Functional Annotation Tool

Introduction

The This tool suite, 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 . 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. 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, with its improved computational power, the new tool accepts customized gene backgrounds, an option rarely found in other Web-based, high- throughput annotation tools for typical gene-annotation enrichment analysis. 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 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. The tool also provides a look at the internal relationships of the clustered terms by comparing it to the typical linear, redundant term report, over which similar annotation terms may be distributed among hundreds or 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.

top

Data Input

Please see the Universal Gene List Manager

top

Gene-Enrichment and Functional Annotation Analysis

1. A Typical Analysis Flow

Load Gene List -> View Summary Page -> Explore details through Chart Report, Table Report, Clustering Report, etc. -> Export and Save Results

2. EASE Score, a modified Fisher Exact P-Value

When members of two independent groups can fall into one of two mutually exclusive categories, Fisher Exact test is used to determine whether the proportions of those falling into each category differs by group. In DAVID annotation system, Fisher Exact is adopted to measure the gene-enrichment in annotation terms.

A Hypothetical Example:

In human genome background (30,000 gene total), 40 genes are involved in p53 signaling pathway. A given gene list has found that 3 out of 300 belong to p53 signaling pathway. Then we ask the question if 3/300 is more than random chance comparing to the human background of 40/30000.

A 2x2 contingency table is built on above numbers:

	User Genes	Genome
In Pathway	3-1	40
Not In Pathway	297	29960

Fisher Exact P-Value = 0.008 (using 3 instead of 3-1). Since P-Value <= 0.01, this user gene list is specifically associated (enriched) in p53 signaling pathway than random chance

However, EASE Score is more conservative to examine the situation. EASE Score = 0.06 (using 3-1 instead of 3). Since P-Value > 0.01, this user gene list is specifically associated (enriched) in p53 signaling pathway no more than random chance

See more discussion at DAVID Forum.

top

4. Functional Annotation Summary

top

4. Functional Annotation Chart Report

Maximum EASE Score/P-Value

Gene list and population background being analyzed

Minimum number of genes for the corresponding term

Maximum number of record per page

Functional Annotation Chart
Current Gene List: demolist1
Current Background: Homo sapiens
171 DAVID IDs

☐ Options

| Count Threshold | 2 | | EASE Threshold | 0.1 | | # of Records Displayed | 1000 |

[Rerun Using Options] [Create Sublist]   📊 Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value
☐	SP_PIR_KEYWORDS	signal	RT	▬	47	27.5%	3.0E-10
☐	SP_PIR_KEYWORDS	glycoprotein	RT	▬	51	29.8%	4.9E-8
☐	GOTERM_CC_ALL	extracellular region	RT	▬	32	18.7%	1.1E-7
☐	SP_PIR_KEYWORDS	alternative splicing	RT	▬	49	28.7%	6.4E-6
☐	SP_PIR_KEYWORDS	chromoprotein	RT	▬	7	4.1%	1.1E-5
☐	SP_PIR_KEYWORDS	direct protein sequencing	RT	▬	33	19.3%	1.2E-5
☐	SP_PIR_KEYWORDS	phosphorylation	RT	▬	31	18.1%	1.6E-5
☐	UP_SEQ_FEATURE	signal peptide	RT	▬	47	27.5%	3.7E-5
☐	SP_PIR_KEYWORDS	metalloprotein	RT	▬	8	4.7%	4.7E-5
☐	GOTERM_BP_ALL	response to chemical stimulus	RT	▬	14	8.2%	6.1E-5

Original database/resource where the terms orient

Enriched terms associated with your gene list

Related Term Search

Genes involved in the term

Percentage, e.g. 14/171=8.2% (involved genes/total genes)

Modified Fisher Exact P-Value, EASE Score. The smaller, the more enriched.
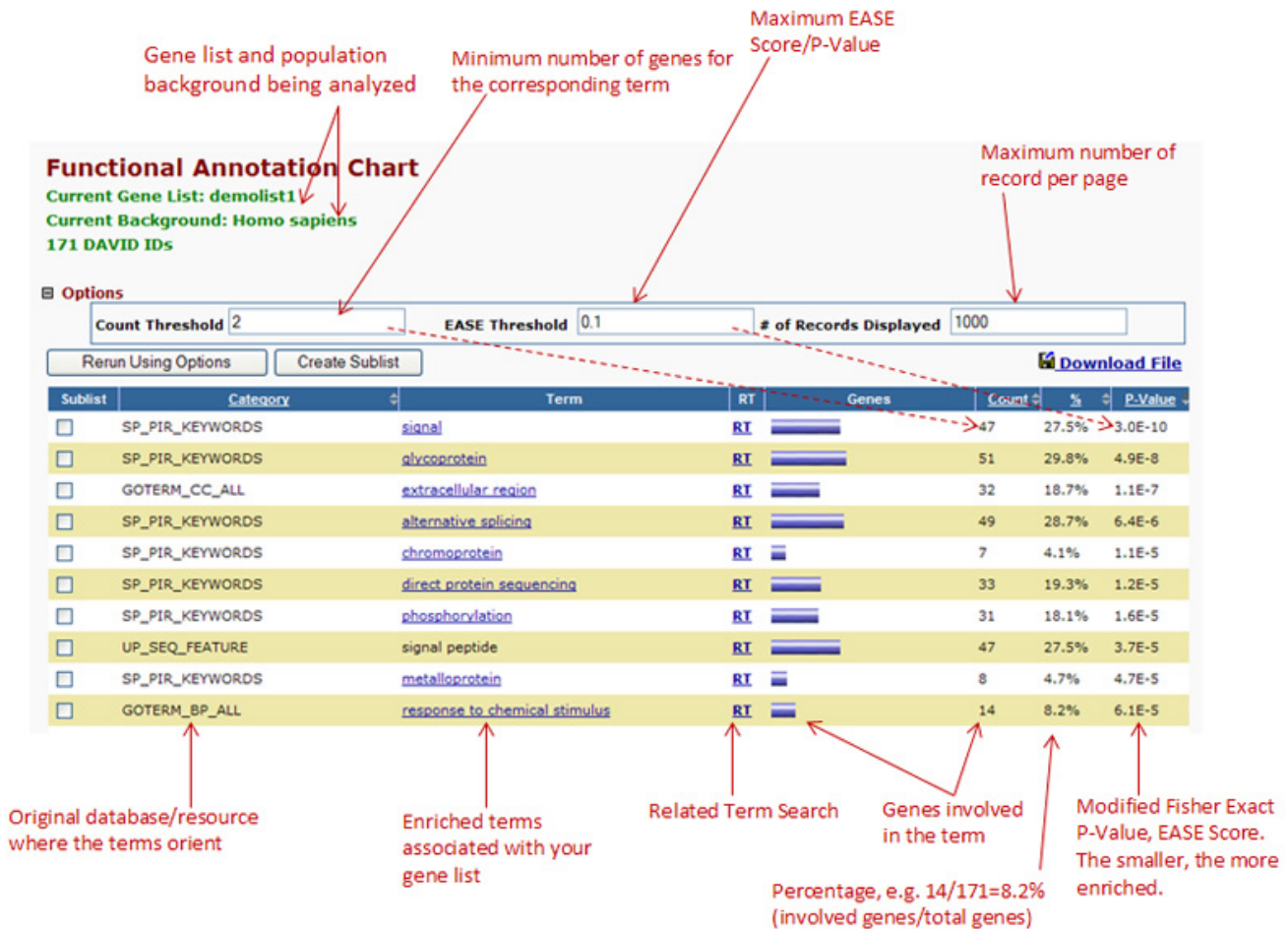
Functional Annotation Chart:
Chart Report is an annotation-term-focused view which lists annotation terms and their associated genes under study. To avoid over counting duplicated genes, the Fisher Exact statistics is calculated based on corresponding DAVID gene IDs by which all redundancies in original IDs are removed. All result of Chart Report has to pass the thresholds (by default, Max.Prob.<=0.1 and Min.Count>=2) in Chart Option section to ensure only statistically significant ones displayed.

EASE Score Threshold (Maximum Probability):
The threshold of EASE Score, a modified Fisher Exact P-Value, for gene-enrichment analysis. It ranges from 0 to 1. Fisher Exact P-Value = 0 represents perfect enrichment. Usually P-Value is equal or smaller than 0.05 to be considered strongly enriched in the annotation categories. Default is 0.1. More details.
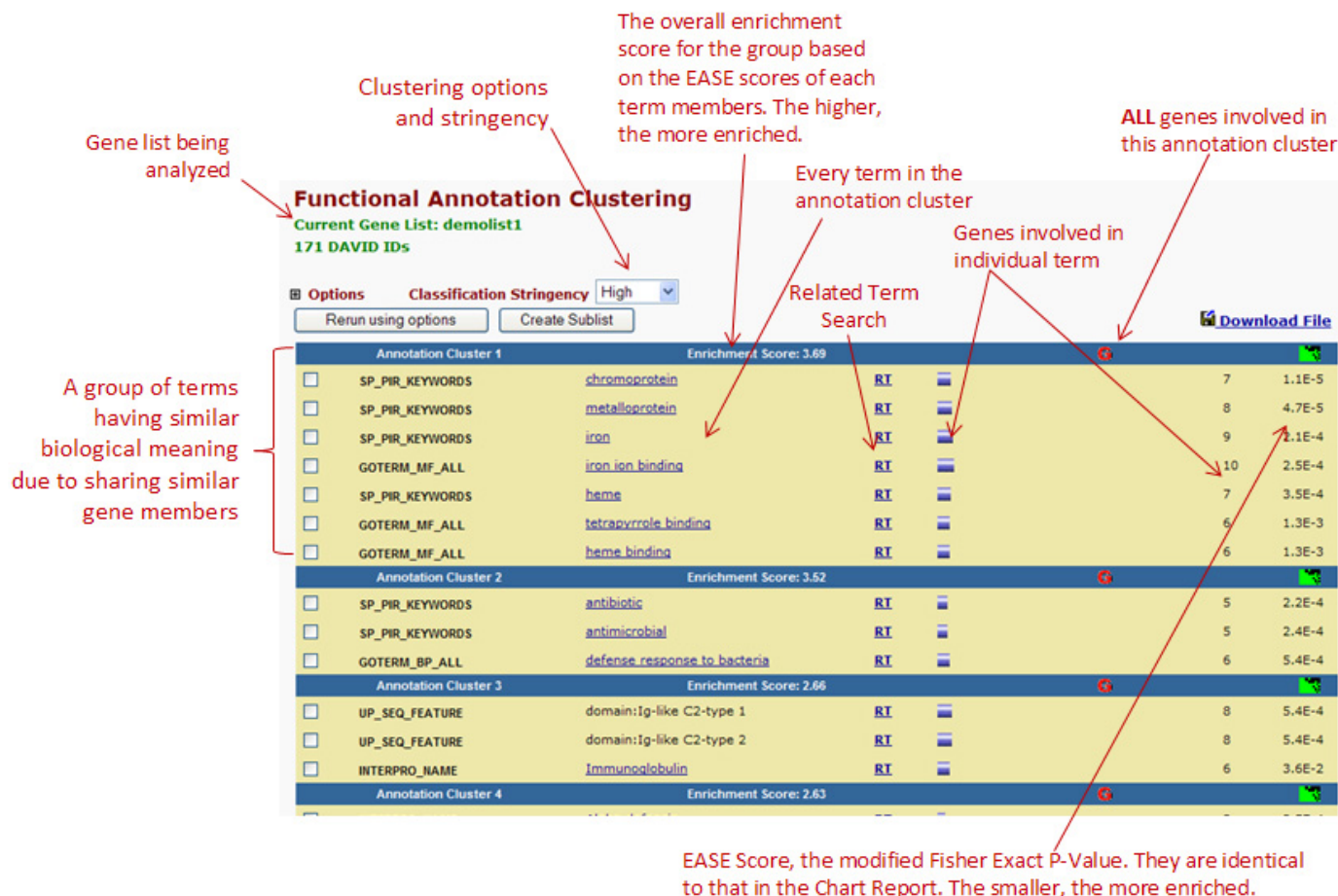
Count Threshold (Minimum Count):
The threshold of minimum gene counts belonging to an annotation term. It has to be equal or greater than 0. Default is 2. In short, you do not trust the term only having one gene involved.

RT (Related Term Search):
Related Term Search can identify other similar terms. More details.

5. Functional Annotation Clustering Report

Functional Annotation Clustering: new!
Due to the redundant nature of annotations, Functional Annotation Chart presents similar/relevant annotations repeatedly. It dilutes the focus of the biology in the report. To reduce the redundancy, the newly developed Functional Annotation Clustering report groups/displays similar annotations together which makes the biology clearer and more focused to be read vs. traditional chart report. The grouping algorithm is based on the hypothesis that similar annotations should have similar gene members. The Functional Annotation Clustering integrates the same techniques of Kappa statistics to measure the degree of the common genes between two annotations, and fuzzy heuristic clustering (used in Gene Functional Classification Tool) to classify the groups of similar annotations according kappa values. In this sense, the more common genes annotations share, the higher chance they will be grouped together.

The p-values associated with each annotation terms inside each clusters are exactly the same meaning/values as p-values (Fisher Exact/EASE Score) shown in the regular chart report for the same terms.

The Group Enrichment Score new! , the geometric mean (in -log scale) of member's p-values in a corresponding annotation cluster, is used to rank their biological significance. Thus, the top ranked annotation groups most likely have consistent lower p-values for their annotation members.

Options:
Similar idea as the options in Gene Functional Classification.

6. Other reports/views

Functional Annotation Table:

Table Report is a gene-centric view which lists the genes and their associated annotation terms (selected only). There is no statistics applied in this report.

Gene Report:

A highly integrated view of a single gene and its general annotations/accessions from multiple resources. It can quickly give a global idea about the gene. The hyperlinks throughout the report will lead to users to original resources for further details.
DAVID Pathway Mapping:

To allocate user input genes on static pathway maps generated by BioCarta and KEGG. This mapping strategy of "View Dynamic Data on Static Map Picture" is uniquely developed in DAVID package.

top

Some Terminology in DAVID System

Annotation Category:
A group of annotation sources collecting similar biological questions, such as: "Pathways" is an annotation category consisting of BioCarta, KEGG, etc.

Annotation Source:
An independent database in a category , such as: BioCarta Pathways.

Term:
A detailed item in an annotation source, such as: p53 signaling pathway in BioCarta.

hierarchical Structure: Category -> Annotation Source -> Term

Pathways -> BioCarta -> p53 signaling pathway

DAVID Gene ID:
It is an internal ID generated on "DAVID Gene Concept" in DAVID system. One DAVID gene ID represents one unique gene cluster belonging to one single gene entry.

Gene-Enrichment: A set of user's input genes is highly associated with certain terms, which is statistically measured by Fisher Exact in DAVID system.

EASE Score:
It is an alternative name of Fisher Exact Statistics in DAVID system, referring to one-tail Fisher Exact Probability Value used for gene-enrichment analysis.

DAVID Id %:
After converting user input gene IDs to corresponding DAVID gene ID, it refers to the percentage of DAVID genes in the list associated with particular annotation term. Since DAVID gene ID is unique per gene, it is more accurate to use DAVID ID% to present the gene-annotation association by removing any redundancy in user gene list, i.e. two user's IDs represent same gene.

DAVID Knowledgebase:
It represents DAVID oracle databases which collect large volume of annotation information from wide range of bioinformatic public resources. It is probably the most largest and comprehensive integrated database in the field.

top

Last Edit: Feb. 2007