DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists

Da Wei Huang1, Brad T. Sherman1, Qina Tan1, Joseph Kir1, David Liu2, David Bryant2, Yongjian Guo5, Robert Stephens2, Michael W. Baseler3, H. Clifford Lane4 and Richard A. Lempicki1,∗

1Laboratory of Immunopathogenesis and Bioinformatics, 2Advanced Biomedical Computing Center and 3Clinical Services Program, SAIC-Frederick, Inc., National Cancer Institute at Frederick, MD 21702, USA, 4Laboratory of Immunoregulation, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, 20892, USA, 5Bioinformatics and Scientific IT Program, NIAID Office of Technology Information Systems, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, 20892, USA

Received January 22, 2007; Revised April 14, 2007; Accepted May 6, 2007

ABSTRACT

All tools in the DAVID Bioinformatics Resources aim to provide functional interpretation of large lists of genes derived from genomic studies. The newly updated DAVID Bioinformatics Resources consists of the DAVID Knowledgebase and five integrated, web-based functional annotation tool suites: the DAVID Gene Functional Classification Tool, the DAVID Functional Annotation Tool, the DAVID Gene ID Conversion Tool, the DAVID Gene Name Viewer and the DAVID NIAID Pathogen Genome Browser. The expanded DAVID Knowledgebase now integrates almost all major and well-known public bioinformatics resources centralized by the DAVID Gene Concept, a single-linkage method to agglomerate tens of millions of diverse gene/protein identifiers and annotation terms from a variety of public bioinformatics databases. For any uploaded gene list, the DAVID Resources now provides not only the typical gene-term enrichment analysis, but also new tools and functions that allow users to condense large gene lists into gene functional groups, convert between gene/protein identifiers, visualize many-genes-to-many-terms relationships, cluster redundant and heterogeneous terms into groups, search for interesting and related genes or terms, dynamically view genes from their lists on bio-pathways and more. With DAVID (http://david.niaid.nih.gov), investigators gain more power to interpret the biological mechanisms associated with large gene lists.

INTRODUCTION

In the post-genomic era, biological interpretation of large gene lists derived from high-throughput experiments, such as genes from microarray experiments, is a challenging task. The first version of DAVID (the Database for Annotation, Visualization and Integration Discovery), released in 2003 (1,2), as well as a number of other similar publicly available high-throughput functional annotation tools (3–23), partially address the challenge by systematically mapping a large number of interesting genes in a list to associated Gene Ontology (GO) terms (10), and then statistically highlighting the most over-represented (enriched) GO terms out of a list of hundreds or thousands of terms. This increases the likelihood that the investigator will identify the biological processes most pertinent to the biological phenomena under study (19). While this tool is extremely useful and has been cited in hundreds of publications during the past three years, the development of other effective data mining algorithms, as additional components to the DAVID Bioinformatics Resources, will improve the power of investigators to analyze their gene lists from different biological angles. The newly added contents, functions

*To whom correspondence should be addressed. Tel: +1-301-846-7114; Fax: 301-846-7672; Email: rlempicki@mail.nih.gov

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
and tool suites in the DAVID Bioinformatics Resources intend to address several issues that other tools have not been able to extensively address: (i) to dramatically expand the biological information coverage in the DAVID Knowledgebase by comprehensively integrating more than 20 types of major gene/protein identifiers and more than 40 well-known functional annotation categories from dozens of public databases; (ii) to address the enriched and redundant relationships among many genes-to-many-terms (i.e. one gene could associate with many different, redundant terms and one term could associate with many genes) by developing a set of novel algorithms, such as the DAVID Gene Functional Classification Tool, the Functional Annotation Clustering Tool, the Linear Searching Tool, the Fuzzy Gene-Term Classification Tool, the Functional Annotation Clustering Tool, the Heat Map Viewer, etc.; (iii) to dynamically visualize genes from a users list within the most relevant KEGG and BioCarta pathways with the DAVID Pathway Viewer; (iv) to allow users to create and use customized gene backgrounds for typical gene-term enrichment analysis; (v) to provide computational power and (v) to facilitate efficient communication and experience exchange within the scientific community by moderating the DAVID Forum.

This article summarizes the key DAVID components and tool suites in the newly released DAVID Bioinformatics Resources, highlighting new or expanded analytic features that provide investigators with additional means to explore and extract biological meaning from large gene lists that users input to the system (Supplementary File 1). For in-depth algorithm information, appropriate references and supplementary materials are provided.

### FEATURES AND FUNCTIONALITIES

#### Computational Infrastructure

The aim of the DAVID software design is to provide users with the simplest usability and fastest exploration speed through better internal software engineering practices. Therefore, the DAVID Bioinformatics Tools, as web-based applications on a Tomcat web server in a Linux machine (4-CPU for 3.5 GHz speed, 8 GB memory), requires no configuration and installation in the client’s computers. Java is the primary language used for all of the server side components of the calculation engines and the Java Server Page (JSP) web interfaces, in a full object-oriented fashion. In-memory Java data objects holding all genes-to-annotation information up to 2.5 GB in size were developed to greatly increase the data IO speed compared to that through typical relational databases (e.g. Oracle). The Java Remote Method Invocation (RMI), a distributed computing technique, is also used to take advantage of multiple computing resources. A set of automated programs monitors many aspects of the web services in order to maximize the performance and minimize the down time period.

<table>
<thead>
<tr>
<th>Gene ID Type</th>
<th>Total ID</th>
<th>Unique Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFFY_ID</td>
<td>2254679</td>
<td>845117</td>
</tr>
<tr>
<td>ENTREZ_GENE_ID</td>
<td>1734858</td>
<td>160233</td>
</tr>
<tr>
<td>GENPEPT_ACCESSION</td>
<td>4065385</td>
<td>2511637</td>
</tr>
<tr>
<td>GENBANK_ACCESSION</td>
<td>16828735</td>
<td>2490120</td>
</tr>
<tr>
<td>GENEBank_ID</td>
<td>20291282</td>
<td>235084</td>
</tr>
<tr>
<td>PIR_ACCESSION</td>
<td>2922818</td>
<td>258079</td>
</tr>
<tr>
<td>PIR_ID</td>
<td>308092</td>
<td>266645</td>
</tr>
<tr>
<td>PIR_NREF_ID</td>
<td>3355759</td>
<td>2657404</td>
</tr>
<tr>
<td>REFSEQ_GENOMIC</td>
<td>1866800</td>
<td>1552597</td>
</tr>
<tr>
<td>REFSEQ_MRNA</td>
<td>645831</td>
<td>561447</td>
</tr>
<tr>
<td>REFSEQ_PROTEIN</td>
<td>1644632</td>
<td>1373467</td>
</tr>
<tr>
<td>REFSEQ_RNA</td>
<td>1364</td>
<td>852</td>
</tr>
<tr>
<td>UNIGENE</td>
<td>161138</td>
<td>158938</td>
</tr>
<tr>
<td>UNIPROT_ACCESSION</td>
<td>2864344</td>
<td>2097488</td>
</tr>
<tr>
<td>UNIPORT_ID</td>
<td>2789453</td>
<td>2096712</td>
</tr>
<tr>
<td>UNIREF100_ID</td>
<td>2552342</td>
<td>2088692</td>
</tr>
<tr>
<td>OFFICIAL_GENE_SYMBOL</td>
<td>1693151</td>
<td>1600966</td>
</tr>
<tr>
<td>FLYBASE_ID</td>
<td>27109</td>
<td>26642</td>
</tr>
<tr>
<td>HAMAP_ID</td>
<td>63925</td>
<td>63822</td>
</tr>
<tr>
<td>HSSP_ID</td>
<td>265000</td>
<td>258750</td>
</tr>
<tr>
<td>TIGR_ID</td>
<td>120117</td>
<td>111699</td>
</tr>
<tr>
<td>WORMBASE_ID</td>
<td>43675</td>
<td>21243</td>
</tr>
<tr>
<td>RGD_ID</td>
<td>25230</td>
<td>25060</td>
</tr>
<tr>
<td>NOT_SURE</td>
<td>ALL IDs</td>
<td></td>
</tr>
</tbody>
</table>

Any of the gene identifier types above can be cross-mapped to the DAVID Knowledgebase. ‘Not Sure’ is a new ID type specifically designed for the DAVID web site. For a given ‘not sure’ ID, all possible matching IDs will be systematically scanned across the entire DAVID collection.

### DAVID Knowledgebase

A highly integrated gene-annotation database with comprehensive data coverage is essential for the success of any high-throughput annotation algorithms. Due to the complex and distributed nature of biological research, our current biological knowledge is distributed among many redundant annotation databases maintained by independent groups. One gene could have several different identifiers within one or more database(s). Similarly, the biological terms associated with different gene identifiers for the same gene could be collected in different levels across different databases. Due to these issues, most high-throughput annotation tools rely on one, or at most a few, resource(s), which limits the analytic comprehensiveness and the level of throughput. The DAVID Knowledgebase is now built around the ‘DAVID Gene Concept’, a single linkage method to agglomerate tens of millions of gene/protein identifiers from a variety of public genomic resources (Table 1), including NCBI, PIR and Uniprot (24–27), into broader secondary gene clusters, called the DAVID Gene Concept (Figure 1, and more technical details at http://david.abcc.ncifcrf.gov/limits/knowledgebase/DAVID_gene.html). Grouping these gene identifiers improves cross-referencing capability, allowing more than 40 categories of publicly available functional annotation to be comprehensively assigned to and centralized by the DAVID Gene Concept (Table 2, see Supplementary File 2 for a complete list of annotation.
sources and more technical details at http://david.abcc.ncifcrf.gov/helps/knowledgebase/DAVID_gene.html). To the best of our knowledge, this annotation coverage far exceeds that of the original DAVID database and those currently used by other similar high-throughput annotation tools. The DAVID knowledgebase not only increases the accessibility to a wide range of heterogeneous annotation data in one centralized location, but also enhances the comprehensiveness of high-throughput gene functional analysis by overlapping multiple biological aspects together. It also provides a solid foundation for the further development of more advanced high-throughput analytic algorithms that may be added to the DAVID Bioinformatics Resources. More importantly, the entire DAVID Knowledgebase, in simple pair-wise text format files containing a broad, highly integrated annotation

---

**Table 2.** The wide-range collection of heterogeneous functional annotations in the DAVID Knowledgebase

| Category                          | GO_BIOLOGICAL_PROCESS | GO_MOLECULAR_FUNCTION | GO_CELLULAR_COMPONENT | PANTHER_BIOLOGICAL_PROCESS | PANTHER_MOLECULAR_FUNCTION | COG_KOG_ONTOLOGY | P-P Interaction | BIND | DIP | MINT | NCICB_CAPATHWAY | TRANSFAC_ID | HIV_INTERACTION | HIV_INTERACTION_CATEGORY | HPRD_INTERACTION | REACTOME_INTERACTION | Disease Association | GENETIC_ASSOCIATION_DB | OMIM_DISEASE | Literature | GENERIF_SUMMARY | PUBMED_ID | HIV_INTERACTION_PUBMED_ID |
|----------------------------------|-----------------------|-----------------------|-----------------------|-----------------------------|-----------------------------|--------------------------|-----------------|-----------------|-----|-----|-----|----------------|-------------|----------------|---------------------|----------------|-----------------|----------------------|-------------------|--------------|------------|-------------|----------|---------------------|
| **Ontology** (>40 million records) | BLOCKS_ID             | COG_KOG_NAME          | INTERPRO_NAME          | PIR_ID                      | PIR_HOMOLOGY_DOMAIN         | PIR_SUPERFAMILY_NAME      | PRINTS_NAME      | PRODOM_NAME    | PROSITE_NAME   | SCOP_ID         | SMART_NAME        | TIGRFAM_S_NAME   | PANTHER_SUBFAMILY | PANTHER_FAMILY   | BioCarta         | KEGG_PATHWAY   | PANTHER_PATHWAY | PDB       | PIR_SUPERFAMILY_NAME |
| **Protein Domain/Family** (>15 millions) | ALIAS_GENE_SYMBOL     | CHROMOSOME            | CYTOBAND               | GENE_SYMBOL                 | HOMOLOGOUS_GENE            | ENTREZ_GENE_SUMMARY     | OMIM_ID          | PROTEIN_MB     | REFSSEQ_PRODUCT | SEQUENCE_LENGTH | SP_COMMENT       | PIR_SEQ_FEATURE  | SP_COMMENT_TYPE  | SP_PIR_KEYWORDS  | GNF Microarray  | UNIGENE EST    | CGAP SAGE     | CGAP EST     | |
| **Sequence Features** (>21 millions) | GNF Microarray         | UNIGENE EST            | CGAP EST                | |

Over 60 functional categories from dozens of independent public sources (databases) (see Supplementary File 2 for a complete list) are collected and integrated in the DAVID Knowledgebase.

---

**Figure 1.** A DAVID gene constructed by a single linkage algorithm. Two UniRef100 clusters, two NRef 100 clusters and one Entrez Gene cluster were systematically found sharing one or more protein identifiers with each other. The single-linkage rule can further iteratively agglomerate them as a whole into one DAVID gene. Thus, for this particular example of tyrosine-protein phosphatase non-receptor type 21 (PTPN21), the resulting DAVID gene is able to collect and integrate all gene/protein identifiers more comprehensively than each original gene cluster.
DAVID Functional Annotation Tool Suite

This tool suite (http://david.abcc.ncifcrf.gov/summary.jsp), introduced in the first version of DAVID, mainly provides typical batch annotation and gene-GO term enrichment analysis to highlight the most relevant GO terms associated with a given gene list (2). The new version of the tool keeps the same enrichment analytic algorithm but with extended annotation content coverage, 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 general features, homologies, gene functional summaries, gene tissue expressions, literatures, etc. (Table 2). The improved annotation coverage alone provides investigators with much more power to analyze their genes using many different biological aspects in a single space. Flexible options are provided to display results in an individual annotation chart report or a combined chart report. In addition to pre-built gene population backgrounds (e.g. Affy U133) used in gene-annotation enrichment analysis, with its improved computational power, the new tool accepts user-defined population gene list, an option rarely found in other similar web-based, high-throughput annotation tools. This feature was added in order to more specifically meet the users’ requirements for the best analytical results.

The DAVID Functional Annotation Clustering is a newly added feature (manuscript submitted, and more details at http://david.abcc.ncifcrf.gov/manuscripts/fuzzy_cluster/) to the DAVID Functional Annotation Tool. This function uses a novel algorithm to measure relationships among the annotation terms based on the degrees of their co-association genes to group the similar, redundant and heterogeneous annotation contents from the same or different resources into annotation groups. This reduces the burden of associating similar redundant terms and makes the biological interpretation more focused in a group level (Figure 2). The tool also provides a look at the internal relationships among the clustered terms. The clustered format is able to give a more insightful view about the relationships of annotations compared to the traditional un-clustered term report, over which similar annotation terms may be spread among hundreds, if not thousands, of other terms. In addition, to take full advantage of the well-known KEGG and BioCarta pathways, the new DAVID Pathway Viewer, another feature of the DAVID Functional Annotation Tool, can display genes from a user’s list on pathway maps to facilitate biological interpretation in a network context.

DAVID Gene Functional Classification Tool Suite

The DAVID Gene Functional Classification Tool (http://david.abcc.ncifcrf.gov/gene2gene.jsp) is a completely new component in the DAVID Bioinformatics Resources. The tool provides a novel way to functionally analyze a large number of genes in a high-throughput fashion by classifying them into gene groups based on their annotation term co-occurrence. This is accomplished and visualized by a set of new fuzzy classification algorithms, including a kappa statistics measurement of gene–gene functional relationship, a fuzzy multi-linkage partitioning method and a fuzzy genes-terms heat map visualization, etc. (manuscript submitted, and more details at http://david.abcc.ncifcrf.gov/manuscripts/fuzzy_cluster/). The power of the tool is that it allows users to simultaneously view the rich and redundant internal relationship of functionally related genes and their annotation terms within biological modules. Investigators are able to functionally analyze their gene list in a highly related many-genes-to-many-terms network context instead of a one-term-to-many-genes or a one-gene-to-many-terms view in the typical gene-annotation enrichment analysis.

DAVID Gene ID Conversion Tool Suite

A significant number of different types of gene/protein identifiers, not mutually mapped to each other across three independent resources, NCBI, PIR and UniProt (25,26,28), are now maximally integrated in the DAVID Knowledgebase (Figure 1, more details at http://david.abcc.ncifcrf.gov/helps/knowledgebase/DAVID_gene.html), whose scope is more expansive than one system only. Even though the DAVID Knowledgebase is used primarily for improvement of annotation terms integration and coverage, such comprehensive gene identifier coverage and cross-referencing capability could itself be very useful for researchers to convert their gene/protein identifiers from one type to another among over 20 major types of identifier systems (Table 1). Thus, with the newly introduced DAVID Gene ID Conversion Tool (http://david.abcc.ncifcrf.gov/conversion.jsp), interesting genes derived from one identifier system can be quickly translated to other gene identifier types preferred by a given annotation resource. In addition, the DAVID Gene ID Conversion Tool provides a ‘not sure’ type for ambiguous gene identifiers, whereby the tool can systematically suggest the potential type(s). For instance, a user has a gene ID ‘3558’ without ID type information. DAVID Gene ID Conversion Tool will scan all possibilities across all gene ID systems collected in the DAVID Knowledgebase. Two choices will be suggested, i.e. ‘3558’ could be an Entrez Gene ID for IL2 (human) or a Genbank ID for CNA1 (yeast). Thus, the user can make a decision based on above information.

DAVID Gene Name Batch Viewer

After obtaining a list of interesting genes, probably the first question researchers will ask is ‘What are the names of my genes?’ Even though it is a simple question,
most high-throughput annotation tools do not answer it in a straightforward way. The new DAVID Gene Name Batch Viewer is designed to simply list the gene names for all given genes. In addition, hyperlinks are provided on each gene entry, allowing users to explore in depth other functional information around the gene. Thus, this tool provides users with a first glance and initial ideas about their interesting genes before proceeding to analysis by other more comprehensive analytic tool. Moreover, hyperlinks, labeled as ‘RT’, are provided for each gene in order to search other functionally related genes in user’s gene list or the entire genome. The search is based on co-occurrence of annotations between genes (more details at http://david.abcc.ncifcrf.gov/helps/linear_search.html).

**DAVID NIAID Pathogen Browser**

The National Institute of Allergy and Infectious Diseases (NIAID) has defined category A, B and C priority pathogens (http://www3.niaid.nih.gov/Biodefense/bande_priority.htm), which have subsequently become important in biodefense research funding, attracting broad interest from the research community. Since the organisms listed in these categories may not be familiar to researchers who
have recently joined the emerging field, the DAVID NIAID Pathogen Browser (http://david.abcc.ncifcrf.gov/GB.jsp) is provided as a quick starting point for them to search the most relevant genes in the organisms by biological key words of interests. A large list of genes retrieved from the search could be further transferred to the DAVID Bioinformatics Resources for in-depth functional analysis with any of the previously mentioned tools. Although the tool is still in its early stage, it may help researchers gain understanding of the genes related to a priority pathogen of interest. More development is ongoing to extend the searching scope to all available genomes and annotations collected in DAVID knowledgebase.

DAVID API Services

DAVID API services (http://david.abcc.ncifcrf.gov/api/) are newly added features that allow users to directly pass gene list to various DAVID tools via a set of predefined URLs instead of DAVID submission forms. Thus, DAVID tools can easily serve as part of the analytic pipeline in other bioinformatics web sites. They can also be used in bioinformatics scripts to automate functional annotation for large number of gene lists, which are too many to be accomplished by the manual procedures.

CONCLUSION

The newly released DAVID Bioinformatics Resources are an expanded version of the original DAVID. It provides a set of powerful, novel tools that researchers can use to explore their large gene lists in depth from many different biological angles (Figure 3) in order to extract associated biological meanings to the greatest extent possible. The advanced data collection in the DAVID Knowledgebase not only creates a solid annotation data foundation for the various DAVID analytic tools, but also is freely available to the public in a simple pair-wise text format to promote the development of novel annotation algorithms and techniques within the scientific community. The DAVID Bioinformatics Resources are accessible at http://david.niaid.nih.gov.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to the referees and editors for their constructive comments. Thanks goes to Melaku Gedil, Ping Ren, and Jun Yang in the LIB group for biological discussions. We also thank Bill Wilton and Mike Tartakovsky for information technology and
network support. This research was supported in whole by the National Institute of Allergy and Infectious Disease. This project has been funded in whole with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01- CO-12400. The content of this publication does 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 U.S. Government. Funding to pay the Open Access publication charges for this article was provided by the same source as above.

Conflict of interest statement. None declared.

REFERENCES


APPENDIX: URLs TO ACCESS MAJOR COMPONENTS IN DAVID

DAVID Home Page: http://david.abcc.ncifcrf.gov
DAVID Knowledgebase Download: http://david.abcc.ncifcrf.gov/knowledgebase
DAVID Functional Annotation Tool Suite: http://david.abcc.ncifcrf.gov/summary.jsp
DAVID Gene ID Conversion Tool: http://david.abcc.ncifcrf.gov/gene2gene.jsp
DAVID Gene Name Batch Viewer: http://david.abcc.ncifcrf.gov/list.jsp
DAVID NIAID Pathogen Browser Tool: http://david.abcc.ncifcrf.gov/GB.jsp
DAVID API Services: http://david.abcc.ncifcrf.gov/api
DAVID Forum: http://david.abcc.ncifcrf.gov/forum