Database

DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis

Brad T Sherman†1, Da Wei Huang†1, Qina Tan1, Yongjian Guo4, Stephan Bour4, David Liu3, Robert Stephens3, Michael W Baseler5, H Clifford Lane2 and Richard A Lempicki*1

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

Email: Brad T Sherman - bsherman@mail.nih.gov; Da Wei Huang - huangdawei@mail.nih.gov; Qina Tan - tanq@mail.nih.gov; Yongjian Guo - guoyo@niaid.nih.gov; Stephan Bour - sbou@niaid.nih.gov; David Liu - dliu@ncifcrf.gov; Robert Stephens - bobs@ncifcrf.gov; Michael W Baseler - mbaseler@mail.nih.gov; H Clifford Lane - clane@niaid.nih.gov; Richard A Lempicki* - rlempicki@niaid.nih.gov

* Corresponding author    †Equal contributors

Abstract

Background: Due to the complex and distributed nature of biological research, our current biological knowledge is spread over many redundant annotation databases maintained by many independent groups. Analysts usually need to visit many of these bioinformatics databases in order to integrate comprehensive annotation information for their genes, which becomes one of the bottlenecks, particularly for the analytic task associated with a large gene list. Thus, a highly centralized and ready-to-use gene-annotation knowledgebase is in demand for high throughput gene functional analysis.

Description: The DAVID Knowledgebase is 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 into DAVID gene clusters. The grouping of such identifiers improves the cross-reference capability, particularly across NCBI and UniProt systems, enabling more than 40 publicly available functional annotation sources to be comprehensively integrated and centralized by the DAVID gene clusters. The simple, pair-wise, text format files which make up the DAVID Knowledgebase are freely downloadable for various data analysis uses. In addition, a well organized web interface allows users to query different types of heterogeneous annotations in a high-throughput manner.

Conclusion: The DAVID Knowledgebase is designed to facilitate high throughput gene functional analysis. For a given gene list, it not only provides the quick accessibility to a wide range of heterogeneous annotation data in a centralized location, but also enriches the level of biological information for an individual gene. Moreover, the entire DAVID Knowledgebase is freely downloadable or searchable at http://david.abcc.ncifcrf.govknowledgebase/.
Background
In the post-genomic era, one of the challenges is to systematically and comprehensively interpret large amounts of data results from experiments with a genome-wide scope, such as gene lists derived from microarray or proteomics studies. Using the biological knowledge accumulated in the past decades and the aid of computing algorithms, it is possible to assemble potential biological pictures associated with these studies. Due to the complex and distributed nature of biological research, our current knowledge is spread over many redundant databases maintained by independent groups. One gene could have different identifiers within one, or many, databases. Similarly, the biological terms associated with different gene identifiers for the same gene could be collected in different levels across different databases. Thus, an integrated gene-annotation database with comprehensive data coverage is essential as the first step of any high-throughput gene functional analytic algorithm. Some integrated databases, such as NCBI Entrez Gene [1], UniProt [2], PIR [3], etc., made great efforts to integrate annotation resources in one centralized location and are considered to be the world-class bioinformatics foundation for general bioinformatics purposes. A couple of other projects, e.g. SOURCE [4], RESOURCER [5], IDconverter [6], BioMart (formerly EnsMart) [7], UCSC Gene Sorter [8], were developed towards being more suitable for high-throughput gene-annotation queries. However, some areas are still needed for further developments in order to better meet the requirements of the high throughput gene analysis: 1) Many types of annotations are not included. e.g. Panther and BioCarta Pathways are not covered in any of above works. 2) The partial cross-reference between NCBI and UniProt systems limits integration capability. e.g. Entrez Gene does not cover PIR ID or Affy ID at all. 3) The resulting format could be better suitable for high-throughput data analysis of multiple genes. 4) The web query is performed on one gene at a time or in a small batch mode. e.g. only 100 gene at-at-time in Entrez Gene. 5) The database download is too large and complicated for regular users. e.g. Entrez Gene is in the range of tens of gigabytes in size and is comprised of a complicated, xml-like structure. 6) All data for a given database is not always available. e.g. SOURCE does not offer downloads. Due to the above limitations, the scope of most high-throughput functional annotation algorithms or data analyses is limited to a small subset of the many annotation resources and ID systems available, which does not maximize the potential analytic power. For example, the gene-annotation enrichment analytic tools, e.g. GOMiner [9], ermine [10], GOStat [11], etc., only use the GO database [12] as a backend annotation source and only NCBI Entrez Gene as a gene ID mapping source. Gene IDs and annotation contents derived from Uniprot are weaker or not acceptable at all in these packages. In addition, each of the tools requires a large amount of redundant efforts to build its own backend database from public resources.

The goal of this work is to create a large gene-centered knowledgebase that integrates the most useful and highly regarded heterogeneous annotation resources in a centralized location with improved cross-referencing capability between NCBI and UniProt systems [1,2], and easy to use pair-wise data structure files for downloads, hence, more comprehensive and suitable for high throughput data analysis. The work was originally conducted years ago to successfully serve as a comprehensive backend knowledgebase for various high throughput gene-annotation enrichment analytic tools in the DAVID and EASE packages [13,14]. The usefulness of the DAVID knowledgebase in our own bioinformatics software motivates us to make it available to the public community in order to benefit the high throughput data analysis projects in other research groups. Now, the entire DAVID Knowledgebase [15] is either freely downloadable or searchable through the DAVID Bioinformatics Resources web site [16].

The paper will describe the DAVID Knowledgebase regarding its unique strategy to integrate the redundant and heterogeneous annotation sources, the improved cross-reference capability across gene ID types, the large annotation content coverage, and the pair-wise text-format files for downloads and easy-to-use web-based query interface.

Construction and content
Most gene functional annotation databases are in a gene-annotation association format, i.e. annotation contents usually associate with corresponding gene or protein identifiers. Such a format provides an opportunity to integrate heterogeneous annotation resources through their common gene identifiers. The construction of the DAVID Knowledgebase consists of two major steps: 1) Improve cross-reference capability across redundant gene/protein IDs, particularly for the IDs found in the NCBI and UniProt systems [1,2], with a novel single-linkage algorithm called the DAVID Gene Concept. As a result, unique DAVID gene clusters are formed to hold the redundant gene/protein IDs belonging to the same gene entries (Figure 1B). 2) Assign heterogeneous annotation contents from different annotation databases, which could be associated with different types of gene/protein IDs (but belonging to the same gene), to the same DAVID gene cluster (Figure 2).

**DAVID gene concept: a novel single-linkage algorithm to agglomerate redundant gene IDs into the DAVID gene clusters in order to improve cross-referencing capability**

There are dozens of types of gene or protein sequence identifiers that are redundant within the same group or
across several independent groups, such as GenBank Accession; GenBank ID; RefSeq Accession; PIR ID; PIR Accession; UniProt ID; UniProt Accession; etc. [1-3,17-19]. The leading organizations, NCBI [1] and UniProt [2], have made significant strides in addressing the cross-reference and redundancy issues associated with gene identifiers. NCBI GenBank, representing the largest redundant database of nucleotide sequences, exchanges data with two other worldwide nucleotide sequence databases, EMBL and DDBJ. In addition, UniProt, as the largest redundant annotated protein sequence database, unites Swiss-Prot, TrEMBL, and PIR. Moreover, the three organizations have been independently constructing non-redundant gene cluster databases, NCBI Entrez Gene [1], UniProt UniRef [2], and PIR-NREF [3], respectively. The resulting databases are presented in a non-redundant format by grouping the different gene/protein IDs for the same gene into one entry. At this point, the redundant nucleotide and protein IDs from different resources have been largely addressed by the leading bioinformatics organizations. However, while the gene clusters are comprehensive for the gene/protein IDs within their own organization, many cannot be cross-referenced with gene identifiers from other independent organizations (Figure [1]).

**Figure 1**
The cross-reference among different types of gene identifiers is improved by the DAVID Gene Concept. A. As global examples, four popular types of protein identifiers (PIR ID, UniProt Accession, RefSeq Protein, and GenPept Accession) are only cross-referenced partially by NCBI Entrez Gene (EG), UniProt UniRef100 (UP), and PIR NRef100 (NF). B. The DAVID Gene Concept, a single-linkage algorithm, iteratively agglomerates all types of gene IDs, belonging to the same gene entry, for example, *ubiquitin specific peptidase 8* (USP8), into one DAVID gene cluster (ID 2859041). Such integration makes the cross-reference of different types of gene IDs more comprehensive than that in each of the original databases, particularly for the IDs across NCBI and UniProt systems.
For example, UniProt does not cover RefSeq IDs; NCBI Entrez Gene does not reference PIR ID at all (Figure 1A). Therefore, the major challenge of annotation integration comes from the weak cross-reference of different types of gene/protein IDs between NCBI and UniProt systems since different annotation databases use one or another system as their major gene identifier systems, e.g. GeneRif [20] adopts NCBI IDs as major associated identifiers; InterPro [21] uses UniProt/SwissProt as major associated identifiers.

To further improve the cross-reference capability among different types of gene/protein identifiers, DAVID gene clusters, as secondary gene clusters, are created by merging the existing gene clusters from three major gene cluster databases, Entrez Gene [1], UniRef100 [19], and PIR-NRRef100 [3], with a single-linkage algorithm (Figure 1B). Any gene clusters from the above three resources with one or more protein IDs in common and from the same species will be considered as the same gene entry. The overlapping gene clusters are iteratively merged into a new gene cluster based on the single-linkage rule until all final gene clusters, or DAVID genes, are stable (Figure 1B). A unique integer number is assigned to each of the newly formed DAVID genes and is used as a centralized gene identifier/index within the DAVID Knowledgebase (Figure 1B). Importantly, the DAVID gene ID is a gene level clustering ID so that any different isoforms or splicing var-

![Diagram](https://example.com/diagram.png)
iants belonging to the same gene will be assigned to the same DAVID ID. The process collects >60 million individual gene/protein identifiers representing over 20 identifier types (Table 1), which are eventually agglomerated into over 3.7 million DAVID genes, for over 90,000 species. The DAVID genes greatly improve the cross-referencing capability for the IDs found in the NCBI and UniProt systems.

Because a DAVID gene is built based on annotated gene clusters, only well-known or studied gene identifiers in the original gene clusters are included. This scope is well aligned with the high throughput functional annotation purpose of the DAVID Knowledgebase in the sense that any unclear or unstudied sequences, such as an EST, are not helpful for automatic analysis of high-throughput functional annotation.

DAVID genes are secondary gene clusters built on well-known and annotated gene clusters from NCBI Entrez Gene, PIR NRef100, and UniProt UniRef100. Thus, the agglomeration quality of a DAVID gene solely relies on the quality of the original databases. Since the original databases have been used by the scientific community for many years, they are well known and regarded as the highest-quality bioinformatics resources in the world. To further detect potential problems inherited from the original sources into DAVID gene clusters, a comprehensive quality control (QC) procedure was conducted by examining the sequence alignment of every protein member within a given DAVID gene cluster using the NCBI BlastClust program [22,23] (Additional File 1 for detailed procedure of QC). The QC examination highlighted poor sequence alignment in ~10% of the DAVID genes, mainly caused by very short sequences (less than 20 amino acids), which were not handled well, or at all, by the BlastClust program. After filtering out those short sequences, less than 0.1% of the DAVID genes with poor alignment members needed to be corrected. The QC procedure reflects the high quality of the original resources, which is passed on to the DAVID Knowledgebase.

Collection and integration of functional annotation contents: the heterogeneous annotation contents and

| Gene Identifiers (> 60 millions) | Annotation Contents (> 90 millions in total) |
|---------------------------------|-----------------------------------------------|
| AFFY_ID                         | Ontology (>40 million records)                |
| ENTREZ_GENE_ID                  | GO_BIOLOGICAL_PROCESS                         |
| GENPEPT_ACCESSION               | GO_MOLECULAR_FUNCTION                        |
| GENBANK_ACCESSION               | GO_CELLULAR_COMPONENT                        |
| GI                              | PANTHER_BIOLOGICAL_PROCESS                   |
| PIR_ACCESSION                   | PANTHER_MOLECULAR_FUNCTION                   |
| PIR_ID                          | COG_KOG_ONTOLOGY                             |
| PIR_NREF_ID                     | P-P Interaction (> 4 millions)               |
| REFSSEQ_GENOMIC                | BIND                                          |
| REFSSEQ_MRNA                   | DIP                                           |
| REFSSEQ_PROTEIN                | MINT                                          |
| REFSSEQ_RNA                    | NCICB_CAPPATHWAY                             |
| UNIGENE                        | TRANSFAC_ID                                   |
| UNIPROT_ACCESSION              | HIV_INTERACTION                              |
| UNIPROT_ID                     | HIV_INTERACTION_CATEGORY                     |
| UNIRF100_ID                    | HPDR_INTERACTION                             |
| OFFICIAL_GENE_SYMBOL           | REACTOME_INTERACTION                         |
| ENSSEMBLE_ID                   | Disease Association (~9,000)                  |
| FLYBASE_ID                     | GENETIC_ASSOCIATION_DB                       |
| HAMAP_ID                       | OMIM_PHENOTYPE                                |
| HSSP_ID                        | Literature (>2.8 millions)                   |
| TIGR_ID                        | GENERIF_SUMMARY                              |
| WORMBASE_ID                    | PUBMED_ID                                     |
| RGD_ID                         | HIV_INTERACTION_PUBMED_ID                    |
| UNIPROT_ACCESSION              |                                               |

Table 1: Data coverage in the DAVID Knowledgebase.

Comprehensive collection and integration of different types of gene identifiers and heterogeneous annotation categories. Importantly, any of the gene identifier types and associated annotation categories above are comprehensively cross-mapped to each other in the DAVID Knowledgebase by the unique DAVID Gene Concept procedure.
their IDs from different annotation databases are assigned to and centralized by the common DAVID genes

NCBI Entrez Gene, PIR, and UniProt databases [1-3] also collect annotation contents associated with corresponding gene IDs. But the integration is dependent on the gene ID system, i.e. one database may not fully integrate some annotation contents that are associated with gene identifiers not well cross-referenced within its system. For instance, InterPro associated with SwissProt/UniProt IDs may not be comprehensively integrated into the NCBI system. However, the situation is improved with the larger collection and higher integration of different types of gene identifiers within the DAVID genes (Table 2). Since all major types of gene identifiers eventually can be translated to a corresponding DAVID gene, the heterogeneous annotation contents, as long as they are in a gene-annotation association format, have a much better chance of being integrated by the common DAVID gene, hence improving the integration of annotation contents for individual gene (Table 2), as well as for entire genome as a whole (Figure 3). The DAVID Knowledgebase collects a wide range of well-known and high-quality annotation contents from dozens of databases including: Gene Ontology; Protein Domains; Bio-pathways; Gene Expression; Disease Association; PubMed; Protein-Protein interactions; Affymetrix; Gene General Features; NCI Thesaurus; Panther Family; and more (Table 1) (also see Additional File 2 for the annotation coverage comparison across databases). Furthermore, the structure of the DAVID Knowledgebase is open to any new annotation database as long as it is in a gene-annotation association format (Figure 2).

Inaccurate and conflicting annotation has been considered due to its potential negative impact on the knowledgebase. This impact is taken into consideration and corrected during DAVID cluster creation within the DAVID Quality control pipeline as stated previously. While the incorrect annotation is not used in the creation of the DAVID clusters, thereby stopping any magnification of error, the annotation is still maintained in the knowledgebase. Due to the significant amount of data available within the knowledgebase, including redundant and complimentary annotation, the inaccurate data becomes highly diluted. Considering that the DAVID Knowledgebase is intended for high throughput gene functional analysis, when many biology aspects are considered together in either high-throughput analysis or for any one gene, the negative impact of any errors is negligible since the true biology is generally overwhelming and supporting from the various sources. This may not be the case with each individual database if there is not a majority of supporting evidence given for the true biology or a systematic error has occurred with the source's annotation process for a given gene.

Considering the DAVID Knowledgebase is designed for high throughput gene functional analysis, the larger collection and integration of heterogeneous annotation sources and the quick accessibility to the larger amount of data are more important than the timely update, simply because the high throughput gene functional analysis relies on the global annotation profiles rather than an individual annotation source. Through automation of several tasks where appropriate and personnel additions, the goal of the DAVID knowledgebase update is set to occur quarterly. A complete list of the public databases contributing to the DAVID Knowledgebase are provided within Additional File 3 while a detailed update procedure can be found in Additional File 4.

Utility and discussion

The data structure of the entire knowledgebase for downloads

The DAVID Knowledgebase is available in two categories of pair-wise text files: gene index files (gene id knowledge) and annotation index files (annotation knowledge) [15]. Gene index files with a naming convention such as david2affy_id, david2genbank_accession, etc., are in a pair-wise format, linking a DAVID gene identifier to a public gene identifier (Figure 2). These relationships were built based on the results from the DAVID gene agglomeration step described in the "Construction and Content" section. Thus, any given public gene identifier can be converted to a corresponding DAVID gene identifier that represents a unique gene entry and internal linker to all available annotation contents within the DAVID Knowledgebase. Annotation index files with a naming conven-
tion such as david2pfam, david2kegg_pathway, david2omim, etc., are also in a format that pairs a DAVID gene identifier with an annotation term. All genes and annotation terms within the DAVID Knowledgebase are centralized by the common DAVID gene identifier. The unified DAVID gene identifier not only normalizes the cross-reference among heterogeneous databases, but also makes any search and calculation simpler and more efficient. Therefore, for any given public gene identifier, the corresponding DAVID gene identifier can be obtained with the DAVID gene index files. Then, any annotation terms within the DAVID Knowledgebase can be further queried from the DAVID annotation index files using the DAVID identifier (Figure 2B). In addition, each independent annotation resource and public gene identifier system is separated into independent files, making it easier for users to focus on the data they are most interested in, as well as reducing the file size and simplifying the format for easy processing. The easily interpreted file names, such as david2genbank_accession and david2pfam, allow users to quickly identify the data that will help them interpret their data. Of course, users can further combine their data files of interest into one large file or database table to search the data in a way best suited to the individual. The simple pair-wise text format provides the flexibility to either directly query the files or insert them into tables in a customized, in-house relational database without much, if any, file parsing or re-formatting. The format should be simple to regular users with some computational skills as well as extendable for expert users. Moreover, users may add any new annotation sources to the DAVID Knowledgebase as additional independent files, as long as the files are in a gene-annotation associated format. The DAVID Knowledgebase, with its large, diverse annotation categories and flexible format, provides the scientific community with a single, comprehensive platform for gathering annotation for specific studies.

### Web interface for batch query

In addition to the database download for genome-wide large scale analysis, the DAVID Knowledgebase can be queried in a gene-centric way (i.e. from gene to annotation) for small (e.g. a couple of hundreds) to medium size (e.g. a couple of thousands) of gene lists through an easy-to-use web interface [24] (Figure 4). The annotation contents from different databases on the web site interface are organized into 10 groups such as Pathways, Protein-Protein Interaction, Gene Ontology, etc. which organization is consistent with that in the entire database downloads (Table 1). Flexible options are provided in order for users to define the query scope of the annotation categories which relate to their interests. After a gene list is submitted, the web interface efficiently queries the chosen annotation contents for gene/protein IDs in the given list.

### Table 2: An example of improved annotation coverage for an individual gene USP8 in the DAVID Knowledgebase.

| Category (Resource)                  | Selected Annotation                                      | Entrez Gene | UniProt | PIR | iProClass | DAVID |
|-------------------------------------|----------------------------------------------------------|-------------|---------|-----|----------|-------|
| Molecular Function (Gene Ontology)  | GO:0004197: cysteine-type endopeptidase activity          | √           | √       |     |           | √     |
|                                     | GO:0005515: protein binding                               | √           |         |     |           | √     |
|                                     | GO:0004843: ubiquitin-specific protease activity          | √           | √       |     |           | √     |
|                                     | GO:0008233: peptidase activity                            | √           |         |     |           | √     |
|                                     | GO:0004221: ubiquitin thiolesterase activity              | √           | √       |     |           | √     |
|                                     | GO:0016787: hydrolase activity                            | √           |         |     |           | √     |
|                                     | GO:0008234: cysteine-type peptidase activity              | √           |         |     |           | √     |
| Biological Process (Gene Ontology)  | GO:0006512: ubiquitin cycle                               | √           |         |     |           | √     |
|                                     | GO:0008283: cell proliferation                            | √           |         |     |           | √     |
|                                     | GO:0006511: ubiquitin-dependent protein catabolic process | √           |         |     |           | √     |
|                                     | GO:0007265: Ras protein signal transduction               | √           |         |     |           | √     |
| Cellular Component (Gene Ontology)  | GO:0005622: intracellular                                  | √           |         |     |           | √     |
| Protein Domain (InterPro)           | IPR001394: Peptidase_C19                                  | √           |         |     |           |       |
|                                     | IPR001763: Rhodanese-like                                 | √           |         |     |           | √     |
| Enzyme (EC)                        | EC 3.1.2.15: Ubiquitin thiolesterase                      | √           |         |     |           | √     |
| Protein 3-D Structure(PDB)          | 1WHB: Solution structure of the Rhodanese-like domain in human UBP8 | √           |         |     |           | √     |
|                                     | 2A9U: Structure of the N-terminal domain of Human USP8   | √           |         |     |           | √     |
| Disease Association (OMIM)          | OMIM: 603158: chronic myeloproliferative disorder         | √           |         |     |           | √     |

For the gene of USP8, the annotation coverage is partially overlapped among the independent resources of Entrez Gene, UniProt and PIR. After the integration around DAVID Gene Concept (Figure 1B), the DAVID Knowledgebase provides more comprehensive coverage of annotation in one centralized location for USP8 than that in each of the individual sources.
The query results for all genes are organized in one single page/file, which is more suitable for high throughput data analysis. Then, users may either further explore the annotation contents on the html report or download the gene-annotation results in a flat text file. In addition, for light duty jobs, a web-based API service [25] is also available for users to query the DAVID Knowledgebase through a URL link on a third party web site or an automatic script.

**A query comparison between DAVID Knowledgebase and Entrez Gene**

Due to the comprehensive integration procedure, the DAVID Knowledgebase provides not only quick access to a wide range of annotation contents in one location but also enriched annotation contents and improved ID cross-referencing capability. To the best of our knowledge, these features exceed that of other similar integrated sources. A list of 400 Affymetrix IDs (see Additional File 6) derived from a HIV microarray study (labeled as demolist 2 on the DAVID web site) [26] is used to query the DAVID Knowledgebase and NCBI Entrez Gene Database respectively in order to obtain associated biological annotations. The first issue arises due to the fact that the Entrez Gene database does not recognize Affymetrix IDs. Thus, we have to convert the Affymetrix IDs to their corresponding Entrez Gene IDs in order to take advantage of the annotation contents in the Entrez Gene database. In contrast, DAVID can directly recognize Affymetrix IDs and most other types of IDs. Secondly, since the Entrez Gene database is designed for general bioinformatics purposes (maximally 100 gene IDs at-a-time), the 400 IDs have to be split into a couple of smaller batches in order to perform the queries. Then, the results have to be manually merged and combined. Conversely, DAVID Knowledgebase supports high throughput access, and the results are put into one single file. Thirdly, the heterogeneous annotation contents in Entrez Gene Report appear in multiple rows, in many big blocks, across multiple genes. Conversely, the annotation contents in the DAVID download page are organized into one row for one gene where annotations are separated by tab delimitation. This organization can immediately be entered into MS Excel for further processing. Finally, the most important advantage of the DAVID Knowledgebase is that the annotation categories are much extended and the annotation contents are enriched. For the example Affymetrix list, 10–20% more GO terms are enriched in the DAVID Knowledgebase compared to each of the individual resources (e.g. Entrez Gene) (Figure 5). A similar conclusion was obtained when we conducted comparisons to SOURCE [4], RESOURCER [5], IDconverter [6], BioMart [7], and UCSC Gene Sorter [8] with the same gene list [26] (see Additional File 7 for detailed results).

**Key advantages of the DAVID Knowledgebase**

The DAVID Knowledgebase intends to integrate and organize the high quality, world-class bioinformatics databases into a centralized location in a gene-centric format. The work is particularly useful for high throughput data analysis with the following advantages: 1) Improved ID cross-referencing capability enhances comprehensiveness of integration of heterogeneous annotation resources, hence enriching the annotation coverage for individual genes. 2) The integration allows quick access to a wide range of annotation contents in a batch manner. 3) Simple pair-wise, gene-centric formatted files simplify the process of data analysis.
data structure so that all users may benefit. 4) The pair-wise data structure is more flexible and suitable for high throughput data access.

**Special attention on the DAVID Knowledgebase**

After several QC procedures, a certain error rate still exists in the DAVID Knowledgebase as it exists in any other bio-databases. Users may report such errors directly to us through email or the DAVID forum [27]. In addition, users should be aware that the DAVID Knowledgebase is designed for high throughput gene functional screening for large gene lists on a gene-centric level, rather than replacing original annotation databases, which may contain additional details, such as gene isofom specific annotations, for drill-down analysis. Moreover, the quarterly update schedule of the DAVID Knowledgebase could result in a slight time gap as compared with the member database updates.

**Conclusion**

The DAVID Gene Concept agglomerates diverse types of gene identifiers belonging to the same gene into one gene cluster. It allows large collections of heterogeneous annotations that are associated with different types of gene identifiers to be comprehensively integrated by a common DAVID gene. Combined with the simple pair-wise text format, the DAVID Knowledgebase provides not only a comprehensive, high-quality collection of gene annotation resources, but also the flexibility to cross-reference identifiers and annotations from several world-class, heterogeneous databases within one resource. To the best of our knowledge, the annotation data coverage and gene/protein ID cross-referencing capability far exceeds that of backend data sources of other high throughput gene functional annotation tools. Therefore, it can be used as the backend gene-annotation database of existing high throughput gene functional analysis tools to improve their discovery power. The DAVID Knowledgebase also aids the researcher in focusing on data analysis or the core development of new high-throughput functional data-mining algorithms, rather than spending time on gene-annotation data collection and integration.

**Availability and requirements**

The DAVID Knowledgebase is freely downloadable for nonprofit use under the URL [http://david.abcc.ncifcrf.gov/knowledgebase](http://david.abcc.ncifcrf.gov/knowledgebase). Data files are available in a tab-delimited text format which can be opened by any text editor in Windows, Mac or Unix systems.

**Abbreviations**

DAVID: Database for Annotation, Visualization and Integrated Discovery

PIR: Protein Information Resource

NCBI: National Center for Biotechnology Information

BLAST: Basic Local Alignment Search Tool

EMBL: European Molecular Biology Laboratory

DDBJ: DNA Data Bank of Japan

RefSeq: Reference Sequence

GO: Gene Ontology

SMART: Simple Modular Architecture Research Tool

DID: DAVID Identifier

EG: Entrez Gene

UP: UniProt-UniRef

NF: PIR-NRef

OMIM: Online Mendelian Inheritance in Man

Pfam: Protein Family

LIB: Laboratory of Immunopathogenesis and Bioinformatics

**Authors’ contributions**

DWH oversaw the project and wrote the manuscript; BS developed the majority of the JAVA programs used to cluster the DAVID genes, and to parse and integrate annotation data; built the database architecture; and revised the manuscript. DWH and QT developed some programs contributing to additional annotation data or web interface; YG and SB developed the QC pipeline for the DAVID gene. JK developed some Perl scripts to add additional annotation data. DL and RS supported the Oracle database. RL and CL are principle investigators, who supervised the project design and implementation, and the manuscript preparation. All authors read and approved the final manuscript.

**Additional material**

**Additional file 1**

*Flow chart of the procedure for DAVID gene QC with NCBI BlastClust program.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-426-S1.doc](http://www.biomedcentral.com/content/supplementary/1471-2105-8-426-S1.doc)
Acknowledgements

The authors are grateful to the referees for their constructive comments and thank David Bryant in the ABC group for web server support. Thanks also go to Melaku Gedil, Ping Ren, Jun Yang in the Laboratory of Immunopathogenesis and Bioinformatics (LIB) group for biological 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. NOI-CO-56000. The contents 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. Maglott D, Ostell J, Pruitt KD, Tatusova T: Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 2005, 33(Database issue):D58-D64.
2. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O’Donovan C, Redaschi N, Yeh LS: UniProt: the Universal Protein Knowledgebase. Nucleic Acids Res 2004, 32(Database issue):D115-9.
3. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtessian P, Ledley RS, Suzek BE, Vinyaray CR, Zhang J, Barker WC: The Protein Information Resource. Nucleic Acids Res 2003, 31(1):345-347.
4. Diehn M, Sherlock G, Binkley G, Jin H, Matice JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botsstein D, Brown PO, Alizadeh AA: SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Res 2003, 31(1):19-223.
5. Tsai J, Sultana R, Lee Y, Pertea G, Karamycheva S, Antonescu V, Cho J, Parvizi B, Cheung F, Quackenbush J: RESOURCER: a database for annotating and linking microarray resources within and across species. Genome Biol 2001, 2(11):SOFTWARE0002.
6. Alibes A, Yankilevich P, Canada A, Diaz-Uriarte R: IDClight: conversion and annotation of gene and protein IDs. BMC Bioinformatics 2007, 8:9.
7. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E: Ensembl: a generic system for fast and flexible access to biological data. Genome Res 2004, 14(1):160-169.
8. Kent WJ, Hsu F, Karolchik D, Kuhn RM, Clawson H, Trumbower H, Haussler D: Exploring relationships and mining data with the UCSC Gene Sorter. Genome Res 2005, 15(5):737-741.
9. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN: Gominer: a resource for biological interpretation of genomic and proteomic data. Genome Biol 2003, 4(4):R28.
10. Lee HK, Braynen W, Keshav K, Pavlidis P, Erminje: tool for functional analysis of gene expression data sets. BMC Bioinformatics 2005, 6:269.
11. Beissbarth T, Speed TP: GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 2004, 20(9):1464-1465.
12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matice JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000, 25(1):25-29.
13. Dennis G Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: DAVID: database forAnnotation, Visualization, and Integrated Discovery. Genome Biol 2003, 4(5):P3.
14. Hosack DA, Dennis G Jr., Sherman BT, Lane HC, Lempicki RA: Identifying biological themes within lists of genes with EASE. Genome Biol 2003, 4(10):R70.
15. DAVID Knowledgebase [http://david.abcc.ncifcrf.gov/knowledgebase]
16. DAVID Bioinformatics Resources [http://david.abcc.ncifcrf.gov]
17. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: GenBank. Nucleic Acids Res 2006, 34(Database issue):D16-20.
18. Pruitt KD, Tatusova T, Maglott DR: NCBI Reference Sequence (RefSeq): a curated non-duplicate sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2005, 33(Database issue):D501-4.
19. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Maragane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B: The Universal Protein Resource (Uniprot): an expanding universe of protein information. Nucleic Acids Res 2006, 34(Database issue):D187-91.

20. Mitchell JA, Aronson AR, Mark G, Folk LC, Humphrey SM, Ward JM: Gene indexing: characterization and analysis of NLM's GeneRIFs. AMIA Annu Symp Proc 2003:460-464.

21. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silvestri V, Studholme DJ, Vaughan R, Wu CH: InterPro, progress and status in 2005. Nucleic Acids Res 2005, 33(Database issue):D201-5.

22. NCBI [http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blastlab.html]

23. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol 1990, 215(3):403-410.

24. DAVID Knowledgebase Web Interface [http://david.abcc.ncifcrf.gov/knowledgebase_summary.jsp]

25. DAVID API Services [http://david.abcc.ncifcrf.gov/content.jsp?file=DAVID_API.html]

26. Cicala C, Arthos J, Selig SM, Dennis G Jr., Hosack DA, Van Ryk D, Spangler ML, Steenbeke TD, Khazanie P, Gupta N, Yang J, Daucher M, Lempicki RA, Fauci AS: HIV envelope induces a cascade of cell signals in non-proliferating target cells that favor virus replication. Proc Natl Acad Sci U S A 2002, 99(14):9380-9385.

27. DAVID Forum [http://david.abcc.ncifcrf.gov/content.jsp?file=Contact.html]