapter 6 06-CO, Evelin Schrock, Stan du Manoir,Tom Reid/National Institutes of Health; F06-02a, Courtesy of the Greenwood Genetic Center, Greenwood, SC; F06-02b, Kristy-Anne Glubish/Alamy; F06-04a, David D. Weaver/Indiana University; F06-08, Courtesy of National Cotton Council of America; F06-11a, Douglas Chapman, University of Washington Medical Center Pathology; F06-11b, Cri du chat Syndrome Support Group, UK; F06-13, Mary Lilly, The Observatories for the Carnegie Institution for Science; F06-17b, Jorge J. Yunis; F06-18, Professor Christine Harrison, Newcastle University, UK

Chapter 7 07-CO,James Kezer C/O Stanley Sessions; F07-14, Dr. Sheldon Wolff & Judy Bodycote/Laboratory of Radiobiology and Environmental Health, University of California, San Francisco

Chapter 8 08-CO, Dr. L. Caro/Science Source; F08-01, Pearson Education; F08-10a, Gopal Murti/Science Source; F08-11a, M. Wurtz/Science Source; F08-13b, Christine Case

Chapter 9 09-CO, Ken Eward/Photo Researchers; F09-03b, Oliver Meckes/Science Source; F09-10,M.H.F. Wilkins; F09-13, Ventana Medical Systems, Inc.

Chapter 10 10-CO,Gopal Murti/Science Source; F10-14, Dr. Harold Weintraub, Howard Hughes Medical Institute, Fred Hutchinson Cancer Center/"Essential Molecular Biology" 2e, Freifelder & Malachinski, Jones & Bartlett, Fig. 7-24, pp. 141

Chapter 11 11-CO, Don. W. Fawcett/Science Source; F11-01a, M. Wurtz/Science Source; F11-01b, William S. Klug; F11-02, Biology Pics/Science Source; F11-03, G. Murti/Science Source; F11-04, Don W. Fawcett/Science Source; F11-05, Dr. Richard D. Kolodnar, Dana-Farber Cancer Institute; F11-06a, Harald Eggert; F11-06b, The Company of Biologists; F11-07b, John Ellison, Richardson Lab, Integrative Biology, The University of Texas at Austin; F11-08a, Omikron/Science Source; F11-08b, William S. Klug; F11-09, William S. Klug; F11-13, William S. Klug

Chapter 12 12-CO, Oscar L. Miller/Science Source; F12-10a Bert W. O'Malley, M.D., Baylor College of Medicine

Chapter 13 FF13-09a, Cold Spring Harbor Laboratory; F13-09b, Elena Kiseleva; F13-11, Sebastian Kaulitzki/Shutterstock; F13-15, Kenneth Eward/Photo Researchers; F13-CO, From Science, Marat M. Yusupov, Gulnara Zh. Yusupova, Albion Baucom, Kate Lieberman, Thomas N. Earnest, J. H. D. Cate, Harry F. Noller. Crystal Structure of the Ribosome at 5.5 Resolution. Reprinted with permission from AAAS

Chapter 14 14-CO, M.G. Neuffer; F14-13, The Xeroderma Pigmentosum Association, otherwise known as The Children of the Moon, organized a trip to Paris on December 17 and 18, 2005 for children with XP. Newscom Photo. Used with permission

Chapter 15 15-CO, T. Cremer/Dr. I. Solovei/Dr. F. Haberman/Biozentrum (LMU)

Chapter 16 16-CO, SPL/Science Source; F16-01a, Courtesy of Hesed M. Padilla-Nash, Antonio Fargiano, and Thomas Ried. Affiliation is Section of Cancer Genomics, Genetics Branch, Center for Research, National Cancer Institute, National Institutes of Health, Bethesda, MD; F16-01b, Courtesy of Hesed M. Padilla-Nash, Antonio Fargiano, and Thomas Ried. Affiliation is Section of Cancer Genomics, Genetics Branch, Center for Research, National Cancer Institute, National Institutes of Health, Bethesda, MD; F16-02a, Roland Birke/Getty Images; F16-02b, Biophoto/Science Source; F16-02c, Biophoto Associates/Science Source; F16-02d, Biophoto Associates/Science Source

Chapter 17 17-CO, Pascal Goetgheluck/Science Source; F17-03a, Gopal Murti/Science Source; F17-05b, Michael Gabridge/Custom Medical Stock Photo; F17-10a, Proceedings of the National Academy of Sciences; F17-10b, Proceedings of the National Academy of Sciences; F17-11, Gerald B. Downes; F17-14a, Scientist microinjecting cloned DNA into a fertilized egg. Copyright 2014 The Regents of the University of California. Used by permission; F17-14b, Courtesy of Ralph Brinster. University of Pennsylvania School of Veterinary Medicine; F17-15, Eye of Science/Science Source;

Chapter 18 F18-11a, Volker Steger/Science Source; F18-11b, Dra Schwartz/Getty Images; F18-12, Swiss Institute of Bioinformatics

Chapter 19 F19-01a, SIU Biomed Com/Custom Medical Stock Photo; F19-02,From Doebley, J. Plant Cell, 2005 Nov; 17(11): 2859-72. Courtesy of John Doebley/University of Wisconsin; F19-CO, GloFish® Fluorescent Fish; F19-09, Affymetrix; F19-10, National Institutes of Health-National Cancer Institute; F19-11 top left, Sebastian Kaultizki/Shutterstock; F19-11 top center, C.D. Humphrey T.G. Ksiazek, Centers for Disease Control and Prevention; F19-11 top right, Janice Haney Carr, Centers for Disease Control and Prevention; F19-11 mice, C.D. Humphrey T.G. Ksiazek, Centers for Disease Control and Prevention; F19-11 gene chips, Affymetrix; F19-11 microarray, M. C. Lorence

Chapter 20 F20-05 Jim Langeland, Stephen Paddock, and Sean Carroll, University of Wisconsin at Madison; F20-07a Peter A. Lawrence; F20-07b, Peter A. Lawrence; F20-08, Jim Langeland, Stephen Paddock, and Sean Carroll, University of Wisconsin at Madison; F20-09a © 2001 The Rockefeller University Press et al. The Journal of Cell Biology. VOL: 153 no. 1 87-100. doi: 10.1083/jcb.153.1.87; F20-09b, © 2001 The Rockefeller University Press et al. The Journal of Cell Biology. VOL: 153 no. 1 87-100. doi: 10.1083/jcb.153.1.87; 20-10a, F Rudolf Turner/Indiana University; F20-10b, F Rudolf Turner/Indiana University; F20-14, P. Barber,RBP /CustomMedical; F20-15a, Darwin Dale/Science Source; F20-15b, Tanya Wolff; F20-15c, Urs Kloter/Georg Halder; F20-16 Dr. Elliot M. Meyerowitz; 20-17a, Max-Planck-Institut fur Entwicklungsbiologie; F20-17b, Max-Planck-Institut fur Entwicklungsbiologie; 20-19a, Dr. Elliot M. Meyerowitz; F20-19b, Dr. Elliot M. Meyerowitz; F20-19c, Dr. Elliot M. Meyerowitz; F20-19d Dr. Elliot M. Meyerowitz; F20-CO, Edward B Lewis/Calfornia institute of technology; F21-08 Rekemp/Getty Images

Chapter 21 F21-CO, 2009fotofriends/Shutterstock; 22-01, Pixshots/Shutterstock; 22-06, Michel Sam-son, Frederick Libert, et al., "Resistance to HIV- 1 infection in Caucasian in-dividuals bearing mutant alleles of the CCR- 5 chemokine receptor gene" Reprinted with permission from Nature [vol. 382, 22 August 1996, p. 725, Fig. 3]. © 1996 Macmillan Magazines Limited

Chapter 22 F22-23, Ria Novosti/Science Source; F22-20b, Peter Scoones/Getty Images; 22-13, American Journal of Human Genetics, Elsevier; 22-16, Smithsonian Institution Libraries; 22-17a, Niels Poulsen/Alamy; F22-17b, Niklas Lindqvist/National Ciklid Society; F22-20a, Mchugh Tom/Getty Images; F22-CO, Simon Booth/Science Source

Special Topic 1 ST1-07, Environmental Health Perspectives

Special Topic 2 ST2-09a, Brosnan and Voinnet, 2011 Current Opinions in Plant Biology (Fig 1B) Original experiment: Brosnan et al 2007 PNAS; ST2-09b, Brosnan and Voinnet, 2011 Current Opinions in Plant Biology (Fig 1B) Original experiment: Brosnan et al 2007 PNAS

Special Topic 3 ST3-05, Viessmann Manufacturing Company, Inc.

Special Topic 4 ST4-02a, Reproduced with permission of Dako Denmark A/S, a subsidiary of Agilent Technologies, Inc., Santa Clara, California, USA. All rights reserved; ST4-02b, Reproduced with permission of Dako Denmark A/S, a subsidiary of Agilent Technologies, Inc., Santa Clara, California, USA. All rights reserved

Special Topic 5 ST5-01, Pascal Pavani/AFP/Getty Images; ST5-03, International Rice Research Institute; ST5-06a, Helios Gene Gun Courtesy of Bio-Rad Laboratories, Inc.; ST5-09, Bob Hartzler, Department of Agronomy, Iowa State University; ST5-Box 1, AquaBounty Technologies Inc; ST5-Box 2, S.A. Ferreira/US Pacific Basin Agricultural Research Center

Special Topic 6 ST6–04a, Van DeSilva; ST6-06, Jennifer Doudna, H. Adam Steinberg, artforscience.com

TEXT

Chapter 1 p. 019, N. Hartsoeker, Essay de dioptrique Paris, 1694, p. 246. National Library of Medicine © 1964 National Library of Medicine

Chapter 7 p. 140, 7A History of Genetics, by Alfred H. Sturtevant. New York: Harper & Row, 1965

Chapter 10 p. 213, Geron Corporation

Chapter 12 p. 234, M.W. Nirenberg and H.J. Matthaei (1961). "The Dependence Of Cell-Free Protein Synthesis In E. coli Upon Naturally Occurring Or Synthetic Polyribonucleotides". Proceedings of the National Academy of Sciences of the USA 47 (10): 1588–160

Chapter 16 p. 326, Data from Vogelstein, B. et al. 2013. Cancer Genome Landscapes. Science 339: 1546-1558

Chapter 18 p. 378, Adapted from Palladino, M. A. Understanding the Human Genome Project, 2nd ed. Benjamin Cummings, 2006.

Chapter 19 p. 410, "A genome-wide association study of brain lesion distribution in multiple sclerosis", Pierre-Antoine Gourraud, Michael Sdika, Pouya Khankhanian, Roland G. Henry, Azadeh Beheshtian, Paul M. Matthews, Stephen L. Hauser, Jorge R. Oksenberg, Daniel Pelletier and Sergio E. Baranzini. Brain, a Neurology Journal, February 13, 2013. Published by Oxford University Press on behalf of the Guarantors of Brain. All rights reserved. **p.417**, Data from Lee et al. 2001. Infect. and Immunity 69: 5786–5793.

Chapter 20 p. **432**, Gerhart, J. 1999. 1998 Warkany lecture: Signaling pathways in development. Teratology 60: 226–239

Chapter 21 p. 452, "Mapping Polygenes" by S.D. Tanksley, from ANNUAL REVIEW OF GENETICS, December 1993, Volume 27. Reproduced with permission of Annual Review of Genetics, Volume © 1993 by Annual Reviews, http://www.annualreviews.org"

Chapter 22 p. 459, From Noonan, J.P. et al. 'Sequencing and analysis of Neanderthal genomic DNA.' SCIENCE 314: 1113–1118

Special Topic 2 p. 492, Based on http://mcb.illinois.edu/faculty/profile/cfanderp. **p.500**, Pearson Education, Inc. **p.499**, Wang et. al. The Long Arm of Long Noncoding RNAs: Roles as Sensors Regulating Gene Transcriptional Programs. Cold Spring Harbor. Perspectives of Biology 2011 Jan; 3(1):a003756;5. Copyright Cold Spring Harbor Laboratory Press. Reprinted with permission

Special Topic 3 p. 505, Reprinted with permission form The Journal of Forensic Sciences, Vol. 48, No. 4, copyright ASTM International, 100 Barr Harbor Drive, West Conshohocken, PA 19428

Special Topic 4 p. 519, Based on a story reported in Kolata, G., In treatment for leukemia, glimpses of the future. New York Times, July 7, 2012. **p.514**, "Personalized Medicine Coalition." The Case for Personalized Medicine: 4th ed, fig.2 Copyright (c) 2014 Personalized Medical Coalition. Used by permission. **p.516**, "Personalized Medicine Coalition." The Case for Personalized Medicine: 4th ed., fig.2 Copyright (c) 2014 Personalized Medical Coalition. Used by permission

Special Topic 5 p. 524, Information from the International Service for the Acquisition of Agri-Biotech Applications, www.isaaa.org

Index

Note: Pages numbers followed by f, t, and b indicate figures, tables, and boxes, respectively.

C

D

E

F

G

H

I

N

O

P

Q

R

S

T

X

Y

Z

EVOLVING CONCEPT OF A GENE

The Evolving Concept of the Gene is a new feature, integrated in key chapters, that highlights how scientists' understanding of the gene has changed over time. By underscoring how the conceptualization of the gene has evolved, our goal is to help students appreciate the process of discovery that has led to an ever more sophisticated understanding of hereditary information.

CHAPTER 3 pg. 58 Based on the pioneering work of Gregor Mendel, we can view the gene as a heritable unit factor that determines the expression of an observable trait, or phenotype. ■

CHAPTER 4 pg. 74 Based on the work of many geneticists following the rediscovery of Mendel's work in the very early part of the twentieth century, the chromosome theory of inheritance was put forward, which hypothesized that chromosomes are the carriers of genes and that meiosis is the physical basis of Mendel's postulates. In the ensuing 40 years, the concept of a gene evolved to reflect that this hereditary unit can exist in multiple forms, or alleles, each of which can impact on the phenotype in different ways, leading to incomplete dominance, codominance, and even lethality. It became clear that the process of mutation was the source of new alleles. ■

CHAPTER 7 pg. 152 Based on the gene-mapping studies in *Drosophila* and many other organisms from the 1920s through the mid-1950s, geneticists regarded genes as hereditary units organized in a specific sequence along chromosomes, between which recombination could occur. Genes were thus viewed as indivisible "beads on a string." ■

CHAPTER 9 pg. 190 Based on the model of DNA put forward by Watson and Crick in 1953, the gene was viewed for the first time in molecular terms as a sequence of nucleotides in a DNA helix that encodes genetic information. ■

CHAPTER 12 pg. 249 The elucidation of the genetic code in the 1960s supported the concept that the gene is composed of a linear series of triplet nucleotides encoding the amino acid sequence of a protein. While this is indeed the case in prokaryotes and viruses, in 1977, it became apparent that in eukaryotes, the gene is divided into coding sequences, called exons, which are interrupted by noncoding sequences, called introns (intervening sequences), which must be spliced out during production of the mature mRNA. ■

CHAPTER 13 pg. 267 In the 1940s, a time when the molecular nature of the gene had yet to be defined, groundbreaking work of Beadle and Tatum provided the first experimental evidence concerning the product of genes, their "one-gene:one-enzyme" hypothesis. This idea received further support and was later modified to indicate that one gene specifies one polypeptide chain. ■

CHAPTER 15 pg. 304 The groundbreaking work of Jacob, Monod, and Lwoff in the early 1960s, which established the operon model for the regulation of gene expression in bacteria, expanded the concept of the gene to include noncoding regulatory sequences that are present upstream (5′) from the coding region. In bacterial operons, the transcription of several contiguous structural genes whose products are involved in the same biochemical pathway are regulated by a single set of regulatory sequences. ■

CHAPTER 18 pg. 374 Based on the work of the ENCODE project, we now know that DNA sequences that have previously been thought of as "junk DNA," which do not encode proteins, are nonetheless often transcribed into what we call noncoding RNA (ncRNA). Since the function of some these RNAs is now being determined, we must consider whether the concept of the gene should be expanded to include DNA sequences that encode ncRNAs. At this writing, there is no consensus, but it is important for you to be aware of these current findings ■

A Selection of Nobel Prizes Awarded for Research in Genetics or Genetics-Related Areas

Year	Recipients	Nobel Prize*	Discovery/Research Topic
2012	J.B. Gurdon, S. Yamanaka	P/M	Differentiated cells can be reprogrammed to become pluripotent
2009	E. H. Blackburn C. W. Greider J. W. Szostak	P/M	The nature and replication of the DNA of telomeres, and the discovery of the telomere-replenishing ribonucleoprotein enzyme telomerase
2008	O. Shimomura, M. Chalfie, R. Tsien	C	Discovery and development of the green fluorescent protein (GFP) technology as a tool for genetic research
2007	M. R. Capecchi M. J. Evans O. Smithies	P/M	Gene-targeting technology essential to the creation of knockout mice serving as animal models of human disease
2006	R. Kornberg	C	Molecular basis of eukaryotic transcription
2006	A. Z. Fire, C. C. Mello	P/M	Gene silencing using RNA interference (RNAi)
2002	S. Brenner, H. R. Horvitz, J. E. Sulston	P/M	Genetic regulation of organ development and programmed cell death (apoptosis)
2001	L. Hartwell, T. Hunt, P. Nurse	P/M	Genes and regulatory molecules controlling the cell cycle
1997	S. Prusiner	P/M	Prions, a new biological principle of infection
1995	E. B. Lewis, C. Nusslein-Volhard, E. Wieschaus	P/M	Genetic control of early development in *Drosophila*
1993	R. Roberts, P. Sharp	P/M	RNA processing of split genes
1993	K. Mullis M. Smith	C	Development of polymerase chain reaction (PCR) and site-directed mutagenesis (SDM)
1989	J. M. Bishop, H. E. Varmus	P/M	Role of retroviruses and oncogenes in cancer
1989	T. R. Cech, S. Altman	C	Catalytic properties of RNA
1987	S. Tonegawa	P/M	Genetic basis of antibody diversity
1983	B. McClintock	P/M	Mobile genetic elements in maize
1982	A. Klug	C	Crystalline structure analysis of significant complexes, including tRNA and nucleosomes
1980	P. Berg, W. Gilbert, F. Sanger	C	Development of recombinant DNA and DNA sequencing technology
1978	W. Arber, D. Nathans, H. O. Smith	P/M	Recombinant DNA technology using restriction endonuclease technology
1976	B. S. Blumberg D. C. Gajdusek	P/M	Elucidation of the human prion-based diseases, kuru and Creutzfeldt-Jakob disease
1975	D. Baltimore, R. Delbecco, H. Temin	P/M	Molecular genetics of tumor viruses
1970	N. Borlaug	PP	Genetic improvement of Mexican wheat
1969	M. Delbrück, A. D. Hershey, S. E. Luria	P/M	Replication mechanisms and genetic structure of bacteriophages
1968	H. G. Khorana, M. W. Nirenberg	P/M	Deciphering the genetic code
1968	R. W. Holley	P/M	Structure and nucleotide sequence of transfer RNA
1965	F. Jacob, A. M. Lwoff, J. L. Monod	P/M	Genetic regulation of enzyme synthesis in bacteria
1962	F. H. C. Crick, J. D. Watson, M. H. F. Wilkins	P/M	Double helical model of DNA
1959	A. Kornberg, S. Ochoa	P/M	Biological synthesis of DNA and RNA
1958	G. W. Beadle, E. L. Tatum	P/M	Genetic control of biochemical processes
1958	J. Lederberg	P/M	Genetic recombination in bacteria
1954	L. Pauling	C	Alpha helical structure of proteins
1946	H. J. Muller	P/M	X-ray induction of mutations in *Drosophila*
1933	T. H. Morgan	P/M	Chromosomal theory of inheritance

*C = Chemistry; P/M = Physiology or Medicine; PP = Peace Prize