Next-generation DNA sequencing techniques

Wilhelm J. Ansorge

Ecole Polytechnique Federal Lausanne, EPFL, Switzerland

Next-generation high-throughput DNA sequencing techniques are opening fascinating opportunities in the life sciences. Novel fields and applications in biology and medicine are becoming a reality, beyond the genomic sequencing which was original development goal and application. Serving as examples are: personal genomics with detailed analysis of individual genome stretches; precise analysis of RNA transcripts for gene expression, surpassing and replacing in several respects analysis by various microarray platforms, for instance in reliable and precise quantification of transcripts and as a tool for identification and analysis of DNA regions interacting with regulatory proteins in functional regulation of gene expression. The next-generation sequencing technologies offer novel and rapid ways for genome-wide characterisation and profiling of mRNAs, small RNAs, transcription factor regions, structure of chromatin and DNA methylation patterns, microbiology and metagenomics. In this article, development of commercial sequencing devices is reviewed and some European contributions to the field are mentioned. Presently commercially available very high-throughput DNA sequencing platforms, as well as techniques under development, are described and their applications in bio-medical fields discussed.

Introduction

Next-generation high-throughput DNA sequencing techniques, which are opening fascinating new opportunities in biomedicine, were selected by *Nature Methods* as the method of the year in 2007 [1]. However, the path to gaining acceptance of the novel technology was not an easy one. Until a few years ago the methods used for the sequencing were the Sanger enzymatic dideoxy technique first described in 1977 [2] and the Maxam and Gilbert chemical degradation method described in the same year [3], which was used in sequence cases which could not easily be resolved with the Sanger technique. The two laboratories where the first automated DNA sequencers were produced, simultaneously, were those of Leroy Hood at Caltech [4], commercialised by Applied Biosystems, and Wilhelm Ansorge at the European Molecular Biology Laboratory EMBL [5,6] and commercialised by Pharmacia-Amersham, later General Electric (GE) Healthcare. The Sanger method was used in the first automated fluorescent project for sequencing of a genome region, in which sequence determination of the complete gene locus for the HPRT gene was performed using the EMBL technique; in that project the important concept of paired-end sequencing was also introduced for the first time [7]. The achievement of successful and unambiguous sequencing of a real genomic DNA region, loaded with many sequence pitfalls like Alu sequences in both directions of the HPRT gene locus, demonstrated the feasibility of using an automated fluorescence-based technique for the sequencing of entire genomes, and in principle the feasibility of the technical sequencing part of the Human Genome project.

When the international community decided on determination of the whole human genome sequence, the goal triggered the development of techniques allowing higher sequencing throughput. In Japan, the work on fluorescent DNA sequencing technology by the team of H. Kambara (http://www.hitachi.com/rd/fellow_kambara.html) in the Hitachi laboratories resulted in the development after 1996 of a high-throughput capillary array DNA sequencer. Two

E-mail address: wilhelm.ansorge@epfl.ch.

1871-6784/$ - see front matter © 2009 Published by Elsevier B.V. doi:10.1016/j.nbt.2008.12.009

companies, ABI (commercialising the Kambara system) and Amersham (taking over and developing further the system set up in the US by the Molecular Dynamics company), commercialised automated sequencing using parallel analysis in systems of up to 384 capillaries at that time. Together with partial miniaturisation of the robotic sample preparation, large efforts in automation of laboratory processes and advances in new enzymes and biochemicals, the Sanger technique made possible the determination of the sequence of the human genome by two consortia working in parallel. It was the unique method used for DNA sequencing, with innumerable applications in biology and medicine.

As the users and developers of the DNA sequencing techniques realised, the great limitations of the Sanger sequencing protocols for even larger sequence output were the need for gels or polymers used as sieving separation media for the fluorescently labelled DNA fragments, the relatively low number of samples which could be analysed in parallel and the difficulty of total automation of the sample preparation methods. These limitations initiated efforts to develop techniques without gels, which would allow sequence determination on very large numbers (i.e. millions) of samples in parallel. One of the first developments of such a technique was at the EMBL (at that time one of the two world leaders in DNA sequencing technology) from 1988 to 1990. A patent application by EMBL [8] described a large-scale DNA sequencing technique without gels, extending primers in 'sequencing-by-synthesis, addition and detection of the incorporated base', proposing and describing the use of the so-called 'reversible terminators' for speed and efficiency [8]. The first step of the technique consisted in detecting the next added fluorescently labelled base (reversible terminator) in the growing DNA chain by means of a sensitive CCD camera. This was performed on a large number of DNA samples in parallel, attached either to a planar support or to beads, on DNA chips, minimising reaction volumes in a miniaturised microsystem. In the next step the terminator was converted to a standard nucleotide and the dye removed from it. This cycle and the process were repeated to determine the next base in the sequence. The principle described in the patent application is in part very similar to that used today in the so-called next-generation devices, with many additional original developments commercialised by Illumina-Solexa, Helicos and other companies.

Since 2000, focused developments have continued in several groups. Various institutions, particularly European laboratories, considered the capillary systems as the high point and in a less visionary decision ceased developments of even the most promising novel sequencing techniques, turning their attention exclusively to arrays. By contrast, in the US, funding for development and testing of novel, non-gel-based high-throughput sequencing technologies were provided by the large granting agencies and private companies. Efforts to bring the platforms to maturity were under way. The resulting devices and platforms available on the market in mid-2008, as well as some interesting parallel developments, are described in more detail below. The EU has recently initiated significant support for the development of novel high-throughput DNA sequencing technologies, among others the READNA initiative (www.cng.fr/READNA).

Next-generation DNA sequencing platforms

Novel DNA sequencing techniques provide high speed and throughput, such that genome sequencing projects that took several years with the Sanger technique can now be completed in a matter of weeks. The advantage of these platforms is the determination of the sequence data from amplified single DNA fragments, avoiding the need for cloning of DNA fragments. A limiting factor of the new technology remains the overall high cost for generating the sequence with very high-throughput, even though compared with Sanger sequencing the cost per base is lower by several orders of magnitude. Reduction of sequencing errors is another factor; in this respect the Sanger sequencing technique remains competitive in the immediate future. Other limitations in some applications are short read lengths, non-uniform confidence in base calling in sequence reads, particularly deteriorating 3′-sequence quality in technologies with short read lengths and generally lower reading accuracy in homopolar stretches of identical bases. The huge amount of data generated by these systems (over a gigabase per run) in the form of short reads presents another challenge to developers of software and more efficient computer algorithms.

The 454 GenomeSequencer FLX instrument (Roche Applied Science)

The principle of pyrophosphate detection, the basis of this device, was described in 1985 [9], and a system using this principle in a new method for DNA sequencing was reported in 1988 [10]. The technique was further developed into a routinely functioning method by the teams of M. Ronaghi, M. Uhlen, and P. Nyren in Stockholm [11], leading to a technique commercialised for the analysis of 96 samples in parallel in a microtiter plate.

The GS instrument was introduced in 2005, developed by 454 Life Sciences, as the first next-generation system on the market. In this system (Fig. 1), DNA fragments are ligated with specific adapters that cause the binding of one fragment to a bead. Emulsion PCR is carried out for fragment amplification, with water droplets containing one bead and PCR reagents immersed in oil. The amplification is necessary to obtain sufficient light signal intensity for reliable detection in the sequencing-by-synthesis reaction steps. When PCR amplification cycles are completed and after denaturation, each bead with its one amplified fragment is placed at the top end of an etched fibre in an optical fibre chip, created from glass fibre bundles. The individual glass fibres are excellent light guides, with the other end facing a sensitive CCD camera, enabling positional detection of emitted light. Each bead thus sits on an addressable position in the light guide chip, containing several hundred thousand fibres with attached beads. In the next step polymerase enzyme and primer are added to the beads, and one unlabelled nucleotide only is supplied to the reaction mixture to all beads on the chip, so that synthesis of the complementary strand can start. Incorporation of a following base by the polymerase enzyme in the growing chain releases a pyrophosphate group, which can be detected as emitted light. Knowing the identity of the nucleotide supplied in each step, the presence of a light signal indicates the next base incorporated into the sequence of the growing DNA strand.

The method has recently increased the achieved reading length to the 400–500 base range, with paired-end reads, and as such is being applied to genome (bacterial, animal, human) sequencing. One spectacular application of the system was the identification of the culprit in the recent honey-bee disease epidemics (see company web pages below). A relatively high cost of operation

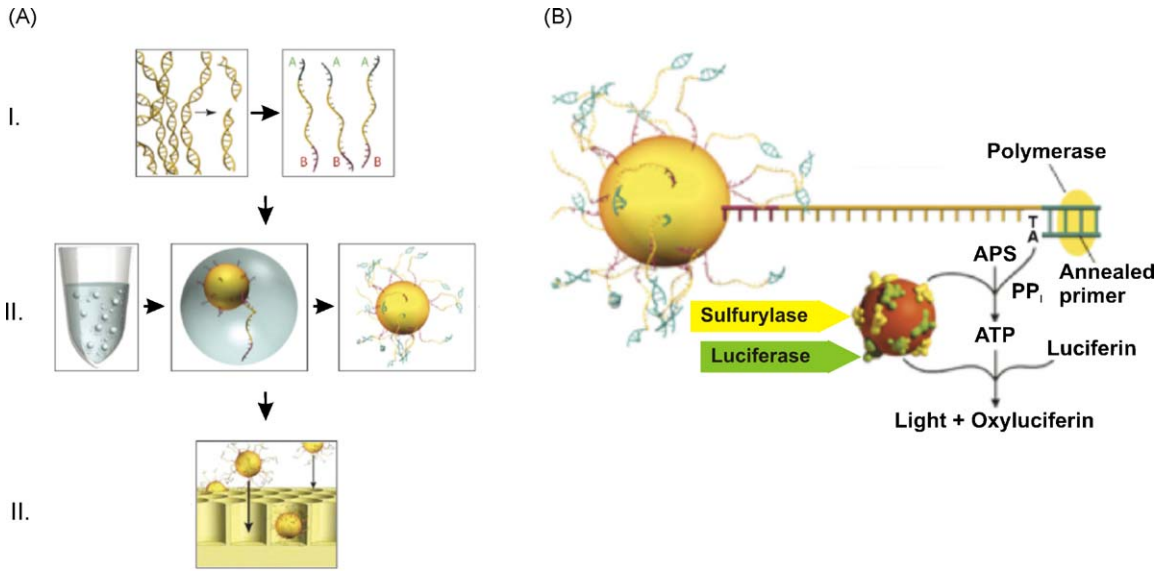

FIGURE 1

(A) Outline of the GS 454 DNA sequencer workflow. Library construction (I) ligates 454-specific adapters to DNA fragments (indicated as A and B) and couples amplification beads with DNA in an emulsion PCR to amplify fragments before sequencing (II). The beads are loaded into the picotiter plate (III). **(B)** Schematic illustration of the pyrosequencing reaction which occurs on nucleotide incorporation to report sequencing-by-synthesis. (Adapted from http://www.454.com.)

and generally lower reading accuracy in homopolar stretches of identical bases are mentioned presently as the few drawbacks of the method. The next upgrade 454 FLX Titanium will quintuple the data output from 100 Mb to about 500 Mb, and the new picotiter plate in the device uses smaller beads about 1 μm diameter. The device, schema of operation, its further developments and list of publications with applications can be found at http://www.454.com/index.asp and in [1].

The Illumina (Solexa) Genome Analyzer

The Solexa sequencing platform was commercialised in 2006, with Illumina acquiring Solexa in early 2007. The principle (Fig. 2) is on the basis of sequencing-by-synthesis chemistry, with novel reversible terminator nucleotides for the four bases each labelled with a different fluorescent dye, and a special DNA polymerase enzyme able to incorporate them. DNA fragments are ligated at both ends to adapters and, after denaturation, immobilised at one end on a solid support. The surface of the support is coated densely with the adapters and the complementary adapters. Each single-stranded fragment, immobilised at one end on the surface, creates a 'bridge' structure by hybridising with its free end to the complementary adapter on the surface of the support. In the mixture containing the PCR amplification reagents, the adapters on the surface act as primers for the following PCR amplification. Again, amplification is needed to obtain sufficient light signal intensity for reliable detection of the added bases. After several PCR cycles, random clusters of about 1000 copies of single-stranded DNA fragments (termed DNA 'polonies', resembling cell colonies after polymerase amplification) are created on the surface. The reaction mixture for the sequencing reactions and DNA synthesis is supplied onto the surface and contains primers, four reversible terminator nucleotides each labelled with a different fluorescent dye and the DNA polymerase. After incorporation into the DNA strand, the terminator nucleotide, as well as its position on the support surface, is detected and identified via its fluorescent dye by the CCD camera. The terminator group at the 3′-end of the base and the fluorescent dye are then removed from the base and the synthesis cycle is repeated. The sequence read length achieved in the repetitive reactions is about 35 nucleotides. The sequence of at least 40 million polonies can be simultaneously determined in parallel, resulting in a very high sequence throughput, on the order of Gigabases per support.

In 2008 Illumina introduced an upgrade, the Genome Analyzer II that triples output compared to the previous Genome Analyzer instrument. A paired-end module for the sequencer was introduced, and with new optics and camera components that allow the system to image DNA clusters more efficiently over larger areas, the new instrument triples the output per paired-end run from 1 to 3 Gb. The system generates at least 1.5 Gb of single-read data per run, at least 3 Gb of data in a paired-end run, recording data from more than 50 million reads per flow cell. The run time for a 36-cycle run was decreased to two days for a single-read run, and four days for a paired-end run. Information on the Genome Analyzer system can be found at http://www.solexa.com/ and in [1].

The Applied Biosystems ABI SOLiD system

The ABI SOLiD sequencing system, a platform using chemistry based upon ligation, was introduced in Autumn 2007. The generation of a DNA fragment library and the sequencing process by subsequent ligation steps are shown schematically in Figs 3,4. In this technique, DNA fragments are ligated to adapters then bound to beads. A water droplet in oil emulsion contains the amplification reagents and only one fragment bound per bead; DNA fragments on the beads are amplified by the emulsion PCR. After DNA denaturation, the beads are deposited onto a glass support surface. In a first step, a primer is hybridised to the adapter. Next, a mixture of oligonucleotide octamers is also hybridised to the DNA fragments and ligation mixture added. In these octamers, the doublet

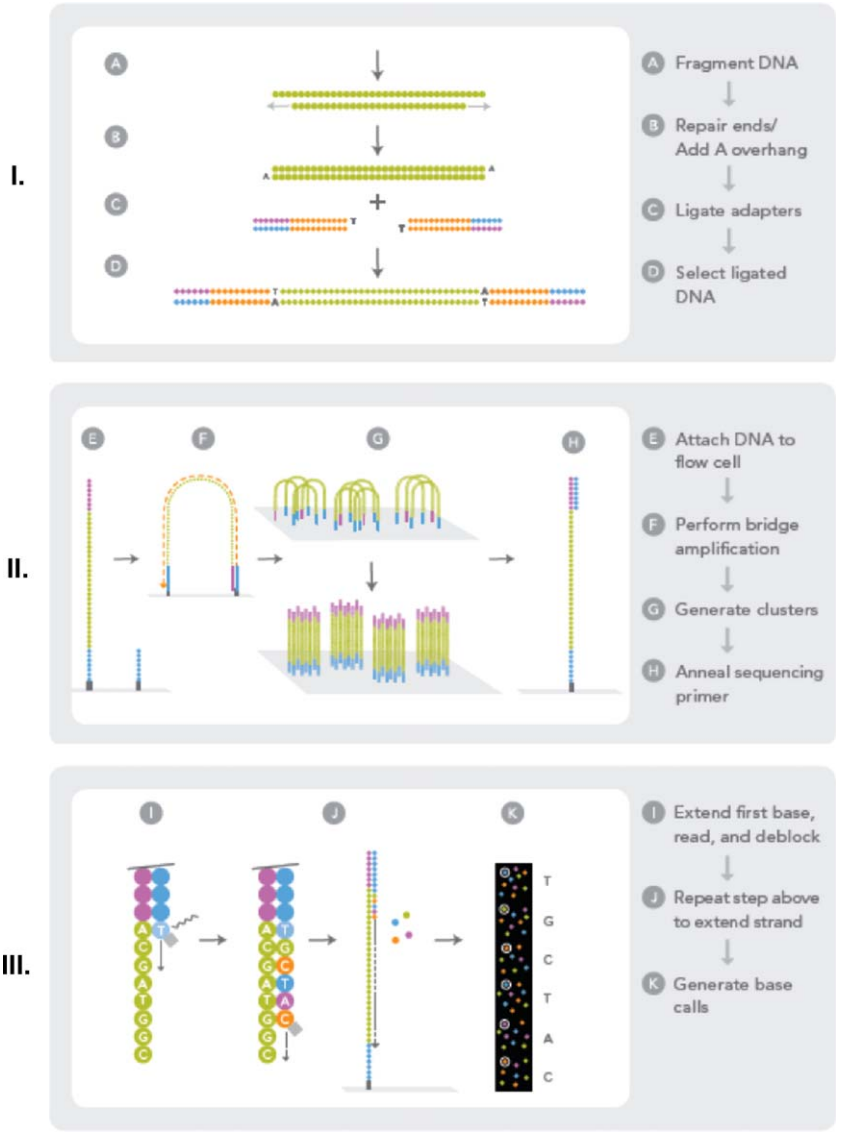

FIGURE 2

Outline of the Illumina Genome Analyzer workflow. Similar fragmentation and adapter ligation steps take place (I), before applying the library onto the solid surface of a flow cell. Attached DNA fragments form 'bridge' molecules which are subsequently amplified via an isothermal amplification process, leading to a cluster of identical fragments that are subsequently denatured for sequencing primer annealing (II). Amplified DNA fragments are subjected to sequencing-by-synthesis using 3′ blocked labelled nucleotides (III). (Adapted from the Genome Analyzer brochure, http://www.solexa.com.)

of fourth and fifth bases is characterised by one of four fluorescent labels at the end of the octamer. After the detection of the fluorescence from the label, bases 4 and 5 in the sequence are thus determined. The ligated octamer oligonucleotides are cleaved off after the fifth base, removing the fluorescent label, then hybridisation and ligation cycles are repeated, this time determining bases 9 and 10 in the sequence; in the subsequent cycle bases 14 and 15 are determined, and so on. The sequencing process may be continued in the same way with another primer, shorter by one base than the previous one, allowing one to determine, in the successive cycles, bases 3 and 4, 8 and 9, 13 and 14. The achieved sequence reading length is at present about 35 bases. Because each base is determined with a different fluorescent label, error rate is reduced. Sequences can be determined in parallel for more than 50 million bead clusters, resulting in a very high throughput of the order of Gigabases per run.

Applied Biosystems produced an updated version in 2008, the SOLiD 2.0 platform, which may increase the output of the instrument from 3 to 10 Gb per run. This change will reduce the overall run time of a fragment library on the new system to 4.5 days from 8.5 days on the existing machine. For further information see www3.appliedbiosystems.com/index.htm, and in [1]

The Helicos single-molecule sequencing device, HeliScope

The systems discussed above require the emulsion PCR amplification step of DNA fragments, to make the light signal strong enough for reliable base detection by the CCD cameras. PCR amplification has revolutionised DNA analysis, but in some instances it may introduce base sequence errors into the copied DNA strands, or favour certain sequences over others, thus changing the relative frequency and abundance of various DNA fragments that existed before amplification. Ultimate miniaturisation

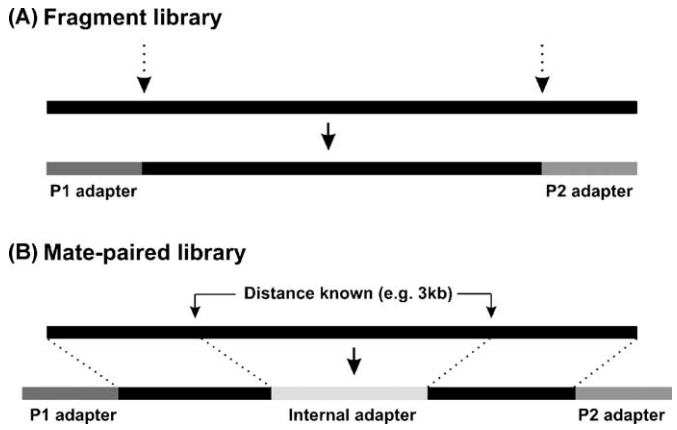

FIGURE 3

Library preparation for DNA sequencing using the SOLiD DNA sequencing platform. **(A)** *Fragment library*: After whole genome DNA is randomly fragmented (indicated by the dashed arrows), two different 25 bp DNA adapters (P1 and P2) are ligated at the 5′- and 3′-ends of the DNA fragments generated. **(B)** *Mate-paired library*: In this case, DNA fragments that are separated from another DNA fragment of known length (e.g. 3 kb for this example) are ligated such that they encompass an internal adapter. Subsequently, two different DNA adapters are ligated at the 5′- and 3′-ends, similarly to (A). (Adapted and modified from http://www.appliedbiosystems.com.)

into the nanoscale, and the minimal use of biochemicals, would be achieved if the sequence could be determined directly from a single DNA molecule, without the need for PCR amplification and its potential for distortion of abundance levels. This requires a very sensitive light detection system and a physical arrangement capable of detecting and identifying light from a single dye molecule. Techniques for the detection and analysis of single molecules have been under intensive development over past decades, and several very sensitive systems for single photon detection have been produced and tested. One of the first techniques for sequencing from a single DNA molecule was described by the team of S. Quake [12], and licensed by Helicos Biosciences.

Helicos introduced the first commercial single-molecule DNA sequencing system in 2007. The nucleic acid fragments are hybridised to primers covalently anchored in random positions on a glass cover slip in a flow cell. The primer, polymerase enzyme and labelled nucleotides are added to the glass support. The next base incorporated into the synthesised strand is determined by analysis of the emitted light signal, in the sequencing-by-synthesis technique (similar to Fig. 2, but on only one DNA fragment, without amplification). This system also analyses many millions of single DNA fragments simultaneously, resulting in sequence throughput in the Gigabase range. Although still in the first years of operation, the system has been tested and validated in several applications with promising results, for example in the pre-natal trisomy-21 (Down Syndrome) test, using only the maternal blood sample, potentially replacing the standard test which is associated with some risk to the foetus [13].

When the Helicos system was used to sequence the genome of M13 phage, read lengths averaged about 23 bases. There were still some limitations in the single-molecule technology, on the basis of the first generation of the chemistry. In the homopolar regions, multiple fluorophore incorporations could decrease emissions, sometimes below the level of detection; when errors did occur, most were deletions. Helicos announced that it has recently developed a new generation of 'one-base-at-a-time' nucleotides which allow more accurate homopolymer sequencing, and lower overall error rates. For further information, see http://www.helicosbio.com/

Novel DNA sequencing techniques in development

Developments of novel DNA sequencing techniques are taking place in many groups worldwide. In the laboratory of Church [14] a technique similar to the sequencing-by-synthesis method above has been developed, with multiplex polony technology. Several hundred sequencing templates are deposited onto thin agarose layers, and sequences are determined in parallel. This presents increase of several orders of magnitudes in the number of samples which can be analysed simultaneously. A further advantage is the large reduction of reaction volumes, the smaller amounts of reagents needed and the resulting lower cost. The group continues development of their platform and offers this technique to academic laboratories using off-the-shelf optics, hardware and reagents.

Another promising approach, attempting to use real-time single-molecule DNA sequence determination, is being developed by VisiGen Biotechnologies http://visigenbio.com/. They have produced a specially engineered DNA polymerase (acting as a 'real-time sensor' for modified nucleotides) with a donor fluorescent dye incorporated close to the active site involved in selection of the nucleotides during synthesis. All four nucleotides to be integrated have been modified, each with a different acceptor dye. During the synthesis, when the correct nucleotide is found, selected and enters the active site of the enzyme, the donor dye label in the polymerase comes into close proximity with the acceptor dye on the nucleotides and energy is transferred from donor to acceptor dye giving rise to a fluorescent resonant energy transfer (FRET) light signal. The frequency of this signal varies depending on the label incorporated in the nucleotides, so that by recording frequencies of emitted FRET signals it will be possible to determine base sequences, at the speed at which the polymerase can integrate the nucleotides during the synthesis process (usually a few hundred per second). The acceptor fluorophore is removed during nucleotide incorporation, which ensures that there are no DNA modifications that might slow down the polymerase during synthesis. VisiGen plans to offer a service on the basis of its real-time single-molecule nanosequencing technology by end 2009, followed by the launch of equipment and reagents 18 months later. The technology could eventually enable researchers to sequence an entire human genome in less than a day for under $1000. The company is currently working on its first version of the instrument, which can generate around 4 Gb of data per day. The single-molecule approach requires no cloning and no amplification, which eliminates a large part of the cost relative to current technologies. In addition, read lengths for the instrument are expected to be around 1 kb, longer than any current platform.

Another US company, Pacific Biosciences (http://www.pacificbiosciences.com/index.php), announced recently that it is working on a next-generation DNA sequencing instrument that will eventually be able to produce 100 Gb of sequence data per hour, or a diploid human genome at onefold coverage in about 4 min. They plan to sell their first systems during 2010. The company's single-molecule real-time (SMRT) technology is based on zero mode waveguides (ZMWs) which were originally developed at Cornell

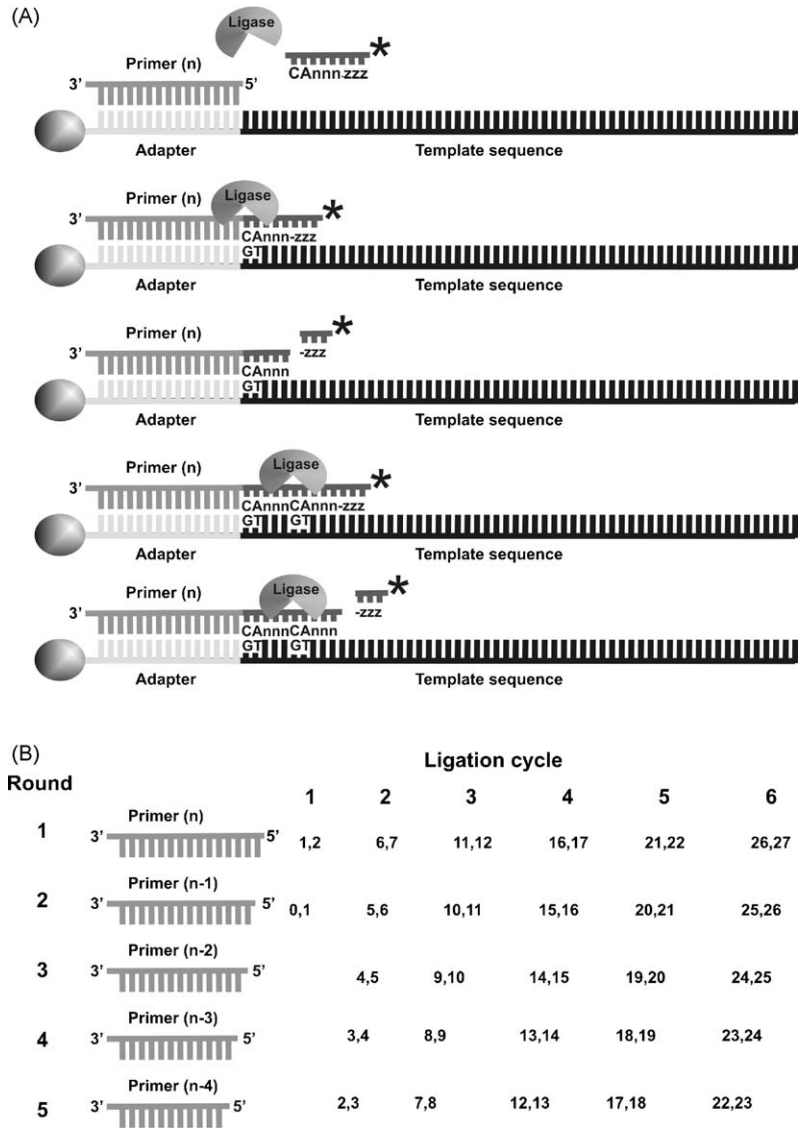

FIGURE 4

Sequencing-by-ligation, using the SOLiD DNA sequencing platform. **(A)** Primers hybridise to the P1 adapter within the library template. A set of four fluorescence-labelled di-base probes competes for ligation to the sequencing primer. These probes have partly degenerated DNA sequence (indicated by *n* and *z*) and for simplicity only one probe is shown (labelling is denoted by asterisk). Specificity of the di-base probe is achieved by interrogating the first and second base in each ligation reaction (CA in this case for the complementary strand). Following ligation, the fluorescent label is enzymatically removed together with the three last bases of the octamer. **(B)** Sequence determination by the SOLiD DNA sequencing platform is performed in multiple ligation cycles, using different primers, each one shorter from the previous one by a single base. The number of ligation cycles (six for this example) determines the eventual read length, whilst for each sequence tag, six rounds of primer reset occur [from primer (*n*) to primer ($n - 4$)]. The dinucleotide positions on the template sequence that are interrogated each time, are depicted underneath each ligation cycle and are separated by 5-bp from the dinucleotide position interrogated in the subsequent ligation cycle. (Adapted and modified from http://www.appliedbiosystems.com.)

University Nanobiotechnology Center. ZMWs are nanometre-scale aperture chambers in a 100 nm metal film deposited on a clear substrate. Owing to the behaviour of light aimed at such a small chamber, the observation volume is only 20 zeptolitres, enabling researchers to measure the fluorescence of nucleotides incorporated by a single DNA polymerase enzyme into a growing DNA strand in real time. The developers have so far observed read lengths of about 1500 bases and a rate of 10 bases/s, and have been able to analyse up to 3000 ZMWs in parallel.

Another single-molecule sequencing technique may develop from studies on translocation of DNA through various artificial nanopores. The work in this field was pioneered at Harvard by D. Branton, G. Church and J. Golovchenko, at UC Santa Cruz by D. Deamer and M. Akeson and at the NI Standards and Technology by J. Kasianowicz. The approach is based on the modulation of the ionic current through the pore as a DNA molecule traverses it, revealing characteristics and parameters (diameter, length and conformation) of the molecule. Recent study in this direction is the work of Trepanier and colleagues [15], which contains references to previous studies on this subject. In their work [15] they analyze several limitations of the method. One limitation to single-base resolution in nanopore-based DNA sequencing approaches is the insufficient control of the translocation speed of the molecule, for example during electrophoresis of the DNA molecules through the nanopore. This was overcome by the integration of an optical trapping system, which enables control

and lowering of the translocation speed by several hundred-fold. For the demonstration, a known DNA fragment was used, attached via streptavidin–biotin to a polystyrene bead with a diameter of 10 μm. The bead was placed into the optical trap, and translocation speed was reduced about 200-fold, giving more time for analysis of the DNA molecule passing through the nanopore. It was also possible to control the motion of the molecule and return it back to its starting point before the translocation, thus making possible repeated measurements and analysis.

Another of these approaches, studied by several teams collaborating as part of an EU consortium on nano-DNA-sequencing coordinated by R. Zikic from Belgrade, L. Forro and A. Radenovic (EPFL Lausanne), is the development of a nano-electronic device for high-throughput single-molecule DNA sequencing, with the potential to determine long genomic sequences. This is on the basis of the electrical characterisation of individual nucleotides, whilst DNA passes through a nanopore (similar to [15]), with integrated nanotube side-electrodes developed at EPFL, Lausanne. A lithographically fabricated nanogap is produced with single-nanometre precision and allows characterisation of the tunnelling conductance across DNA bases and the electrical response of DNA molecule translocation between two carbon nanotube electrodes. The translocation rate of DNA through the nanopore will be varied by an optical tweezers system (in addition to standard techniques of applied voltage, viscosity change and DNA charge at various pH), aiming to achieve single-base resolution. Further improvements and modifications of the technique, increasing the number of parameters measured during the translocation of the DNA enabling single-base resolution, could lead to a rapid nanopore-based DNA sequencing technique.

Sequenom (http://www.sequenom.com) has licensed technology from Harvard University, to develop a nanopore-based sequencing platform that will be faster and cheaper than currently available technologies. In the near term they plan to use it for large-scale genotyping applications, RNA and epigenetic analyses. In the long term it has the potential to provide a commercially viable, rapid, sub-1000 dollar human genome sequencing solution. The technology has also been licensed by Oxford Nanopore Technologies, UK (http://www.nanoporetech.com/).

BioNanomatrix and Complete Genomics (http://bionanomatrix.com/) announced in 2007 the formation of a joint venture to develop technology to sequence a human genome in eight hours for less than $100. The proposed platform will use Complete Genomics's sequencing chemistry and BioNanomatrix's nanofluidic technology. They plan to adapt DNA sequencing chemistry with linearised nanoscale DNA imaging to create a system that can read DNA sequences greater than 100,000 bases. With their design and price they target the possible sequencing of many genomes. Complete Genomics company (http://www.completegenomics.com) presented recently a new method, using rolling circle PCR amplification resulting in DNA nanoballs, and a modified ligation technique, for fast and non expensive sequencing of human genomes.

A very different approach to single-molecule DNA sequencing, using RNA polymerase (RNAP), has been presented recently [16]. In the planned method, RNAP is attached to one polystyrene bead, whilst the distal end of a DNA fragment is attached to another bead. Each bead is placed in an optical trap and the pair of optical traps levitates the beads. The RNAP interacts with the DNA fragment and the transcriptional motion of RNAP along the template changes the length of the DNA between the two beads. This leads to displacement of the two beads that can be registered with precision in the Angstrom range, resulting in single-base resolution on a single DNA molecule. By aligning four displacement records, each with a lower concentration of one of the four nucleotides, in a role analogous to the primers used in Sanger sequencing, and using for calibration the known sequences flanking the unknown fragment to be sequenced, it is possible to deduce the sequence information. Thirty out of 32 bases were correctly identified in about 2 min. The technique demonstrates that the movement of a nucleic acid enzyme, and the very sensitive optical trap method, may allow extraction of sequence information directly from a single DNA molecule.

Applications of high-throughput DNA sequencing

Novel fields and applications in biology and medicine are becoming a reality, beyond genomic sequencing as the original development goal and application. Examples include personal genomics with detailed analysis of individual genomic stretches; precise analysis of RNA transcripts for gene expression, surpassing and replacing in several aspects analysis carried out by various microarray platforms, for example in reliable and precise transcript quantification; and as a tool for identification and analysis of DNA regions that interact with regulatory proteins in functional regulation of gene expression. Next-generation sequencing technologies offer novel, rapid ways for genome-wide characterisation and profiling of mRNAs, small RNAs, transcription factor regions, chromatin structure and DNA methylation patterns, in microbiology and metagenomics.

Personal genomics, project human diversity in 1000 genomes

The cost of genome sequencing, an important factor in future studies, is becoming low enough to make personal genomics a close reality. Reduction of cost by two orders of magnitude is needed to be able to realise the potential of personal genomics, for which the goal of $1000 for a human genome sequence has been set. The impressive results obtained so far in various projects with the new technology are very convincing and will lead to lower cost. The analysis of the first two available human genomes [17,18] has demonstrated, how difficult it still is to draw medically or biologically relevant conclusions from individual sequences. More genomes need to be sequenced, to learn how genotype correlates with phenotype. A plan for a project to sequence 1000 human genomes has been prepared, which will allow creation of a reference standard for the analysis of human genomic variations that is expected to contribute to studies of disease (http://www.1000genomes.org/). Illumina, Roche 454 Life Sciences and Applied Biosystems will take part in the project and generate the equivalent of 25 human genomes each per year over a period of three years. This significant sequence contribution will enable the team to analyse the human genome with deeper sequencing and shorten its completion time. The 1000 Genomes Project will identify variants present at a frequency of 1% over most of the genome, and as low as 0.5% within genes.

The immediate applications and relevance of next-generation sequencing techniques in the medical field have been demonstrated already, by the ability to detect cancer alleles with deep

sequencing of genomic DNA in cancerous tissues (carefully isolated by laser microdissection and capture techniques), which would have presented a very tedious task for the Sanger technique.

RNA sequencing, analysis of gene expression

The high throughput of next-generation sequencing technology, rapidly producing huge numbers of short sequencing reads, made possible the analysis of a complex sample containing a mixture of a large number of nucleic acids, by sequencing simultaneously the entire sample content. This is now possible without the tedious and time-consuming bacterial cloning, avoiding associated disadvantages. It may also be applied to the characterisation of mRNAs, methylated DNA, DNA or RNA regions bound by certain proteins and other DNA or RNA regions involved in gene expression and regulation. The original SAGE technique [19] demonstrated novelty and powerful analysis, but was limited in applications because of the need for difficult ligation of a huge number of short DNA transcripts, subsequent cloning and Sanger sequencing. Using next-generation technology, the concept of the SAGE method now allows the analysis of RNA transcripts in a biological sample by obtaining short sequence tags, 20–35 bases long, directly from each transcript in the sample. With this technique, transcripts are characterised through their sequence [20], in contrast to the probe hybridisation employed in DNA chip techniques, with their inherent difficulties of cross-hybridisation and quantitation. Owing to the huge number of samples analysed simultaneously, sequence-based techniques can detect low abundance RNAs, small RNAs, or the presence of rare cells contained in the sample. Another advantage of this approach is that it does not require prior knowledge of the genome sequence. The technique has been applied recently to transcriptome profiling in stem cells [21] and to RNA-Seq study into alternative splicing in human cells [22].

Chromatin immunoprecipitation, ChIP-Seq technique

Next-generation sequencing technology allowed replacement of microarrays in the mapping step with high-throughput sequencing of DNA binding sites, and their direct mapping to a reference genome in the database [23]. The sequence of the binding site is mapped with high resolution to regions shorter than 40 bases, a resolution not achievable by microarray mapping. Moreover, the ChiP-Seq technique is not biased and allows the identification of unknown protein binding sites, which is not the case with the ChiP-on-chip approach, where the sequence of the DNA fragments on the microarray is pre-determined, e.g. in promoter arrays, exon arrays, etc.

Prospects for future DNA sequencing technology and applications

The availability of ultra-deep sequencing of genomic DNA will transform the biological and medical fields in the near future, in analysis of the causes of disease, development of new drugs and diagnostics. It may become a promising tool in the analysis of mental and developmental disorders such as schizophrenia and autism [24–26]. It is anticipated that DNA sequencing of whole genomes for clinical purposes using these new technologies will probably occur in the next couple of decades. Some of the most recent applications can be found in the proceedings of the AGBT conference (Advances in Genome Biology and Technology), February 2009.

The novel sequencing technologies will be also useful in microbial genomics, for example in the metagenomics measuring the genetic diversity encoded by microbial life in organisms inhabiting a common environment [27]. Many microbial sequencing projects have been already completed or are being prepared and several comparative genome analyses are under way to link genotype and phenotype at the genomic level. The proposed Human Microbiome Project (also called The Second Human Genome Project), analysing the collection of microbes in and on the human body, will contribute to understanding human health and disease [28]. Changes in microbial communities in the body have been generally linked to immune system function, obesity and cancer. In future, each individual's microbiome could eventually become a medical biometric.

An important application is planned by the US DOE Joint Genome Institute, JGI (http://www.jgi.doe.gov/), which will focus its sequencing efforts on new plant and microbial targets that may be of use in the development of alternative energies. The JGI plans to sequence the genome of the marine red alga, which may play an important environmental role in removing carbon dioxide from the atmosphere.

Genomics, proteomics and medical research all benefit from recent advances and novel techniques for high-throughput analysis (e.g. DNA and protein microarrays, quantitative PCR, mass spectrometry, novel DNA sequencing techniques and others). Devices with short DNA sequence reads (25–50 bases) have already found many applications, but for genomic sequencing, and for analysis of the ever more important structural genetic variations in genomes, such as copy number variations, chromosomal translocations, inversions, large deletions, insertions and duplications, it would be a great advantage if sequence read length on the original single DNA molecule could be increased to several 1000 bases and more per second. Ideally, the goal would be the sequence determination of a whole chromosome from a single original DNA molecule. Hopes for future in this direction may provide novel developments in several physical techniques (e.g. various advanced AFM methods, electron microscopy, soft X-rays, various spectroscopic techniques, nanopores and nano-edges), with many improvements needed and under intense development.

Acknowledgement

I am grateful to Dr George Patrinos for discussions and help with preparation of the figures.

References

1 Schuster, S.C. *et al.* (2008) Method of the year, next-generation DNA sequencing. Functional genomics and medical applications. *Nat. Methods* 5, 11–21

2 Sanger, F. *et al.* (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–5467

3 Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* 74, 560–564

4 Smith, L.M. *et al.* (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* 321, 674–679

5 Ansorge, W. *et al.* (1986) A non-radioactive automated method for DNA sequence determination. *J. Biochem. Biophys. Methods* 13, 315–323
6 Ansorge, W. *et al.* (1987) Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Res.* 15, 4593–4602
7 Edwards, A. *et al.* (1990) Automated DNA sequencing of the human HPRT locus. *Genomics* 6, 593–608
8 Ansorge, W., EMBL Heidelberg (1991), Process for sequencing nucleic acids without gel sieving media on solid support and DNA chips (Verfahren zur Sequenzierung von Nukleinsauren ohne Gele). German Patent Application DE 41 41 178 A1 and Corresponding Worldwide Patent Applications.
9 Nyren, P. and Lundin, A. (1985) Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal. Biochem.* 151, 504–509
10 Hyman, E.D. (1988) A new method of sequencing DNA. *Anal. Biochem.* 174, 423–436
11 Ronaghi, M. *et al.* (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* 242, 84–89
12 Braslavsky, I. *et al.* (2003) Sequence information can be obtained from single DNA molecule. *Proc. Natl. Acad. Sci. U. S. A.* 100, 3960–3964
13 Quake, S.R. *et al.* (2008) Pre-natal trisomia-21-test from maternal blood sample. *Proc. Natl. Acad. Sci. U. S. A.* 105, 16266–16271
14 Shendure, J. *et al.* (2005) *Science* 309, 1728–1732
15 Trepagnier, E.H. *et al.* (2007) Controlling DNA capture and propagation through artificial nanopores. *Nano Lett.* 7, 2824–2830
16 Greenleaf, W.J. and Block, S.M. (2006) Single-molecule, motion-based DNA sequencing using RNA polymerase. *Science* 313, 801
17 Wheeler, D.A. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872–876
18 Levy, S. *et al.* (2007) The diploid genome sequence of an individual human. *PloS* 5, e254
19 Velculescu, V.E. *et al.* (1995) Serial analysis of gene expression. *Science* 270, 484–487
20 Mortazavi, A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628
21 Cloonan, N. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5, 613–619
22 Sultan, M. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321, 956–960
23 Robertson, G. *et al.* (2007) ChIP-Seq techniques. *Nat. Methods* 4, 651–657
24 Morrow, E.M. *et al.* (2008) Identifying autism loci and genes by tracing recent shared ancestry. *Science* 321, 218–223
25 Geschwind, D.H. (2008) Autism – family connections. *Nature* 454, 838–839
26 Sutcliffe, J.S. (2008) Insights into the pathogenesis of autism. *Science* 321, 208–209
27 Hugenholtz, P. and Tyson, G.W. (2008) Metagenomics. *Nature* 455, 481–483
28 Turnbaugh, P.J. *et al.* (2007) The human microbiome project. *Nature* 449, 804–810