Nature Protocols

Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources

Da Wei Huang[1], Brad T. Sherman[1], Richard A. Lempicki[*]

Laboratory of Immunopathogenesis and Bioinformatics, Clinical Services Program, SAIC-Frederick, Inc., National Cancer Institute at Frederick, Frederick, MD 21702.

[1] These authors contributed equally to this study.

* Correspondence: Dr. Richard A Lempicki.

E-mail: rlempicki@mail.nih.gov
Ph. (301) 846-5093
Lab Web Site: http://david.niaid.nih.gov or http://david.abcc.ncifcrf.gov

Lab Fax: (301) 846-6762

E-mail Addresses for Authors:

Da Wei Huang huangdawei@mail.nih.gov
Brad T. Sherman bsherman@mail.nih.gov
Richard A Lempicki rlempicki@niaid.nih.gov

ABSTRACT

DAVID Bioinformatics Resources (DAVID) at http://david.abcc.ncifcrf.gov consists of an integrated biological knowledgebase and analytic tools aiming at systematically extracting biological meaning from large gene/protein lists. This protocol explains how to use DAVID, a high throughput and integrated data mining environment, to analyze gene lists derived from high throughput genomic experiments. The procedure first requires uploading a gene list containing any number of common gene identifiers followed by analysis using one or more text and pathway mining tools such as Gene Functional Classification, Functional Annotation Chart or Clustering, and Functional Annotation Table. By following this protocol, investigators are able to gain an in-depth understanding of the biological themes in lists of genes that are enriched in genome-scale studies.

SEARCH TERMS

Computational Biology; Bioinformatics; Genomics; Microarray data analysis; Bio-knowledge databases; Gene functional annotation; High throughput gene functional analysis**;** Gene functional classification.

INTRODUCTION

High-throughput genomic, proteomic and bioinformatics scanning approaches, such as, expression microarray, promoter microarray, proteomic data, and ChIP-on-CHIPs, provide significant capabilities to study a large variety of biological mechanisms including associations with diseases. These technologies usually result in a large 'interesting' gene list (ranging in size from hundreds to thousands of genes) involved in studied biological conditions. Data analysis of the large gene lists is a very important downstream task following the above example high throughput technologies in order to understand the biological meaning of the output gene lists. The data analysis of such high complex and large volume datasets is a challenging task, which requires support from special bioinformatics software packages. In this protocol, we introduce DAVID (the Database for Annotation, Visualization and Integrated Discovery) Bioinformatics Resources[1,2], which is able to extract biological features/meaning associated with large gene lists. DAVID is able to handle any type of gene list, no matter which genomic platform or software package generated them.

DAVID, released in 2003[2,3], as well as a number of other similar publicly available tools, including, but not limited to, GoMiner[4], GOstat[5], Onto-express[6], GoToolBox[7], FatiGO[8], GFINDer[9], GOBar[10], and GSEA[11], (See Supplementary Data 1 for a complete list), address various aspects of the challenge of functionally analyzing large gene lists. Although each tool has distinct features and strengths, as reviewed by Khatri et al.[12], they all adopt a common core strategy to systematically map a large number of interesting genes in a list to the associated biological annotation (e.g. Gene Ontology Terms), and then statistically highlight the most over-represented (enriched)

biological annotation out of thousands of linked terms and contents. Enrichment analysis is a promising strategy that increases the likelihood for investigators to identify biological processes most pertinent to the biological phenomena under study.

The analysis of large gene lists is indeed more of an exploratory, computational procedure rather than a purely statistical solution. As compared to other similar services, DAVID provides some unique features and capabilities, such as, an integrated and expanded backend annotation database[13], advanced modular enrichment algorithms[14], and powerful exploratory ability in an integrated data mining environment[1]**.** Even though users can learn more in-depth information about DAVID algorithms in our original publications[1-3,13-15], we now briefly summarize the rationale regarding the key DAVID modules, as well as the analytic limitations (also see Table 1 for across comparisons), so that readers may be able to quickly follow the protocol.

Large Gene Lists Ready for Functional Analysis by DAVID

In this protocol, we use a previously published gene list[16] (Supplementary Data 2) as an example to illustrate the results obtained from the various DAVID analytic modules. To obtain this list, freshly isolated peripheral blood mononuclear cells (PBMCs) were treated with an HIV envelope protein (gp120) and genome-wide gene expression changes were observed using Affymetrix U95A microarray chips[16]. The aim of the experiment was to investigate cellular responses to viral envelope protein infection, which may help in understanding the mechanisms for HIV replication in resting or sub-optimally activated PBMCs.

The quality of large gene lists derived from high-throughput biological studies is one of the most important foundations that directly influences the success of the following functional analysis in DAVID. Due to the complexity of the data mining situations involved in biological studies, there is no good systematic way, at the present time, to quantitatively estimate the quality of the gene list ahead of time (i.e. before the gene functional analysis). However, based on real-life data analysis experiences during the past several years, a 'good' gene list may exhibit most, if not all of following characteristics:

1) Contain many important genes (marker genes) as expected for given study (e.g. IL8, CCL4, and TNFSF8 from the example gene list)

2) Reasonable number of genes ranging from hundreds to thousands (e.g. 100 to 2000 genes), not extremely low or high.

3) Most of the genes significantly pass the statistical threshold (e.g. selecting genes by comparing gene expression between control and experimental cells with t-test statistics: fold changes >=2 and P-values <=0.05) for selection. Importantly, statistical thresholds do not have to be sacrificed (e.g. fold changes >=1.1 and p-value<=0.2) in order to reach a comfortable gene size.

4) Notable portion of up/down-regulated genes are involved in certain interesting biological processes, rather than randomly spread throughout all possible biological processes.

5) A 'good' gene list should consistently contain more enriched biology than that of a random list in the same size range during analysis in DAVID (Supplementary Data 3 for detailed discussions).

6) High reproducibility (e.g. by independent experiments under the same conditions or by leave-one-out statistical test) to generate a similar gene list under the same conditions.
7) The high quality of the high-throughput data can be confirmed by other independent wet lab tests/experiments.
Some of the estimating points (2, 3, 6, & 7) come from upstream analysis while DAVID may help in examining others (1, 4 & 5).

Moreover, for enrichment analysis, in general, a larger gene list can have higher statistical power resulting in a higher sensitivity (more significant p-values) to slightly enriched terms, as well as to more specific terms. Otherwise, the sensitivity is decreased toward largely enriched terms and broader/general terms. Although the size of the gene list influences (in a non-linear way) the absolute enrichment p-values, which makes it difficult to directly compare the absolute enrichment p-values across gene lists, the enrichment p-values are fairly comparable within the same or same size of gene list. In addition, when different sizes of gene lists are generated from the same dataset with different threshold stringencies (within a reasonable range), the absolute enrichment p-values may vary from list to list. However, the relative rank/order of the enriched terms may remain fairly stable, which will lead to consistent global conclusions of functional annotations across the different sizes of gene lists derived from the same dataset (data not shown). This kind of reproducibility and consistency should be expected using DAVID tools if the underlying high-throughput biological studies are robust.

Interestingly, we found that many gene lists input to DAVID are in the size range of 1 to 10 genes. The enrichment statistic's power will be very limited in such extreme cases. However, the unique exploratory capability of DAVID could still be very powerful

for analyzing such small gene lists. Since the analysis is most likely in a very focused and small scope, analysts may take advantage of the unique exploratory capability of DAVID to navigate through all of the well organized heterogeneous annotation contents around the focused genes regardless of the statistics.

Submission of User's Gene Identifiers to DAVID

Comprehensively mapping a user's gene identifiers (gene IDs) to the relevant biological annotation in the DAVID database is an essential foundation for the success of any high-throughput gene functional analysis. Gene IDs and biological annotations are highly redundant within the vast array of public databases. The DAVID Knowledgebase[13], was designed to collect and integrate diverse gene identifiers as well as more than 40 well-known publicly available annotation categories (Supplementary Data 4), which are then centralized by internal DAVID identifiers in a non-redundant manner. The wide range of biological annotation coverage and the non-redundant integration of gene IDs in the DAVID Knowledgebase enables a user's gene ID to be mapped across the entire database, thus providing comprehensive coverage of gene-associated annotation. If a significant portion (>=20%) of input gene IDs fail to be mapped to an internal DAVID ID, a specially designed module, the DAVID Gene ID Conversion Tool[15], will start up in order to help map such IDs.

Principle of 'Gene Population Background' in enrichment analysis

The principle foundation of enrichment analysis is that if a biological process is abnormal in a given study, the co-functioning genes should have a higher potential (enriched) to be selected as a relevant group by the high-throughput screening technologies. To decide the degree of enrichment, a certain background must be set up in order to perform the comparison (also see Step 1 in Table 2). For example, 10% of the user's genes are kinases vs. 1% of the genes in the human genome (this is the gene population background) that are kinases. The enrichment can therefore be quantitatively measured by some common and well-known statistical methods including Chi-square, Fisher's exact test, Binomial probability, and Hypergeometric distribution. Thus, a conclusion may be obtained for the particular example, that is, kinases are enriched in the user's study, and therefore play important roles in the study. However, 10% alone cannot make such a conclusion without comparing it to the background information (i.e. 1%).

In this sense, the background is one of the critical factors that impact the conclusion to a certain degree, particularly when two ratios are close. There are many ways to set the backgrounds, e.g. all genome genes; genes on an Affymetrix chip; and a sub-set of genome genes that the user used in their study. In general, larger backgrounds, e.g. the total genes in the genome as a population background, intends to give more significant p-values, as compared to a narrowed-down set of genes as a population background, such as genes only existing on a microarray. Even though there is no gold standard for the population background, a general guideline is to set up the population background as the pool of genes which have a chance to be selected for the studied annotation category in the scope of the users' particular study.

One of the advantages of DAVID is its flexibility of setting different population backgrounds to meet different situations. DAVID has an automatic procedure to ‘guess’ the background as the global set of genes in the genome based on user's uploaded gene list. Thus, in a regular situation, users do not have to set up a population background by themselves. We found that it works generally well simply because most of the studies analyzed by DAVID are genome-wide or close to genome-wide studies. Moreover, other options are also available for user's choices including all genes in the studied genome, genes in various microarray chips, and most importantly any gene set that users define and upload. The latter feature requires significant computational power so that it is rarely found in similar web-based applications. In summary, various settings and options for population backgrounds can meet the range of needs of general users to those of power users.

DAVID Gene Name Batch Viewer

Gene IDs, such as Entrez Gene 3558, typically do not convey biological meaning in and of itself. The Gene Name Batch Viewer[1] is able to quickly attach meaning to a list of gene IDs by rapidly translating them into their corresponding gene names (Figure 3, Slide 4 of Supplementary Data 5 for more detail). Thus, before proceeding to analysis with other more comprehensive analytic tools, investigators can quickly glance at the gene names to further gain insight about their study and to answer questions such as, "Does my gene list contain important genes relevant to the study?". In addition, a set of

hyperlinks are provided for each gene entry, allowing users to further explore additional functional information about each gene.

DAVID Gene Functional Classification

As the analysis proceeds, Gene Functional Classification[14] provides the distinct ability for investigators to explore and view functionally related genes together, as a unit, in order to concentrate on the larger biological network rather than at the level of an individual gene. In fact, the majority of co-functioning genes may have diversified names so that genes cannot be simply classified into functional groups according to their names. However, Gene Functional Classification, accomplished with a set of novel fuzzy clustering techniques, is able to classify input genes into functionally related gene groups (or classes) based on their annotation term co-occurrence rather than on gene names. Condensing large gene lists into biologically meaningful modules greatly improves one's ability to assimilate large amounts of information and thus switches functional annotation analysis from a gene-centric analysis to a biological module-centric analysis (Figure 4, Slide 5 and 6 of Supplementary Data 5 for more detail). Taken together with the 'drill-down' function associated with each biological module and visualizations to view the relationships between the many-genes-to-many-terms associations, investigators are able to more comprehensively understand how genes are associated with each other and with the functional annotation.

DAVID Functional Annotation Chart

Functional Annotation Chart[1-3] provides typical gene-term enrichment (over-represented) analysis, that is also provided by other similar tools, to identify the most relevant (over-represented) biological terms associated with a given gene list (Figure 5, Slide 8 of Supplementary Data 5 for more detail). Compared to other similar enrichment analysis tools, the notable difference of this function provided by DAVID is its extended annotation coverage[13], increasing from only GO in the original version of DAVID to currently over 40 annotation categories, including GO terms, protein–protein interactions, protein functional domains, disease associations, bio-pathways, sequence features, homology, gene functional summaries, gene tissue expression, and literature. (Supplementary Data 4). The annotation categories can be flexibly included or excluded from the analysis based upon a user's choices (Slide 7 of Supplementary Data 5 for more detail). The enhanced annotation coverage alone increases the analytic power by allowing investigators to analyze their genes from many different biological aspects in a single space. In addition, to take full advantage of the well-known KEGG and BioCarta pathways, DAVID Pathway Viewer, which is accessed by clicking on pathway links within the chart report, can display genes from a user's list on pathway maps to facilitate biological interpretation in a network context (Figure 4). Finally, the choice of pre-built or user-defined gene population backgrounds provides the user with the ability to tailor the enrichment analysis to meet the user's specific analytic situation.

DAVID Functional Annotation Clustering

Functional Annotation Clustering[14] uses a similar fuzzy clustering concept as Functional Classification by measuring relationships among the annotation terms based on the degree of their co-association with genes within the user's list in order to cluster somewhat heterogeneous, yet highly similar annotation into functional annotation groups (Figure 6, Slide 10 of Supplementary Data 5 for more detail). This reduces the burden of associating different terms associated with the similar biological process, thus allowing the biological interpretation to be more focused at the "biological module" level. The 2-D view tool is also provided for examining the internal relationships among the clustered terms and genes (Slide 6 of Supplementary Data 5). This type of grouping of functional annotation is able to give a more insightful view of the relationships between annotation categories and terms compared to the traditional linear list of enriched terms since highly related/redundant annotation terms may be dispersed among hundreds, if not thousands, of other terms.

DAVID Functional Annotation Table

Functional Annotation Table[1,2] is a query engine for the DAVID Knowledgebase, without statistical calculations (Figure 7, Slide 11 of Supplementary Data 5). For a given gene list, the tool can quickly query corresponding annotation for each gene and present them in a table format. Thus, users are able to explore annotation in a gene by gene manner. This is a useful analytic module particularly when users want to closely look at the annotation of highly interesting genes.

Purpose

This paper will mainly describe the protocol of how to use each DAVID analytic module in a logical, sequential order, as well as how to switch among the analytic modules (Figure 1). The example gene list used in this protocol (also available as demo list 2 on DAVID web site) allows new users to quickly test and experience various functions provided by DAVID. The protocol provides a routine analytic flow for new users to begin, as well as the flexibility for experienced users to use the modules in different combinations in order to balance the different focuses and strengths of each module to better meet specific analytical questions (Figure 1). Moreover, table 2 lists major statistical methods and filtering parameters that may influence the DAVID analysis and result interpretation in certain ways, for users to quickly look up specific statistical topics according to their interests.

MATERIALS

EQUIPMENT

A computer with high speed internet access and a web browser.

EQUIPMENT SETUP

Hardware requirements and computer configurations

DAVID is a web-based tool designed so that a computer with a standard web browser using default settings should work well. There is no need for special configuration and installation. Although DAVID was tested with several combinations of internet browsers and operating systems, MS Internet Explorer or Firefox in a Window XP operating system is recommended in order to obtain the most satisfactory usability.

Input Data

A list of gene identifiers is the only required input for all DAVID analytic modules/tools. The gene list may be derived from any type of high-throughput genomic, computational or proteomic study, such as DNA expression microarray, proteomics, CHIP-on-CHIP, SNP array, CHIP-sequence, etc. The format of the gene list to be uploaded is described throughout the web site, and is either one gene ID per line or a list of comma delimited gene IDs in one line (Supplementary Data 6). DAVID supports most common public gene identifiers[13] (see Supplementary Data 4). In addition, after the gene list is submitted to DAVID, all DAVID analytic modules can access the current list from the gene list manager so that there is no need to re-submit the gene list for each DAVID tool.

An example gene list derived from an HIV microarray study[16] is used in this protocol, as well as available as demo_list2 on DAVID web site. The HIV microarray study is briefly

described in the introduction section. More detail can be found in the original publication[16].

Result Download

All results derived from DAVID may be explored and visualized on the web browser. Moreover, all results generated by DAVID can be downloaded in simple flat text formats, thereafter to be edited or plotted by other graphic tools, e.g. MS Excel, for publication purposes, as well as for archive purposes.

TIME TAKEN

The total analysis time varies, ranging from several minutes to hours, and is dependent on the analytical questions being addressed, the number of genes in the list being analyzed and the familiarity with the tools. It is not uncommon to make several visits to focus on different questions regarding a gene list of interest. Indeed computational time is only a small portion of the total time whereas exploring, interpreting and re-exploring both within DAVID and external to DAIVD tends to dominate most of the time. We used a PC computer with the Windows XP operating system, 2 G memory, 2.0 GHz CPU and 1Mbps internet connection for the data analysis of a gene list consisting of ~400 Affymetrix IDs (Supplementary Data 2) derived from an HIV study[16](presented in the Anticipated Results section). During the analysis course, for regular functional calls, each result was typically returned in ~10 seconds. For the most computationally intensive functions, such as Gene Functional Classification, results were typically returned within ~30 seconds, otherwise, never longer than 1 minute.

PROCEDURE

Submission of User's Gene IDs to DAVID

1| Submit a gene list to DAVID (Figure 2 & Slide 2 of Supplementary Data 5). Go to http://david.abcc.ncifcrf.gov or http://david.niaid.nih.gov and click on "Start Analysis" on the header. To do this, use the gene list manager panel that appears on the left side of the page (Figure 2) and perform the following steps: **(i)** Copy and paste a list of gene IDs into the box A or load a text file containing gene IDs to box B. See more details regarding format requirements in the Materials Section and also see Supplementary Data 6. **(ii)** Select the appropriate gene identifier type for your input gene IDs. See more details for supported ID types in the Materials Section**. (iii**) Indicate the list to be submitted as a gene list (i.e. genes to be analyzed) or as background genes (i.e. gene population background). **(iv)** Click the "Submit List" button.

CAUTION It takes ~30 seconds for a typical submission of ~1000 gene IDs; the progress bar, below the header, will disappear after a successful submission and a gene list name should appear in the list manager box; If >=20% input gene IDs cannot be recognized, the submission will be redirected to the DAVID Gene ID Conversion Tool[15] for further diagnosis. By default, the background is automatically set up as the genome-wide set of genes for the species that is found to have the majority of genes in the user's input list. However, it is always a good practice to double check the default, or select a more appropriate pre-built background through the "background" tab on top of the list manager.

2| Access DAVID analytic modules (Figure 2 & Slide 3 of Supplementary Data 5) via the tool menu page. The tool main menu is the central page which lists a set of hyperlinks leading to all available analytic modules. Clicking on each link will lead to the corresponding analytic module for analysis of your current gene list, highlighted in the gene list manager.

CRITICAL SETP: By clicking on "Start Analysis" on the header menu, users can always go back to this page at any time, no matter where they are, for choosing or switching to other analysis modules for current gene list.

Gene Name Batch Viewer

3| Run "Gene Name Batch Viewer" and explore results (Figure 3 & Slide 4 of Supplementary Data 5). Click on the "Gene Name Batch Viewer" link on the tool menu page. All the gene names will be listed by the Gene Name Batch Viewer. For a gene of interest, one or all of following options may be conducted:

(A) Click on the gene name to link to more detailed information.

(B) Click on "RG" (related genes) beside the gene name to search for other functionally related genes.

(C) Use the browser's "Find" function to search for particular items.

Gene Functional Classification

4| Run "Gene Functional Classification" and explore results (Figure 4 & Slide 5 of Supplementary Data 5). Get back to the tool menu page by clicking on "Start analysis" on the header. Click on "Gene Functional Classification Tool" to classify the input gene list

into gene groups. For any gene groups of interest, one or all of following options may be conducted:

(A) Click on the gene name which leads to individual gene reports for in-depth information about the gene.

(B) Click on the red "T" (term reports) to list associated biology of the gene group.

(C) Click on "RG" (related genes) to list all genes functionally related to the particular gene group.

(D) Click on the "green icon" to invoke 2-D (gene-to-term) view.

(E) Create a new sub-gene list for further analysis on a subset of the genes.

TROUBLESHOOTING 2-D view is a Java Applet application that may take awhile to load for the first time; the 2-D view Java Applet may require you to accept the online security certificate.

CAUTION The input genes are classified using the default clustering stringency. Users may re-run the classification function leading to optimal results for the particular case by re-setting the stringency (high, medium or low) in the options on top of the result page.

Functional Annotation Chart

5| Run "Functional Annotation Chart" (Slide 7 of Supplementary Data 5). Go back to the tool menu page by clicking on "Start Analysis" on the header. **(i)** Click on "Functional Annotation Chart" to go the "Summary Page" of the tool suite. **(ii)** Choose functional annotation categories of your interest (Slide 7 of Supplementary Data 5): Accept 7 default functional annotation categories; Or expand the tree beside each main category (i.e. Main Accessions, Gene Ontology, etc) to select or deselect functional annotation

categories of your interest. **(iii)** Click on the “Functional Annotation Chart” button on the bottom of the page leading to a chart report.

6| Explore the results of the "Functional Annotation Chart" (Figure 5 & Slide 8 of Supplementary Data 5). For an annotation term of interest, one or all of following options may be conducted:

(A) Click on the term name linking to a more detailed description.

(B) Click on “RT” (related terms) to list other related terms.

(C) Click on the “blue bar” to list all associated genes.

(D) Click on a pathway name to view genes on the pathway picture.

CAUTION By default, the order of the annotation terms is based on the EASE(enrichment) score. However, results can also be sorted by different values in the columns; The annotation terms with EASE score <= 0.1 are displayed in the results by default. The stringency of this filter (EASE score cutoff) may be set higher or lower through the options provided at the top of the report page in order to include more or less of the annotation terms.

Functional Annotation Clustering

7| Run “Functional Annotation Clustering” (Slide 10 of Supplementary Data 5). Go back to the tool menu page by clicking on “Start Analysis” on the header. **(i)** Click on “Functional Annotation Clustering” to go to the “Summary Page” of the tool suite. **(ii)** Select annotation categories as described in step 5. **(iii)** Click on the “Functional Annotation Clustering” button on the bottom of the page.

8| Explore the results of "Functional Annotation Clustering" (Figure 6 & Slide 10 of Supplementary Data 5). For an annotation term cluster of interest, one or all of following options may be conducted:

(A) Click on the term name linking to a more detailed description.

(B) Click on "RT" (related terms) to list other related terms.

(C) Click on the "blue bar" to list all associated genes of corresponding individual term.

(D) Click on the red "G" to list all associated genes of all terms within the cluster.

(E) Click on the "green icon" to display the 2-D (gene-to-term) view for all genes and terms within the cluster.

CAUTION The annotation terms are clustered using the default clustering stringency. Users may re-run the classification function leading to optimal results for the particular case by resetting the stringency (high, medium or low) in the options on top of the result page.

Functional Annotation Table

9| Run "Functional Annotation Table" (Slide 11 of Supplementary Data 5). Go back to the tool menu page by clicking on "Start Analysis" on the header and perform the following steps**: (i)** Click on "Functional Annotation Table" to go the "Summary Page" of the tool suite. **(ii)** Select annotation categories as described in step 5. **(iii)** Click on "Functional Annotation Table" button on the bottom of the page.

10| Explore the results of "Functional Annotation Table" (Figure 7 & Slide 11 of Supplementary Data 5). For a gene of your interest:

(A) Click on annotation terms for a detailed description.

(B) Click on “Related Genes” to search functionally related genes.

CAUTION When the output is too large to be displayed by internet browsers, only top 500 records are shown on the result page. However, full results are available to be downloaded as a tab delimited text file through the download link on the top of the result page.

TROUBLESHOOTING

Step	Problem	Possible Reason	Solution
1	Gene ID submission is stuck and I got the message "*You are either not sure which identifier type your list contains, or less than 80% of your list has mapped to your chosen identifier type. Please use the Gene Conversion Tool to determine the identifier type."*	User knows the correct gene ID types, but selected wrong one that did not match the actual input IDs.	Go back to re-submit with correct selection of gene ID type; Or move forward with DAVID Gene ID Conversion tool to determine the potential gene ID type.
		User does not know the correct gene ID type corresponding to their gene list.	Submission panel in DAVID offers a special ID type, called "Not sure". Gene ID submission will be redirected to the DAVID Gene ID Conversion tool which has a mechanism to scan the entire ID system in DAVID to help you to determine the potential ID type(s) of your genes.
		There is more than one type of gene ID in the user's list	DAVID Gene ID Conversion Tool can help you determine the gene ID types and translate them to one single type.
		User's gene ids may contain a version number	Remove the version number since DAVID will not recognize them.
		>=20% of your gene IDs belong to low quality or retired IDs.	DAVID Gene ID Conversion Tool may help to identify the problem IDs. User should consider removing them from the gene list, or move forward to analysis ignoring the problem.
1, 3	The gene number that DAVID recognizes does not match the number in my gene list	Repeated IDs in user's list	DAVID ID submission will automatically remove redundancy.
		Particular ID(s) mapped to many different genes	DAVID ID Conversion Tool could help to identify the problem IDs. User should consider removing the 'bad' ID(s) from the gene list
		User's input gene IDs are gene symbol	Gene symbol is not species-specific so that one symbol may be mapped to many homologous genes across different species. You can define particular species matching your study, after the gene symbols are submitted, in the gene list manager.
3, 4, 5, 6,7, 8, 9, 10	Result page is blank or empty	30 minute timeout	If your web browser is inactive longer than 30 minutes, DAVID will clean up all information (your gene list, etc.) on the server side. Thus, you have to restart your analysis by resubmitting your gene list.
		The size of the gene list is too small	Enrichment or clustering algorithms are based on the survey between input genes against background genes. Thus, a reasonable size (e.g. >30) of input genes is required. Otherwise, certain algorithms will not work properly.
		The cutoff or stringency options are too high	Lower down the thresholds accordingly

		Wrong background selected	Background is automatically set up as the genome-wide set of genes corresponding to the species for the majority of genes in the gene list. Sometimes, the system may not choose the appropriate species. User may check and correct the appropriate background through the Gene List Manager.
		Small or minor species	Some small species may have very little annotation for the genes. There is nothing that can be done about this situation. Alternatively, you could map the genes to the homologous genes in a better annotated species.
4, 8	2-D view not displayed	Network certificate	Please accept it by clicking on "Accept". Basically, you are telling your browser to trust the DAVID application.
		Java plug-in not enabled	By default, most browsers should have the Java plug-in enabled. In case yours is not, please turn it on through Internet Options.
N/A	Service too slow	Slow computer and/or internet speed	Make sure that you have a reasonably good computer and internet speed. See recommendations in Material section.
		Gene list too large (>3,000)	Please be patient.
		DAVID server overwhelmed	The DAVID service may sometimes be slow due to too many large, simultaneous requests. We have monitoring programs to auto detect and fix the situation in a short time period. If the situation is not resolved in a reasonable amount of time, please report the problem to the DAVID team through the contact provided on the DAVID website.

ANTICIPATED RESULTS

We now submit the example gene list (~400 genes), derived from an HIV microarray study[16], to DAVID in order to illustrate the results obtained from various DAVID analytic modules. More detailed information and the availability regarding the gene list can be found in the Materials and Introduction sections of this protocol. Moreover, Supplementary Data 5 provides screen shots of each major step of the following analysis.

Submission of User's Gene IDs to DAVID

A successful submission of the gene list to DAVID is shown in Figure 2 (also Slide 2 & 3 of Supplementary Data 5). Users should see the progress bar under the header move through submission and disappear upon completion of the submission. The gene list manager panel on the left side, thereafter displays the list name (e.g. Upload_list_1) and corresponding species information (Home Sapiens [391]). The number (i.e. 391) appended after the species information is the number of genes that are recognized by DAVID. A set of hyperlinks on right side page lists the analytic modules available in the DAVID analytic pipeline. Users may follow the order of the pipeline to conduct analysis or jointly use analytic modules in varying combinations in order to meet the user's specific needs (Figure 1). Most importantly, the page serves as a central page (Figure 2) for users to choose analytic modules. Users may go back to this page at any time by clicking the "Start Analysis" button on the header in order to switch back and forth among analytic modules as needed.

Gene Name Batch Viewer

The corresponding gene names of input gene IDs were displayed as shown in Figure 3 and Slide 4 of Supplementary Data 5. Users may explore the gene names to examine whether there are any interesting study or marker genes in the list. Many immune related genes, containing names like "*interleukin*", "*chemokine*", "*kinase*" and "*tumor necrosis factor*" can be found in the list, which are consistent with that reported in the publicaton[16]. A set of hyperlinks provided for each gene can further lead to more detail information about a given gene. In addition, by clicking on the "RG" (related genes) search function beside a gene name, for example, "interleukin 8", all other functionally related chemokine genes (e.g. *cxcl 1, 2, 3, 4, 20*) will be listed so that users will be able to see other functionally similar genes in the list based on the bait gene.

Gene Functional Classification

The tool classified the example gene list(~400 genes) into 10 functional groups in an easily readable tabular format. An example output is illustrated in Figure 4 as well as in Slide 5 of Supplementary Data 5. Gene groups (with significant enrichment scores >=1), such as cytokines/chemokines (Group 1: 3.39), kinases (group 2: 2.21), clathrin membrane fusion genes (Group 3: 1.86), transcription factors (Group 6: 1.39), etc., can be easily identified. All of these gene groups are highly relevant to an HIV study and are therefore expected biological results[16]. Organizing the large gene list into gene groups allows investigators to quickly focus on the overall major common biology associated

with a gene group rather than one gene at a time, thereby avoiding dilution of focus during the analysis due to too many single genes. Furthermore, the "2-D View" function associated with each group is able to display all related terms and genes in detail in one picture, in order to examine their inter-relationships. For example, for the kinase group (Group 2), a user who is not familiar with kinases may explore the terms of kinase activity, transferase activity, ATP-binding, nucleotide binding, protein metabolism, tyrosine specificity, serine/threonine specificity, regulation of G protein signaling, signal transduction, and so on in one view at the same time (Slide 6 of Supplementary Data 5). Therefore, we can quickly learn the biology for the kinase group, with the above related terms in a single view and also identify the fine differences among them. For example, there are two G-protein coupled receptor kinases, three protein tyrosine kinases and six kinases involved in cell surface receptor-linked signal transduction among the 23 kinases within the group. The fine details may be very important for pinpointing the key biology associated with a study.

Functional Annotation Chart

Over five hundred enriched (over-represented) biological terms were reported (Figure 5 and Slide 8 of Supplementary Data 5). Many of them are highly immune related, such as response to pathogenic bacteria, chemokine activity, cell migration, clathrin coated vesicle membrane, kinase activity, RNA polymerase II transcription factor activity, cell communication. This is consistent with observations previously identified by the other analytic modules, as well as meeting the expectation for the HIV study[16].

The report offers a lot of redundant details regarding the enriched biology associated with the gene list, which certainly helps the interpretation of the biology, but also may dilute the focus. Moreover, a set of hyperlinks provided for each term will lead to more details about each term, such as in-depth description, associated genes, other related terms, directed acyclic graph (DAG) of GO, etc. Notably, the pathway viewer module offers visualization of users' genes on enriched pathways. For example, "IL-10 Anti-inflammatory Signaling Pathway" was reported in the output. We can observe that IL10 was activated as an upstream immune regulator and was then further regulated by HO-1. As result, the IL1/TNFa/IL6 complex was activated leading to further downstream inflammatory responses (Figure 8). Thus, the inter-relationship of input genes was examined on the pathway in a network context.

Functional Annotation Clustering

The tool condensed the input gene list into smaller, much more organized biological annotation modules in a similar format (Figure 6 & Slide 10 of Supplementary Data 5) as that of Gene Annotation Clustering, but in a term-centric manner. Similarly, it allows investigators to focus on the annotation group level by quickly organizing many redundant/similar/hierarchical terms within the group. Annotation clusters, such as immune response, transcriptional regulation, chemokine activity, cytokine activity, kinase acitivity, signaling transduction, cell death, etc., could be found on the top of output as expected for this study[16]. The highly organized and simplified annotation results allow users to quickly focus on the major biology at an annotation cluster level instead of trying

to derive the same conclusions by putting together pieces that are scattered throughout a list of hundreds of terms in a typical term enrichment analysis. In addition, the 'G' (genes) link provided for each cluster can comprehensively pool all related genes from different terms within the cluster. For example, each of the 7 terms within cluster 2 (inflammatory response cluster) associates with both overlapping as well as differing genes. Therefore, a pooled gene list brought together by cluster 2 regarding inflammatory response may be much more comprehensive, compared to the genes selected from one or a few individual terms.

Summary

Collectively, all of the DAVID analytic modules aim to extract biological meaning from the given gene list from different biological angles with highly consistent and expected results for a given study. Integration of the results from the different analytic modules (Figure 1) will take advantage of the different focus and strength of each module, in order to make the overall biological picture, assembled based on the gene list, more comprehensive and detailed. For a given gene list, DAVID Bioinformatics Resources is able to help users to (Figure 1 b):

- Convert gene IDs from one type to another
- Diagnose and fix problems with gene IDs
- Explore gene names in batch
- Discover enriched functionally-related gene groups
- Display relationship of many-genes-to-many-terms on 2-D view.
- Initial glance of major biological functions associated with my gene list
- Identify enriched (over-represented) annotation terms
- Visualize genes on BioCarta & KEGG pathway maps

- Link gene-disease associations
- Highlight protein functional domains and motifs
- Redirect to related literature
- List interacting proteins
- Cluster redundant and heterozygous annotation terms
- Search other functionally similar genes in genome, but not in my list
- Search other annotations functionally similar to one of my interest
- Read all annotation contents associated with a gene
- And more

ACKNOWLEDGMENTS

The authors are grateful to the referees for their constructive comments and thank Robert Stephens, David Bryant and David Liu in the ABCC group for web server support. Thanks also go to Xin Zheng and Jun Yang in the Laboratory of Immunopathogenesis and Bioinformatics (LIB) group for discussion. We also thank Bill Wilton and Mike Tartakovsky for information technology and network support. The project has been funded with federal funds from the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), under Contract No. NO1-CO-56000. The annotation of this tool and publication do not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the United States Government.

REFERENCES

1. **Huang da, W. et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res* 35, W169-75 (2007).**
2. **Dennis, G., Jr. et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4, P3 (2003).**
3. **Hosack, D.A., Dennis, G., Jr., Sherman, B.T., Lane, H.C. & Lempicki, R.A. Identifying biological themes within lists of genes with EASE. *Genome Biol* 4, R70 (2003).**
4. **Zeeberg, B.R. et al. High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). *BMC Bioinformatics* 6, 168 (2005).**
5. **Beissbarth, T. & Speed, T.P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* 20, 1464-5 (2004).**
6. **Khatri, P., Bhavsar, P., Bawa, G. & Draghici, S. Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res* 32, W449-56 (2004).**
7. **Martin, D. et al. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol* 5, R101 (2004).**
8. **Al-Shahrour, F., Diaz-Uriarte, R. & Dopazo, J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20, 578-80 (2004).**
9. **Masseroli, M., Galati, O. & Pinciroli, F. GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. *Nucleic Acids Res* 33, W717-23 (2005).**
10. **Lee, J.S., Katari, G. & Sachidanandam, R. GObar: a gene ontology based analysis and visualization tool for gene sets. *BMC Bioinformatics* 6, 189 (2005).**
11. **Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102, 15545-50 (2005).**
12. **Khatri, P. & Draghici, S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21, 3587-95 (2005).**
13. **Sherman, B.T. et al. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics* 8, 426 (2007).**
14. **Huang da, W. et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol* 8, R183 (2007).**
15. **Huang, D.W., Sherman, B.T. & Lempicki, R.A. DAVID gene ID conversion tool. *Bioinformation* 2, 428-430 (2008).**

16. Cicala, C. et al. HIV envelope induces a cascade of cell signals in non-proliferating target cells that favor virus replication. *Proc Natl Acad Sci U S A* 99, 9380-5 (2002).

TABLE

Table 1: Side-by-side comparisons of DAVID's major analytic modules

	Input user's large gene lists				
DAVID Analytic Module/Tool	**Gene Name Batch Viewer[1]**	**Gene Functional Classification[14]**	**Functional Annotation Chart[1,2]**	**Functional Annotation Clustering[14]**	**Functional Annotation Table[1,2]**
Brief definition /explanation	List names of user's genes	Classify user's genes into gene groups	Identify enriched annotation terms associated with user's gene list	Cluster functionally similar terms associated with user's gene list into groups	Query associated terms for all user's genes
Key Points	Gene-centric singular exploration	Gene-centric modular analysis	Term-centric singular enrichment analysis (typical enrichment analysis)	Term-centric modular enrichment analysis	Large scale query
Example question to ask	What are the genes in my list?	What major gene family in my list?	What are enriched annotation terms for my gene list?	What are enriched annotation groups for my gene list?	What are all selected annotations for my genes?
Main Functions	• Display all gene names in a linear tabular text format • Deep links to more information around given gene • Search other functionally related genes	• Classify functionally related genes into groups • 2-D view for related gene-term relationship • Rank importance of gene groups with enrichment score • Highlight annotation terms for gene groups	• Identify enriched annotation terms in a linear tabular text format • Deep links to more information around terms and associated genes • Search other functionally related terms • View genes on pathway maps	• Cluster functionally related annotations into groups • 2-D view for related gene-term relationship • Rank importance of annotation groups with enrichment score • Pool genes for annotation groups	Query selected annotations for given genes.
Advantages	• Roughly explore genes one-by-one • Quickly check if the expected/important genes are in the list • Quickly learn annotation about genes of interests • All genes included in the analysis	• Explore genes group-by-group rather than singular gene one-by-one • Highlight important gene groups by enrichment scores • Study functionally related genes and their relationship	Simple format to explore all singular enriched terms	• Explore annotations group-by-group rather than singular term one-by-one. • Highlight important annotation groups by enrichment scores • Study	• Quickly Explore all annotations (both enrich and not enriched ones) for given genes • Good for analysis of small # of focused genes • Save entire annotation profile of given gene list in text file ready for other external

		in a network format • Good to catch major biology		functionally related genes and their relationship in a network format • Good to focus on major and fine biology	analysis
Drawbacks	• Related genes scattered in the results losing inter-relationships of genes during exploration • Difficult to judge enriched genes or noisy genes without enrichment calculation	Some genes without strong neighbors will be left out from the analysis	• Related/redundant terms scattered in the results • Some fine biology could be diluted by the redundancy • Lack of term-term relationships during analysis	Some enriched terms without strong neighbors will be left out from the analysis	• Difficult to explore large gene list. • No enrichment analysis

Table 2: Major statistical methods and associated parameters used in DAVID.

Step #	Module/Page	Statistics /Parameters	Explanation/Definition	How to Understand the Value
1	Submission of User's Gene IDs	Background genes (or called population genes)	To decide the degree of enrichment, a certain background must be set up in order to be compared to the user's gene list. For example, 10% of user's genes are kinases vs. 1% of genes in human genome (this is population background) are kinases. Thus, the conclusion is obvious in the particular example that the user's study is highly related to kinase. However, 10% itself alone cannot provide such a conclusion without comparing it to the background information.	• A general guideline is to set up the reference background as the pool of genes which have a chance to be selected for the studied annotation category under the scope of users' particular study. • Default background is the entire genome-wide genes of the species matching the user's input IDs. Pre-built backgrounds such as genes in Affy chips, etc. are available for the user's choice. • In principle, a larger gene background tends to give smaller p-values. Since most of the high-throughput studies are, or at least are close to, genome-wide scope, the default background is good for regular cases in general.
4	Gene Functional Classification [1,14]	Classification Stringency	To control the behavior of DAVID Fuzzy clustering.	• A general guideline is to choose higher stringency settings for tight, clean and smaller numbers of clusters; Otherwise lower for loose , broader and larger numbers of clusters. • Default setting is medium. • Five pre-defined levels from lowest to highest for user's choices. • Users may want to play with different stringency for more satisfactory results.
		Enrichment Score (for each group)	To rank overall importance (enrichment) of gene groups. It is the geometric mean of all the enrichment p-values (EASE scores) for each annotation term associated with the gene members in the group. To emphasize that the geometric mean is a relative score instead of an absolute *p* value, minus log transformation is applied on the average p-values	• A higher score for a group indicates that the gene members in the group are involved in more important (enriched) terms in a given study, therefore more attention should go to them. • Enrichment score of 1.3 is equivalent to non-log scale 0.05. thus, more attention should go to groups with scores >= 1.3 • However, the gene groups with lower scores could be potentially interesting, and should be explored as well if possible.
6	Functional Annotation Chart [1,2]	P-value (or called EASE Score)	To examine the significance of gene-term enrichment with a modified Fisher's Exact Test (EASE Score). For example, 10% of user's genes are kinases vs. 1% of genes in human genome (this is population background) are kinases. Thus, the EASE score is <0.05 which suggests that kinases are significantly more enriched than random chance in the study for this particular example.	• The smaller the p-values, the more significant they are. • Default cutoff is 0.1 • Users could set different levels of cutoff through option panel on the top of result page. • Due to the complexity of biological data mining of this type, p-values are suggested to be treated as score systems, i.e. suggesting roles, rather than decision making roles. Users themselves should play critical roles in judging "are the results making sense or not for expected biology".
		Benjamini	To globally correct enrichment p-values in order to control family-wide false discovery rate under certain rate (e.g. <= 0.05). It is one of the multiple testing correction techniques (Bonferroni, Benjamini and FDR) provided by DAVID.	• More terms examined, more conservative the corrections are. As a result, all the p-values get larger. • It is great if the interesting terms have significant p-values after the corrections. But since the multiple testing correction techniques are known as conservative approaches, it could hurt the sensitivity of discovery if over emphasizing them. Users' judgment could be critical as discussed in EASE Score in Functional Annotation Chart section.
		Fold Enrichment	To measure the magnitude of enrichment. For example, 10% of user's genes are kinases vs. 1% of genes in human genome (this is population background) are kinases. Thus, the fold enrichment is 10 fold. Fold enrichment	• Fold enrichment 1.5 and above are suggested to be considered as interesting. • Fold enrichment and EASE Score should be always examined side-by-side. Terms with larger fold enrichments and smaller may be interesting. • Caution should be taken when big fold enrichment are

			along with EASE Score could rank the enriched terms in a more comprehensive way.	obtained from a small number of genes (e.g. <=3). This situation often happens to the terms with a few genes (more specific terms) orfsmaller size (e.g. <100) of user's input gene list. In this case, the reliability is not as much as those fold enrichment scores obtained from larger numbers of genes.
		%	# of genes involved in given term is divided by the total # of user's input genes, i.e. Percentage of user's input gene hitting a given term. For example, 10% of user's genes hit "kinase activity".	• It gives overall idea of gene distributions among the terms. • The higher percentage does not necessarily have a good EASE Score because it also depends on the percentage of background genes as discussed in the EASE Score in Functional Annotation Chart section.
8	Functional Annotation Clustering[1,14]	Classification Stringency	To control the behalvior of DAVID Fuzzy clustering.	• A general guideline is to choose higher stringency setting for tight, clean and smaller numbers of clusters; Otherwise lower for looser , broader and larger numbers of clusters. • Default setting is medium. • Five pre-defined levels from lowest to highest for user's choices. • Users may want to play with different stringency to obtain more satisfactory results.
		Enrichment Score (for each group)	To rank overall importance (enrichment) of annotation term groups. It is the geometric mean of all the enrichment p-values (EASE scores) of each annotation term in the group. To emphasize that the geometric mean is a relative score instead of an absolute *p* value, minus log transformation is applied on the average p-values	• A higher score for a group indicates that annotation term members in the group are playing more important (enriched) roles in given study, therefore pay more attention toward them. • Enrichment score 1.3 is equivalent to non-log scale 0.05. thus, more attention should go to groups with scores >= 1.3 • However, the annotation groups with lower scores could be potentially interesting, and should be explored as well if possible.
		P-value (or called EASE Score) (for individual term members)	To examine the significance of gene-term enrichment with a modified Fisher's Exact Test (EASE Score). This p-value is calculated in exactly the same way as in the Functional Annotation Chart section.	The explanation is the same as that in Functional Annotation Chart section.
		Benjamini	To globally correct enrichment p-values of individual term members. The idea and calculations are exactly the same as that in Functional Annotation Chart section.	The explanation is the same as that in the Functional Annotation Chart section.

FIGURES

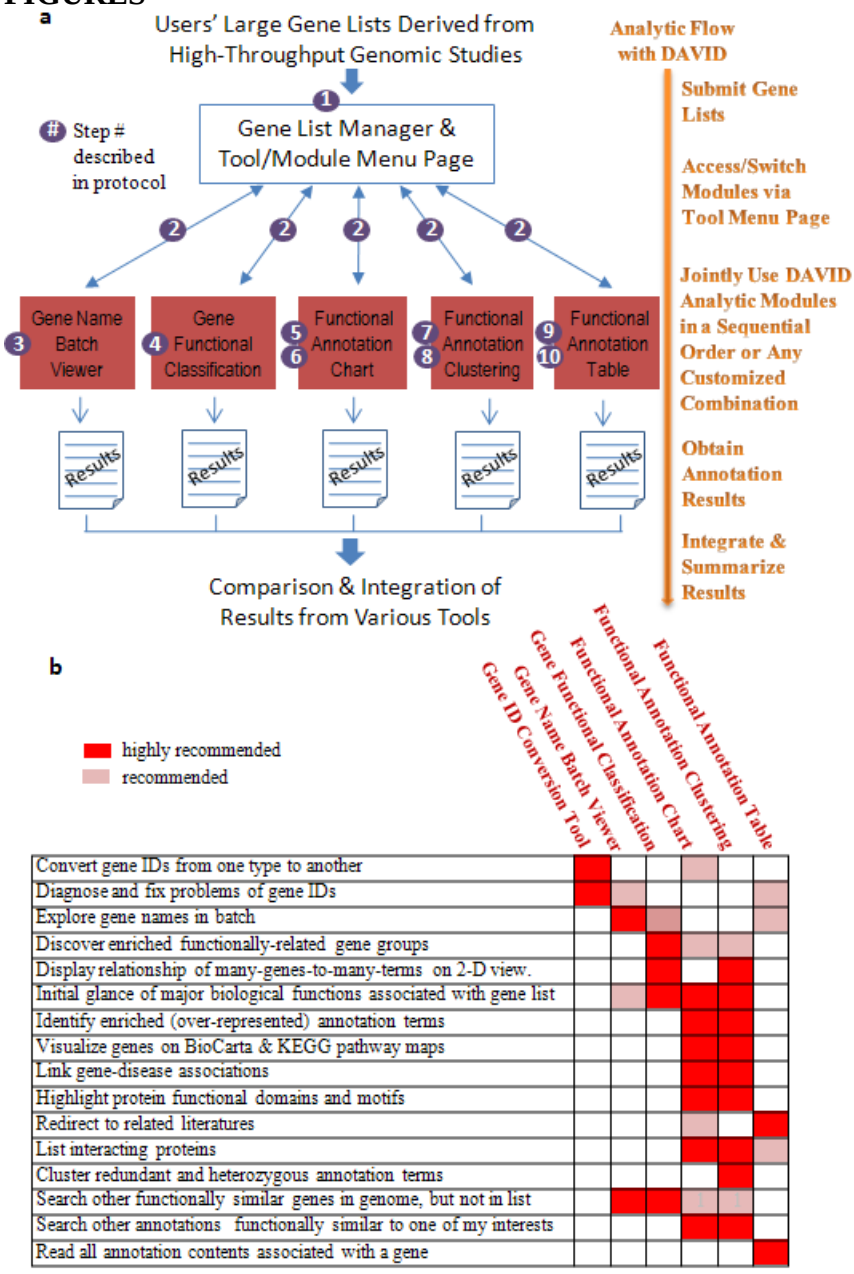

Figure 1. Analytic tools/modules in DAVID. a. After the user's gene list is submitted to DAVID, the gene list manager may be accessed by all DAVID analytic modules (red boxes) at any time. The circled numbers indicate step numbers described in the procedure section to facilitate reading. **b.** DAVID analytic modules, each having different strengths and focus, can be used independently or jointly. A roadmap to help users to choose some or all DAVID analytic modules for the analysis of large gene lists.

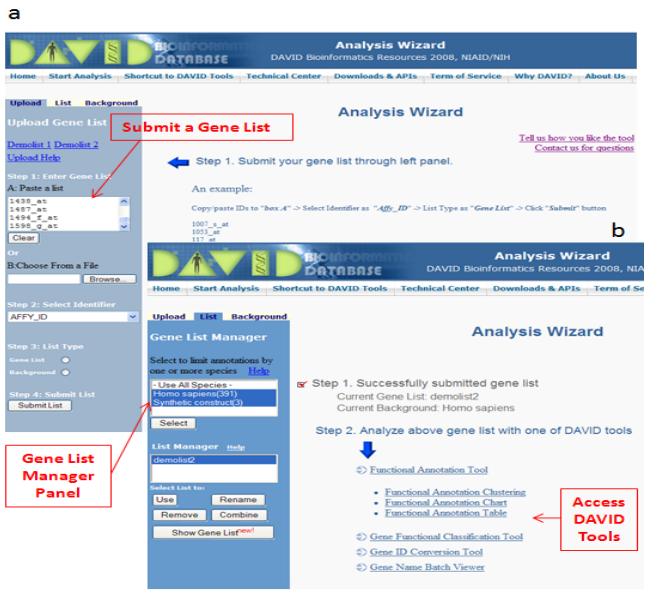

Figure 2. Submit a gene list to DAVID and access various analytic tools/modules. **a.** Following the example input format and steps on the left side uploading panel, a list of genes may be uploaded into DAVID. **b.** After successfully uploading a gene list(s), a set of analytic modules are available for the analysis of the current gene list highlighted in the gene list manager on the left side. Importantly, users may go to this page at any time by clicking "Start Analysis" on the header in order to access any analytic tool of interest.

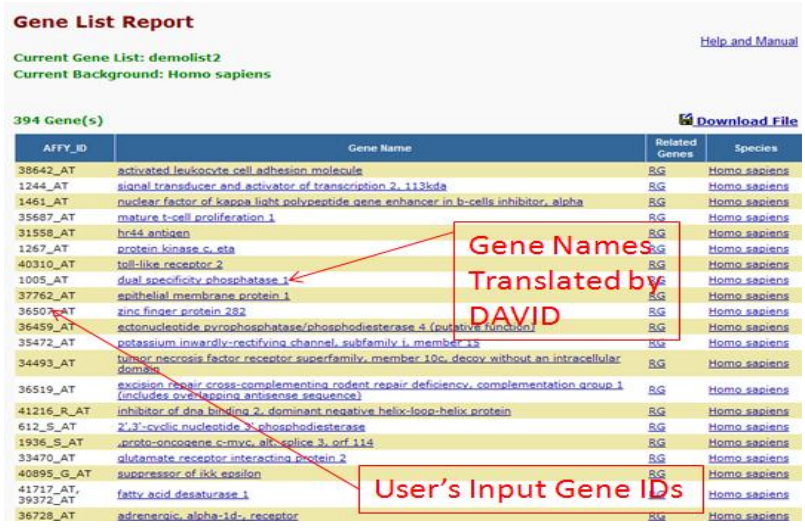

Figure 3. **An example layout of DAVID Gene Name Batch Viewer.** User's input gene IDs are translated into meaningful and readable gene names. The link on each gene name can lead to more in-depth information.

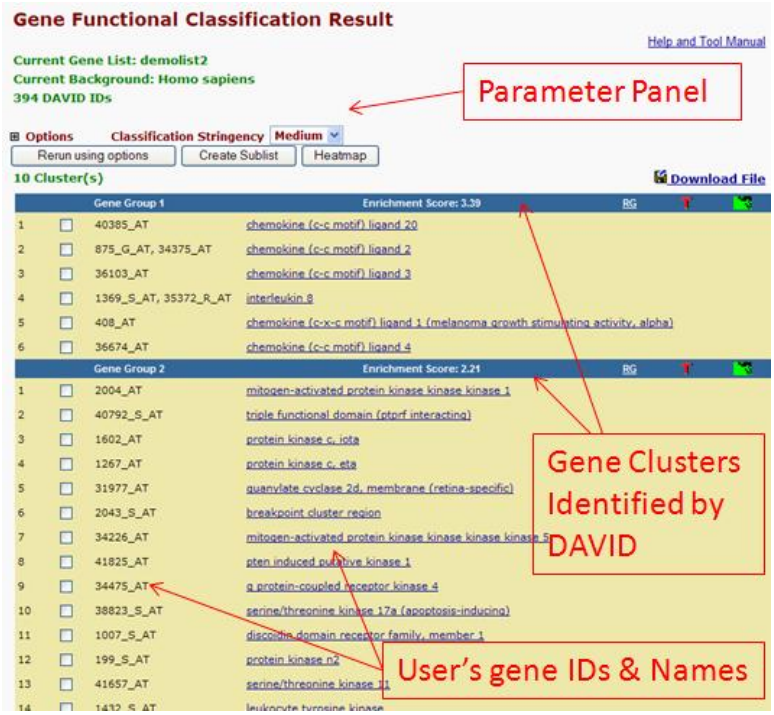

Figure 4. An example layout of DAVID Gene Functional Classification. User's genes were organized and condensed into several functional groups. The gene members in each group share common biological functions. A set of accessory tools provided for each group will further facilitate the 'drill-down' analysis of biological inter-relationships among the gene members within the same group.

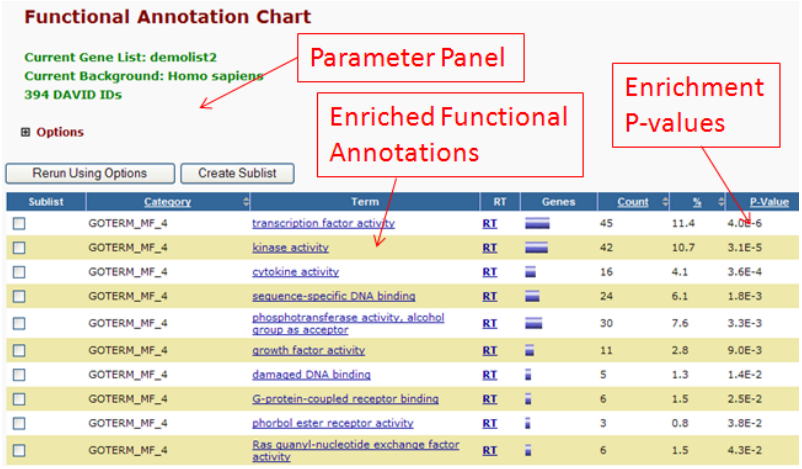

Figure 5. An example layout of DAVID Annotation Chart. The enriched functional annotation terms associated with user's gene list are identified and listed according to their enrichment p-value by DAVID. The links on the page can lead to various detail information regarding corresponding items.

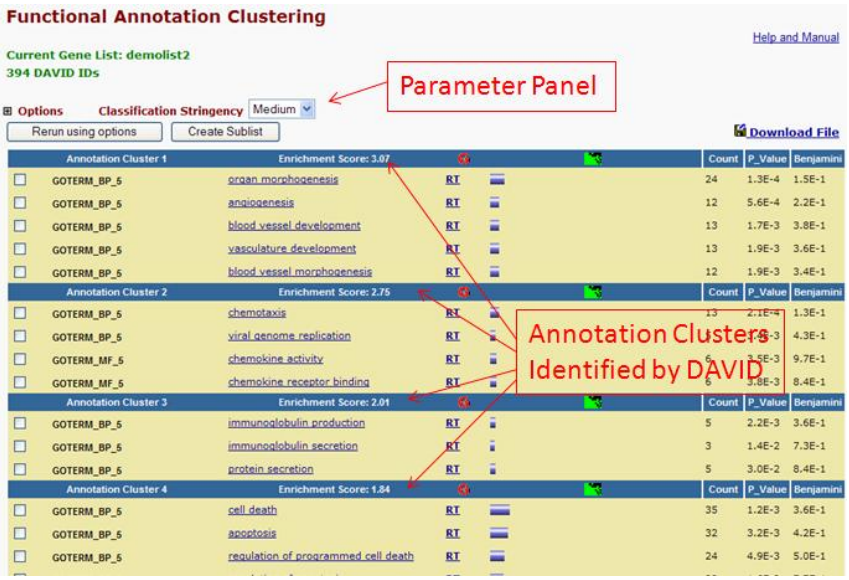

Figure 6. An example layout of DAVID Functional Annotation Clustering. The similar annotation terms are grouped into clusters so that user can read through the important terms in a way of block-by-block instead of individual-by-individual.

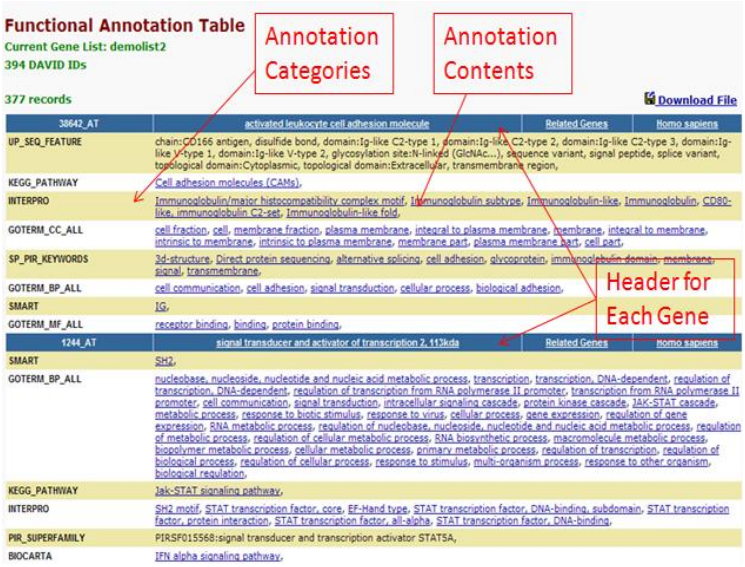

Figure 7. An example layout of DAVID Annotation Table. Various annotation contents for given gene are list in a tabular format. The contents for each gene are separated by the header rows in blue color.

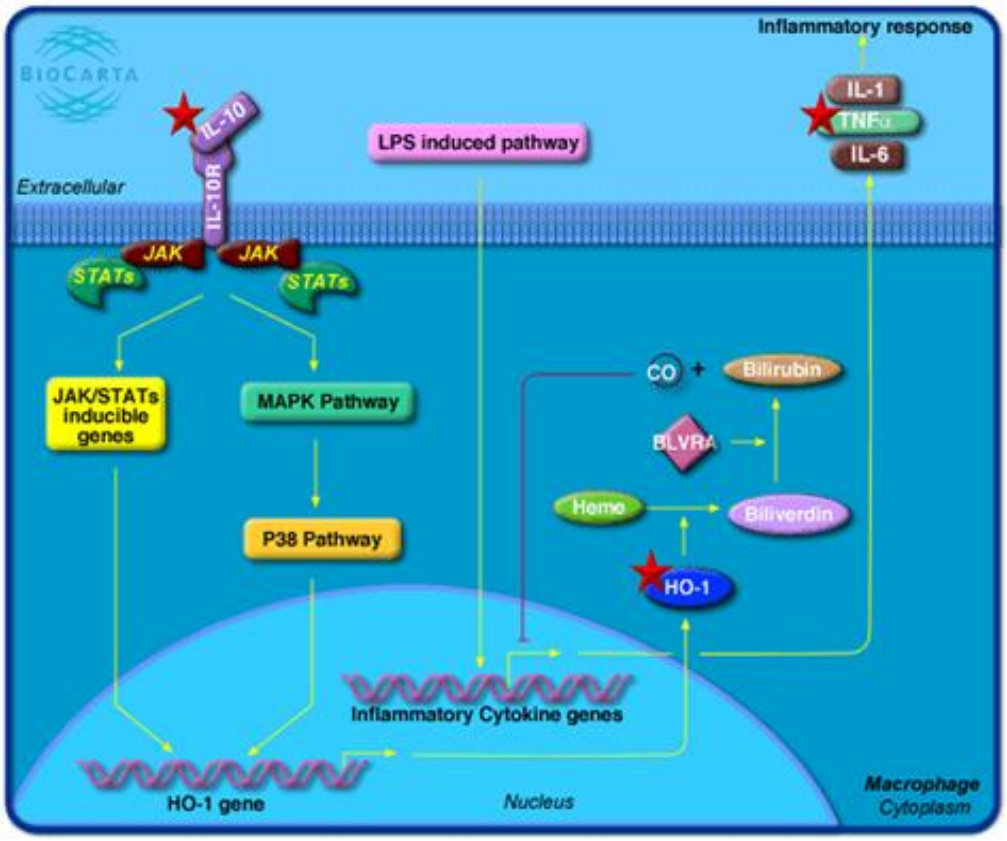

Figure 8. Pathway map viewer. The red star indicates the associations between pathway genes and the user's input genes. Following the pathway flow, IL10 was activated as an upstream immune stimulator. Then the middle stream gene, HO-1, was involved. IL-1/INFa/IL-6, as downstream regulator, was finally activated. Thus, the user's genes may be analyzed in a network context.

SUPPLEMENTARY DATA

Supplementary Data 1. Collection of ~68 similar enrichment analysis tools. The tools are roughly categorized into three classes according to their backend algorithms. Reference links are provided for more information of each tool.

Supplementary Data 2. ~400 Affymetrix IDs[16] used in this paper.

Supplementary Data 3. Comparisons of the enrichment p-values between gene lists derived from microarray study vs. same size of gene lists generated randomly. A 'good' gene lists should consistently contain more enriched biology than that of random list in the same sizes.

Supplementary Data 4. Summaries of gene identifier types and annotation categories supported in the DAVID system.

Supplementary Data 5. Screen shots of each major analysis step according to the description in the manuscript.

Supplementary Data 6. Examples for the input formats of a gene list.