WebGestalt Manual

April 12, 2013

The Web-based Gene Set Analysis Toolkit (WebGestalt) is a suite of tools for functional enrichment analysis in various biological contexts. WebGestalt compares a user uploaded gene list with genes in pre-defined functional categories to identify those categories with enriched numbers of user-uploaded genes.

The original version of WebGestalt was described in the paper "WebGestalt: an integrated system for exploring gene sets in various biological contexts." (Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W741-8.). There are some major changes in the current version compared to the original version (these changes were primarily happened in two major updates in 2010 and 2013):

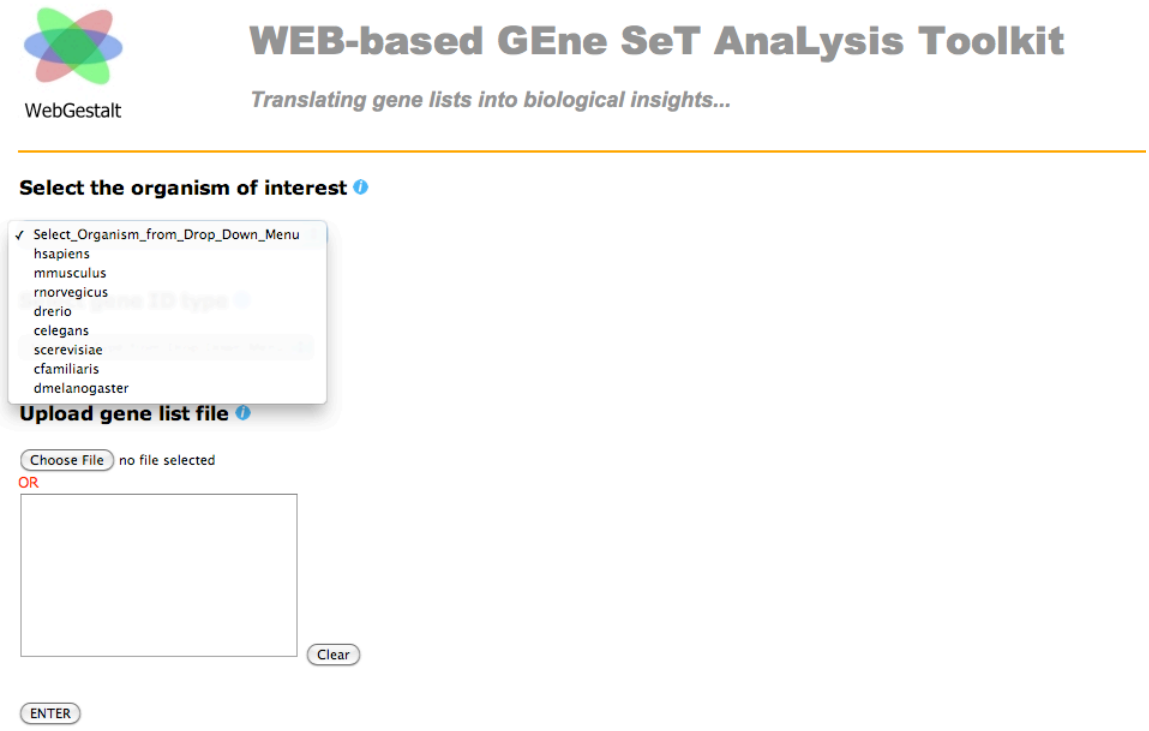
- To simplify the analysis process, the new version requires no user account and log in.
- The new version supports eight organisms including human, mouse, rat, worm, fly, yeast, dog, and zebrafish.
- The new version supports 196 gene identifiers as input, including gene and protein IDs from major public databases and microarray probe IDs from Affematrix, Agilent, Illumina and Codelink. Specially, the new version adds 22 SNP identifiers from SNP arrays, which can allow easy translating of SNP data into biological insights. ID types supported by WebGestalt are listed in Appendix 2 at the end of this document.
- The new version has increased coverage of functional categories in various biological contexts. 78,612 functional categories have been created based on Gene Ontology (GO), KEGG pathways, Pathway Commons pathways, Wikipathways, Transcription factor targets, microRNA targets, protein interaction modules, cytogenetic bands, disease association genes, drug association genes and phenotype association genes.
- New interactive features, such as wikipathway visualization, network DAG and node-link diagram visualization and phenotype DAG visualization, have been added to WebGestalt to help users better understand the enrichment results.
- The new version no longer stores gene sets and analysis results in the system. However, the user is able to save the complete enrichment analysis results.
- Multiple test adjustment is implemented in the new version.

Similar to the original version, the enrichment analysis follows a very simple workflow:

1. **Select** the organism of interest.
2. **Upload** a gene/protein list by uploading text format file or pasting data in the textarea.
3. **Choose** the analysis type (or biological context).
4. **Analyze** the uploaded ID list for Category Enrichment by selecting an appropriate predefined reference set or uploading a user-defined reference set.
5. **Retrieve** results by opening the link directly from this web site. You may also open and/or download a TSV file, or download the zipped results and all results associated with it to a directory on your desktop and unzip.

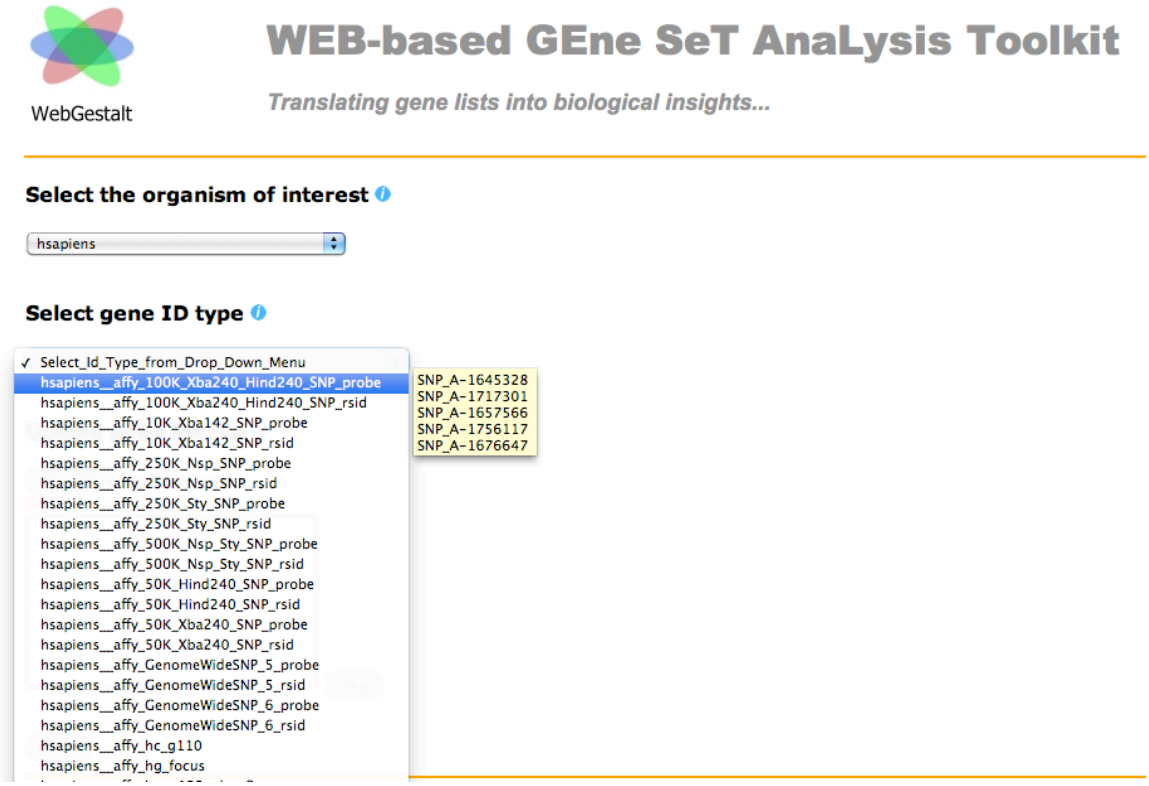
The following screenshots will help illustrate the usage of the tool.

1. Select the organism of interest



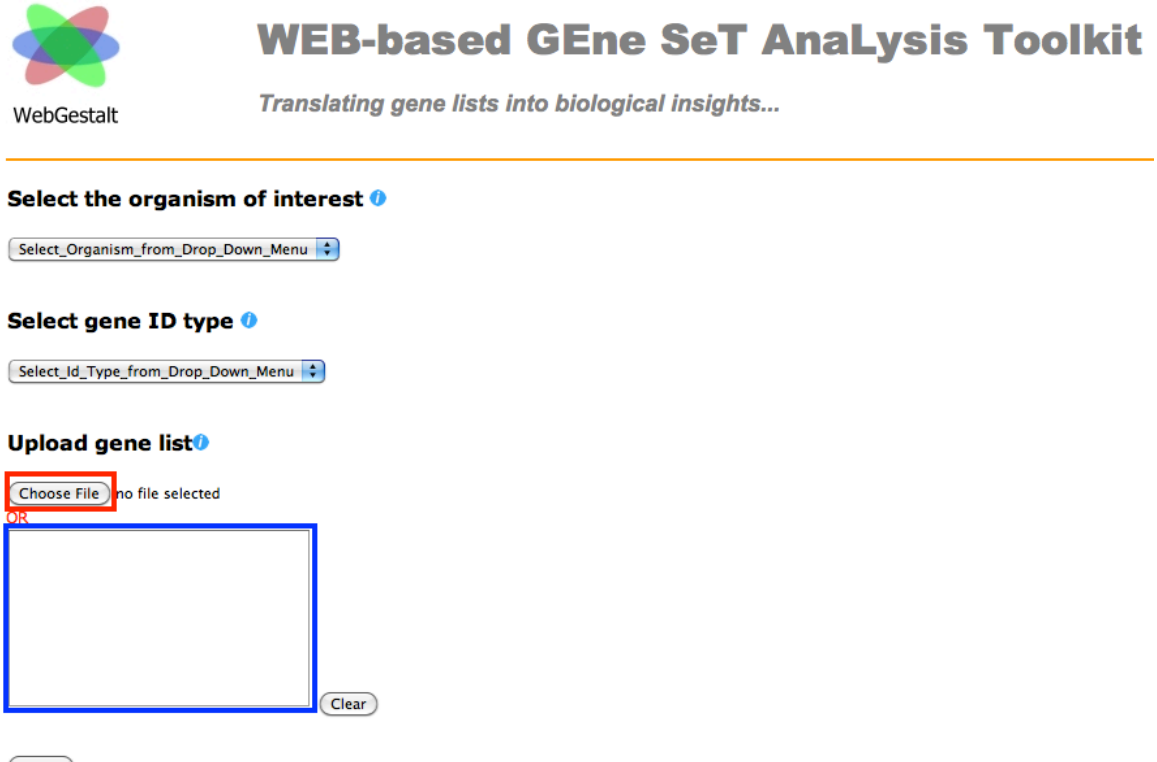
A user needs to select the organism of interest from a drop down menu.

2. Selecting ID type



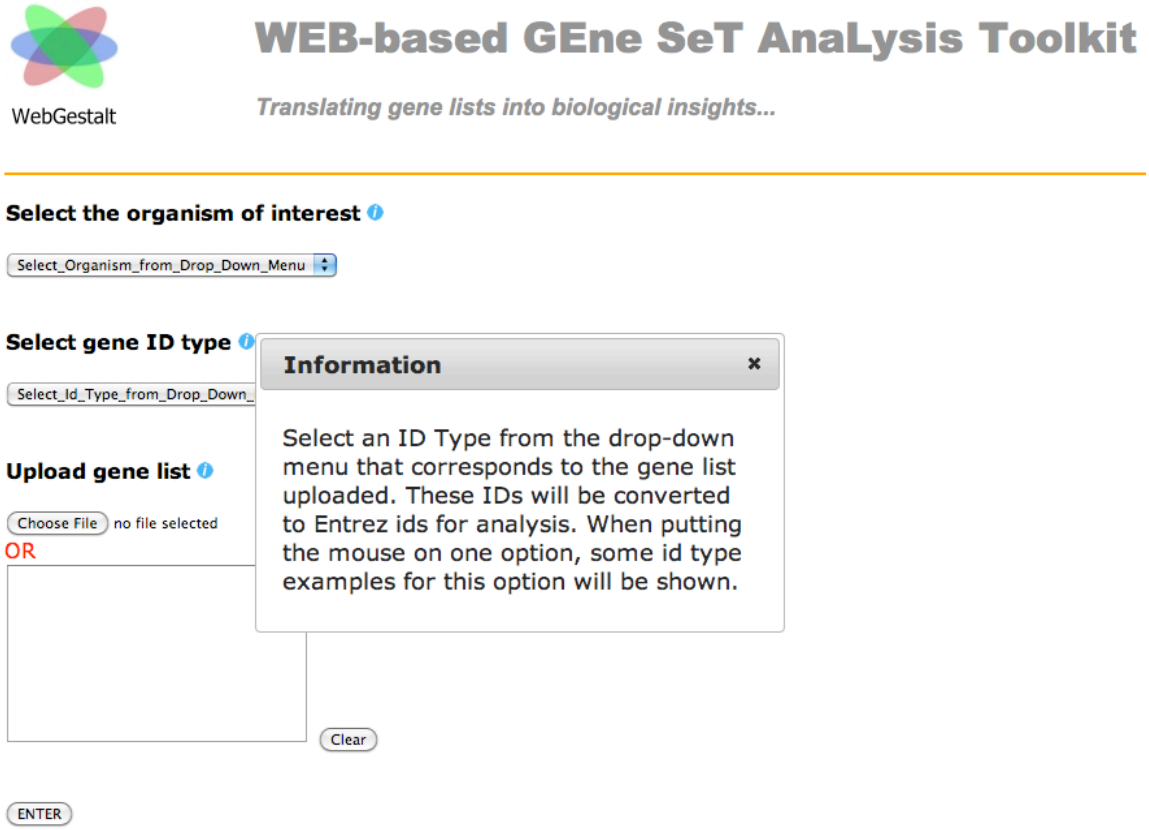
After selecting the organism, WebGestalt will upload all the ID types related to the organism in the drop down menu. To facilitate the users to prepare the data with the correct format of the selected ID type, each ID type has a hint with five identifier examples which can be shown when hovering the mouse on the ID type.

3. Upload gene list



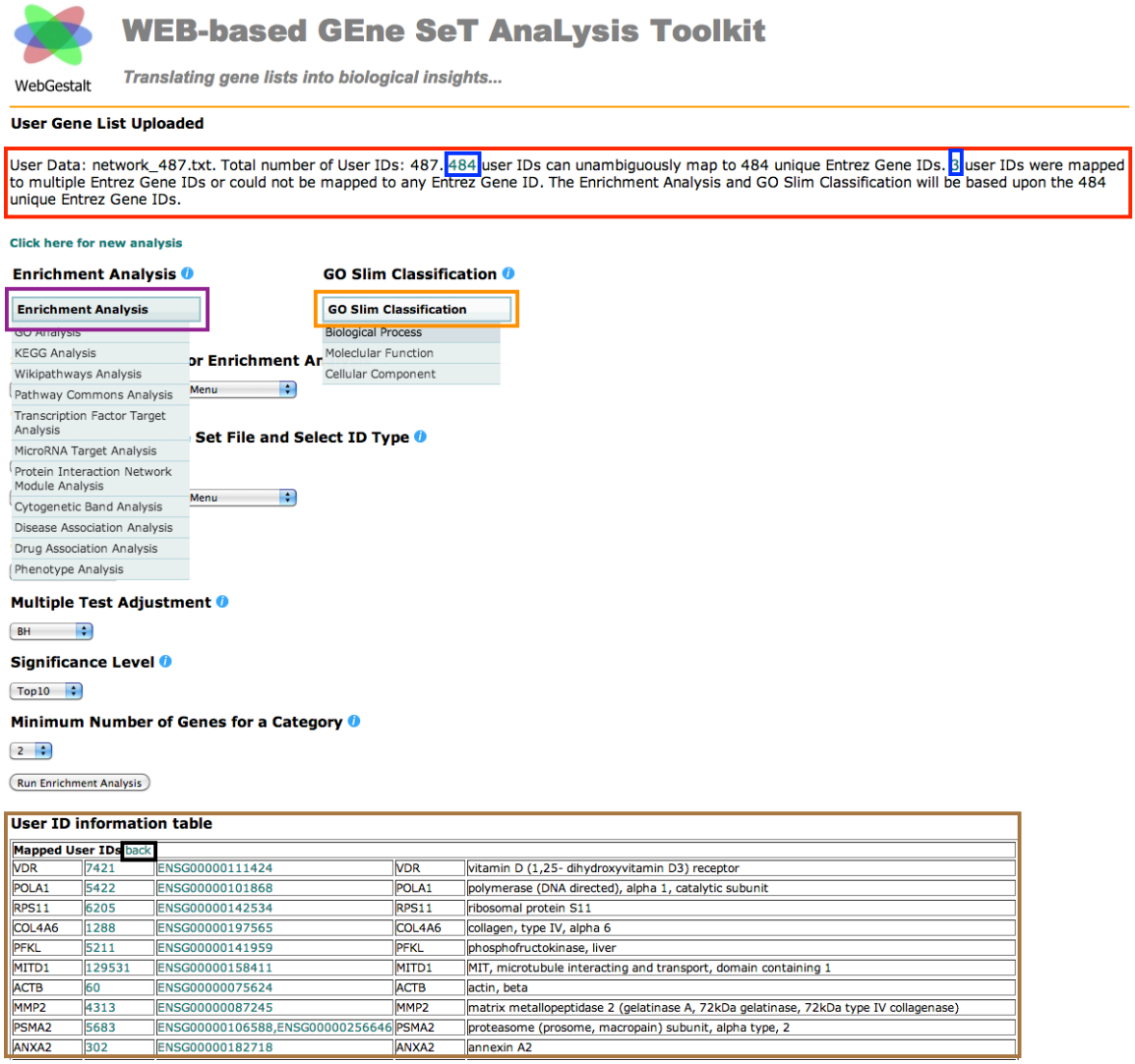
WebGestalt provides two methods to upload a gene/protein list.
1. Upload a gene/protein list file with a text format by clicking the button in the red box, one ID per row. Optionally, put the ID and value in the same row and separate them by a tab. This program only accepts files with a .txt or .tsv extension. Only characters such as a thru z, A thru Z, underscore, numbers, and one period are allowed in the file name. Here are examples of file names: File_2345_name.txt or File_12_17_07.tsv. File names with slashes, extra periods, or spaces etc. are not allowed.
2. Paste a gene/protein list in the text area. One ID per row and values are not supported.
When both options are used simultaneously, WebGestalt will take data in the upload file.

4. Help



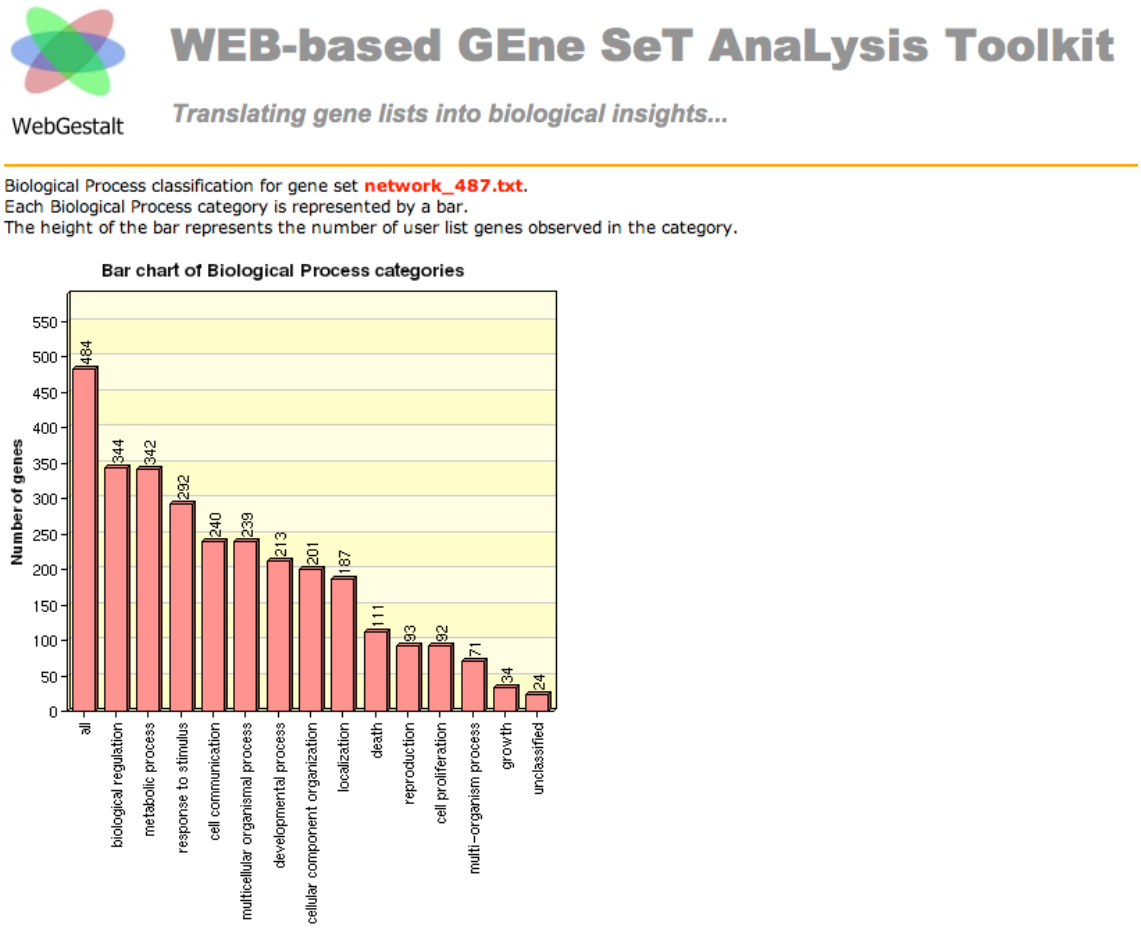
Clicking the “i" button besides the title of each section can show an information dialog which contains the description of this section.

5. Choose the analysis type.



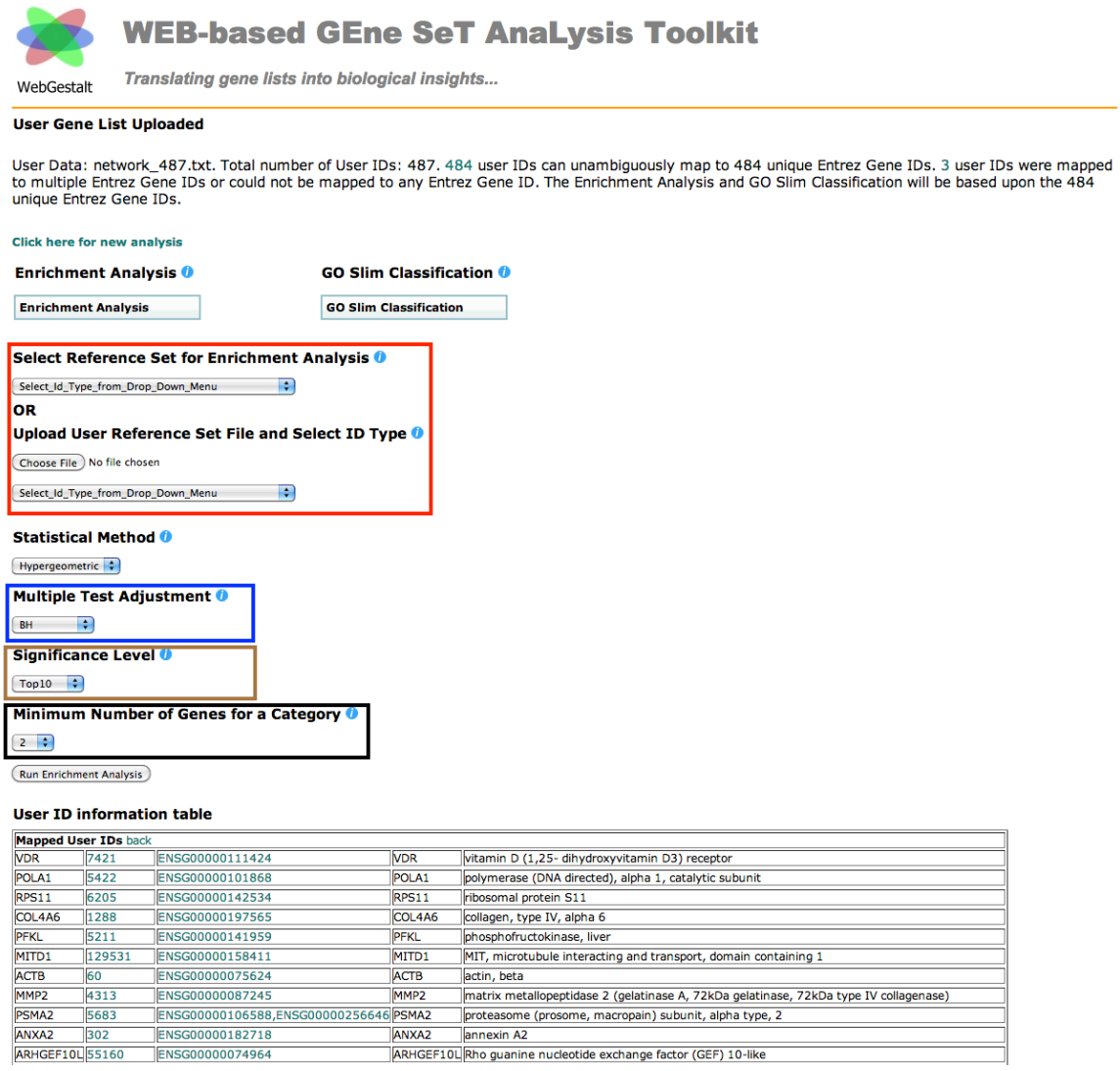
After uploading the gene list into WebGestalt, WebGestalt maps different gene identifiers to Entrez Gene ID because most pathway and network databases use Entrez Gene ID as gene identifier. And the mapping results will be shown at the top of the page (see red box). Only identifiers mapping to a single Entrez Gene ID are used in the enrichment analysis. When multiple identifiers are mapped to the same Entrez Gene ID, they will be counted only once in the enrichment analysis. Identifiers mapping to multiple Entrez Gene IDs and those without Entrez Gene ID mapping are reported in a separate table, and they are not used in the enrichment analysis. Users can click the numbers in the sentence (see blue box) to show the mapping information tables (see brown box). Clicking the "back" button besides the table title (see black box) can jump back to the top of the page. Clicking on the "Enrichment Analysis" button (see purple box) can show the analysis types supported in the WebGestalt.

6. GOSlim classification



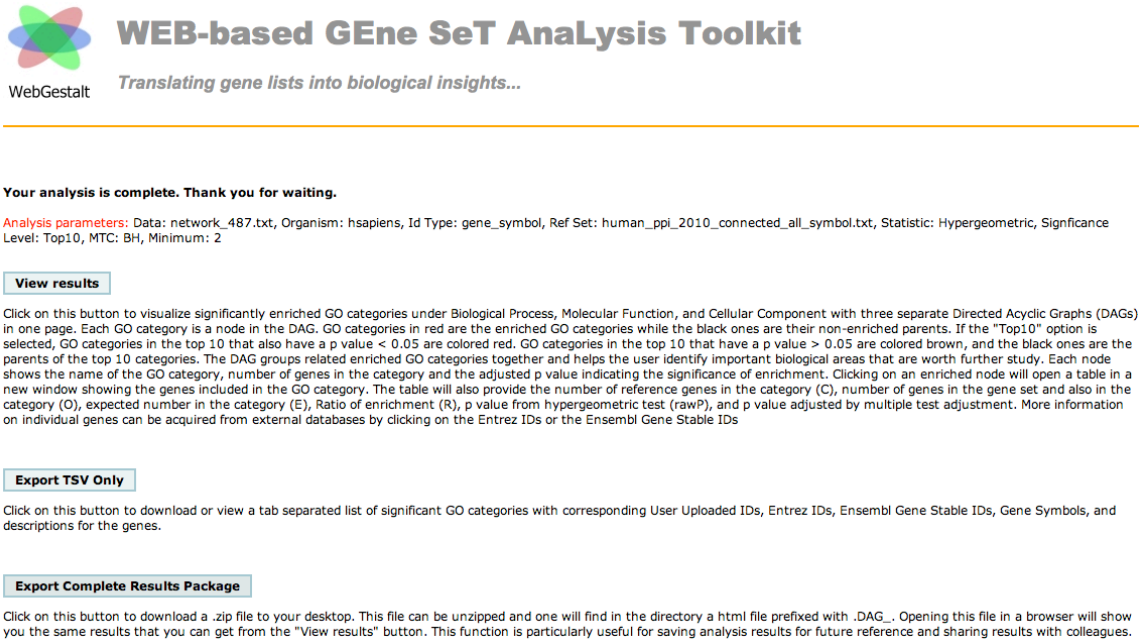
WebGestalt provides GO Slim classification to provide a high-level functional classification of user uploaded genes. Clicking the "GO Slim classification" button (see orange box in the above figure) and selecting one of three ontologies, WebGestalt can plot a bar chart to show the number of user uploaded genes in the GO Slim categories. The label "all" in the bar chart means all Entrez gene IDs that can be derived from the user uploaded genes.

7. Select reference set and choose analysis parameters



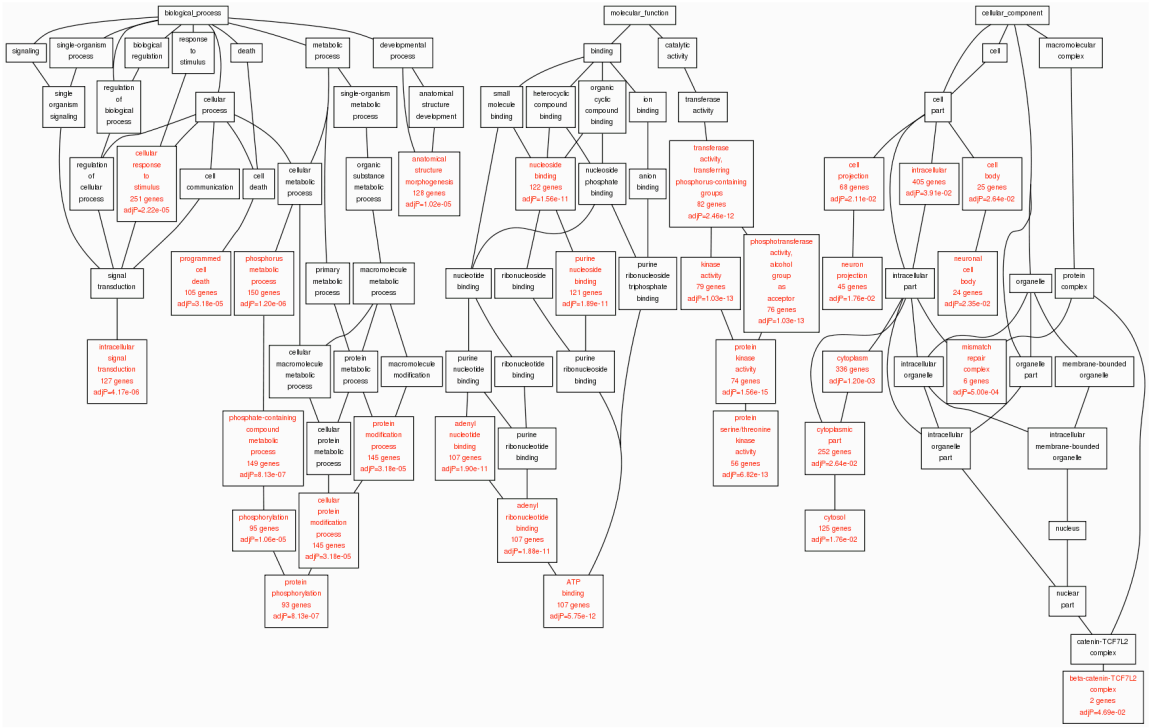
To perform statistical analysis to identify enriched categories, one needs to select a reference gene set from the Reference Set dropdown menu or upload a reference set file and choose the ID Type of the file upload (see red box). The list in the Reference Set dropdown menu provides some commonly used reference gene sets, including all genes in a genome or all genes on a microarray from Affymetrix, Illumina, or Agilent. Alternatively, the user may provide a user-defined reference set by uploading a file with a list of IDs in the text format, one ID per row. In this case, the user also needs to pick an ID type that corresponds to the list of IDs. When both options are used simultaneously, WebGestalt will take data in the upload file. Then the user has the option to choose from five different methods for multiple test adjustment (see blue box). The default method is the one proposed by Benjamini & Hochberg (1995). The user also has the option to use no multiple test adjustment. Next, the user can select a significant level as the cutoff for selecting significantly enriched categories (see brown box). The Gene Set Analysis Toolkit also provides a handy "Top 10" option that always identifies the 10 most significant categories, which can be used as a good start for a gene list of interest. The "Minimum number of genes for a category" can also be set in this page (see black box). The users can also click "i" buttion of each section to get description.

8. Retrieve results.



Once the analysis is complete, the user will be given three buttons to retrieve results. The first button is "View results", which allows the user to browse the results in the WebGestalt. The second one is "Export TSV Only", which allows the user to open a TSV file or download this file and save it to the local computer. The last one is "Export Complete Results Package", which allows the user download the entire zipped results got from the "View results" button. Next, we will introduce the detailed information got from the "View results" button.

8.1 Gene Ontology



Because Gene Ontology has a directed acyclic graph (DAG) structure, to help the user understand the results better, WebGestalt plots the DAG structure for enriched GO categories, in which nodes with red label represents enriched categories and nodes with black label represents their non-enriched parents. If the "top10" option was selected, GO categories in the top 10 and also have a *p* value < 0.05 are colored red, GO categories in the top 10 but have a *p* value > 0.05 are colored brown, and the black ones are the parents of the top 10 categories. The enriched nodes or top 10 nodes show the name of the GO category, number of genes in the category and the adjusted *p* value.

User data and parameters: User data: network_487.txt, Organism: hsapiens, Id Type: gene_symbol, Ref Set: human_ppi_2010_connected_all_symbol.txt, Significance Level: Top10, Statistics Test: Hypergeometric, MTC: BH, Minimum: 2

The results for each enriched GO category are listed in this table. For each GO category, the first row lists its sub-root (biological process, molecular function, or cellular component), category name, and corresponding GO ID. The second row lists number of reference genes in the category (C), number of genes in the gene set and also in the category (O), expected number in the category (E), Ratio of enrichment (R), p value from hypergeometric test (rawP), and p value adjusted by the multiple test adjustment (adjP). Finally, genes in the category are listed. For each gene, the table lists the user uploaded ID and value (optional), Entrez ID, Ensembl Gene Stable ID, Gene symbol, and description. Ensembl Gene Stable ID and Entrez Gene ID are linked to the Ensembl and Entrez Gene databases, respectively.

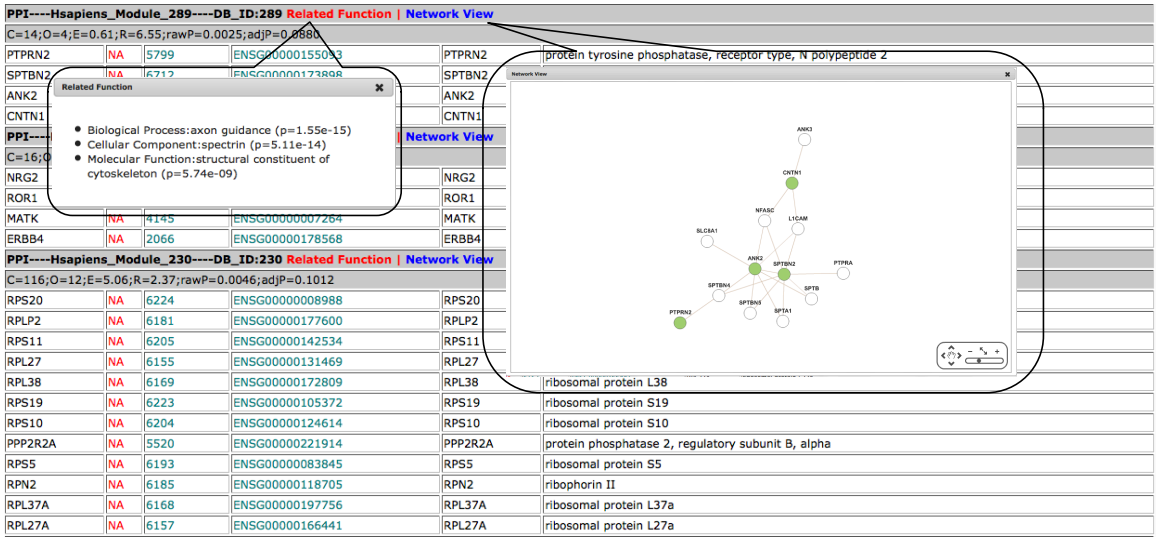
biological process----protein phosphorylation----GO:0006468

C=1023;O=93;E=49.06;R=1.90;rawP=4.46e-10;adjP=8.13e-07

EP300	NA	2033	ENSG00000100393	EP300	E1A binding protein p300
NF1	NA	4763	ENSG00000196712	NF1	neurofibromin 1
KDR	NA	3791	ENSG00000128052	KDR	kinase insert domain receptor (a type III receptor tyrosine kinase)
RET	NA	5979	ENSG00000165731	RET	ret proto-oncogene
TP53	NA	7157	ENSG00000141510	TP53	tumor protein p53
TPX2	NA	22974	ENSG00000088325	TPX2	TPX2, microtubule-associated, homolog (Xenopus laevis)
P2RX7	NA	5027	ENSG00000089041	P2RX7	purinergic receptor P2X, ligand-gated ion channel, 7
RIPK1	NA	8737	ENSG00000137275	RIPK1	receptor (TNFRSF)-interacting serine-threonine kinase 1
CHUK	NA	1147	ENSG00000213341	CHUK	conserved helix-loop-helix ubiquitous kinase
APC	NA	324	ENSG00000134982	APC	adenomatous polyposis coli
CCNB1	NA	891	ENSG00000134057	CCNB1	cyclin B1
ERBB4	NA	2066	ENSG00000178568	ERBB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)
MAP2K3	NA	5606	ENSG00000034152	MAP2K3	mitogen-activated protein kinase kinase 3
EEF2K	NA	29904	ENSG00000103319	EEF2K	eukaryotic elongation factor-2 kinase
CAD	NA	790	ENSG00000084774	CAD	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase
RPS6KA2	NA	6196	ENSG00000071242	RPS6KA2	ribosomal protein S6 kinase, 90kDa, polypeptide 2

Clicking on the nodes with red label, the detailed information about the corresponding category will be shown as a table. The table provides genes in the category, number of reference genes in the category (C), number of genes in the gene set and also in the category (O), expected number in the category (E), Ratio of enrichment (R), *p* value from hypergeometric test (rawP), and *p* value adjusted by multiple test adjustment. More information on individual genes can be acquired from external database by clicking on the Ensembl Gene Stable IDs or the Entrez Gene IDs.

8.2 Hierarchical protein interaction network modules

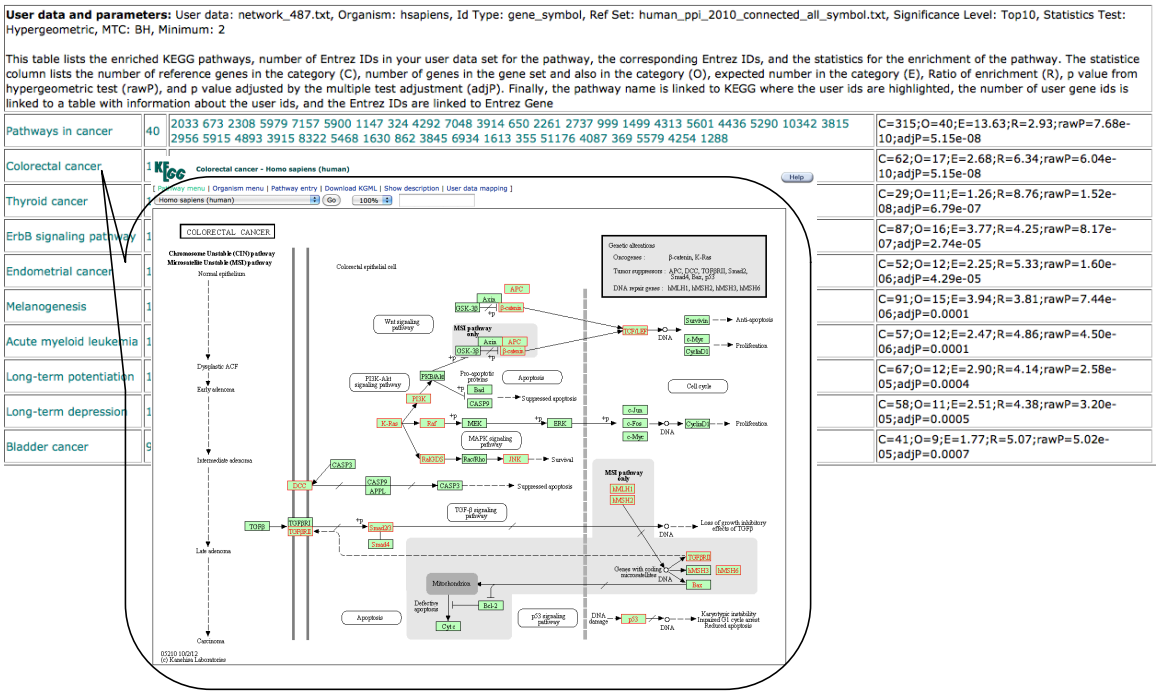


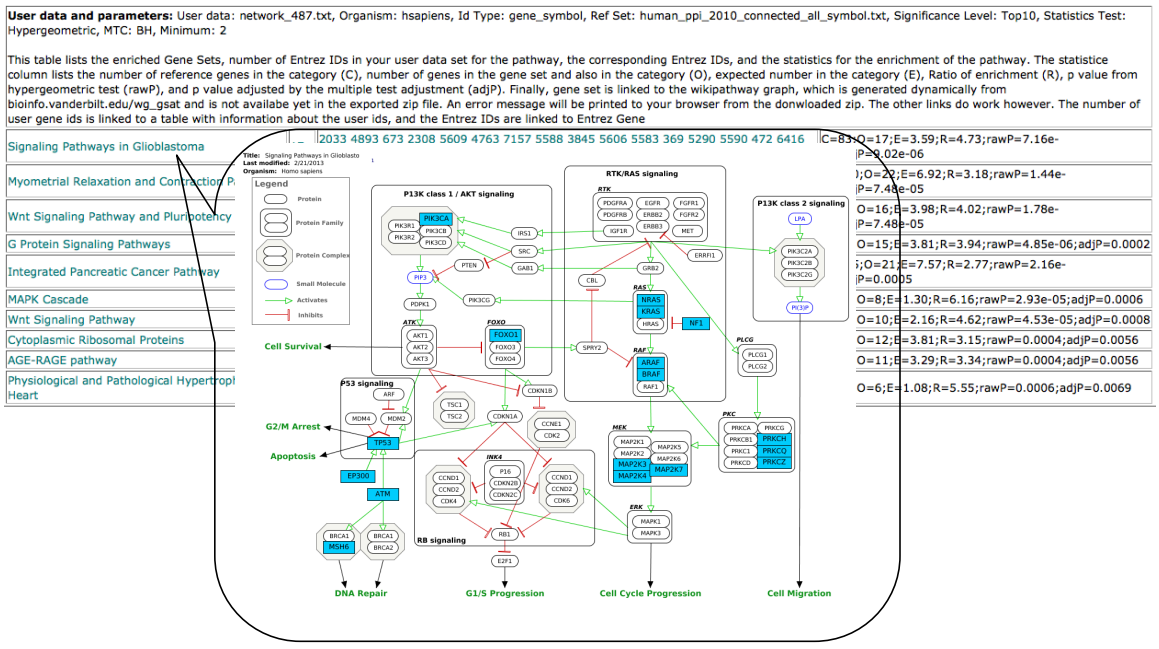
WebGestalt contains human and mouse protein-protein interaction modules. For the human network, we first collected from seven public databases all protein-protein interactions with at least one publication support. After combining these interactions, we removed redundant entries and entries in which one or both interactors had no Entrez Gene IDs. For the mouse network, because the seven databases only contained a limited number of interactions, we used ortholog-based method to infer mouse interactions from curated human interactions. After combining the database curated interactions with inferred interactions, we followed the same cleaning process as described for the human network. Both networks are unweighted. After combining human and mouse interactions, we used the NetSAM package (http://www.bioconductor.org/packages/devel/bioc/html/NetSAM.html) to identify the hierarchical modules. The detailed description of the NetSAM package can be found in our recently published Nature Methods paper "NetGestalt: integrating multidimensional omics data over biological networks" (http://www.nature.com/nmeth/journal/v10/n7/full/nmeth.2517.html).
Because the hierarchical modules also have DAG structure, WebGestalt plots the DAG structure of enriched modules. Clicking nodes with red label can also create table with detailed information about the modules. Clicking "Related Function" button, WebGestalt can show the GO categories most related to the enriched modules. Because protein-protein interaction network has a two-dimensional structure, WebGestalt integrates Cytoscape Web to plot sub-network for each module. Clicking "Network View" besides the module name can visualize in a network graph the uploaded genes (in green) and their direct neighbors (in white) in the enriched module.

8.3 Phenotype

Because phenotype was downloaded from Mammalian Phenotype Ontology and Human Phenotype Ontology which contains phenotype DAG structure, WebGestalt can also plots the DAG structure of enriched phenotype terms. Other operations are similar with GO.

8.4 KEGG pathway, Wikipathway and Pathway Commons pathway





For After clicking “View results” button, there will be a table containing pathway name, the number of genes in the pathway, Entrez gene IDs and statistic values. Because pathway has a detailed structure, WebGestalt will plot the pathway structure when clicking the pathway name. For KEGG pathway, WebGestalt will send the Entrez gene IDs contained in the pathway to KEGG database and get the plot in

which genes in the uploaded gene list will be labeled by the red color. For Wikipathway, WebGestalt will call Wikipathway API to plot the pathway structure in which gene in the uploaded gene list will be filled by blue color. For Pathway Commons pathway, WebGestalt will use Pathway Commons API to search pathway name in the Pathway Commons website and get the detailed information for the searched pathway. Clicking the number of genes will show a table containing detailed gene information.

For other databases, WebGestalt will first show a summary table like table for KEGG pathways. Clicking the number of genes will show the detailed gene information.

Appendix I. Statistical analysis

Statistical analysis is necessary for biological discovery from large gene sets. Statistical tests that have been used for identifying enriched categories include the $\chi 2$ test, the T test, the binomial test and the hypergeometric test. Although being implemented in several existing tools, the T test, the $\chi 2$ test and the binomial test require certain distributions that are usually violated. When analyzing the functional significance of the interesting gene sets produced by for example microarray experiments, all of the genes on the microarray, which represent the population from which the interesting genes are drawn, should be used as the reference. This becomes a sampling without replacement problem and can be appropriately modeled by the hypergeometric distribution. Suppose that we have n genes in the interesting gene set (A) and N genes in the reference gene set (B). Suppose further that there are k genes in A and K genes in B are in a given category (C). If B represents the population from which the genes in A are drew, GSAT uses the hypergeometric test to evaluate the significance of enrichment for category C in gene set A,

$$P = \sum_{i=k}^{n} \frac{\binom{N-K}{n-i}\binom{K}{i}}{\binom{N}{n}}$$

As we are testing multiple categories in a group of functional gene set categories, the p values need to be adjusted for multiple tests. We use the R function p.adjust for this purpose. It provides five different methods for the multiple test adjustment. The default method used in the GSAT is the one proposed by Benjamini & Hochberg (1995). It is one of the less conservative methods, with the bonferroni being the most conservative. Other adjustment methods that are supported are holm, Holm, S. (1979), hommel, Hommel, G. (1988), bonferroni, Hochberg, Y. (1988), BY, Benjamini, Y., and Yekutieli, D. (2001).

Appendix II. Gene/protein ID types supported by the WebGestalt

C.elegans

affy_c_elegans
ensembl_gene_stable_id
ensembl_peptide_id
entrezgene
gene_symbol
genebank
refseq_dna
refseq_dna_all
refseq_peptide
refseq_peptide_all
unigene
uniprot_swissprot_accession
wormbase_locus

C. familiaris

affy_canine_2
ensembl_gene_stable_id
ensembl_peptide_id
entrezgene
gene_symbol
genebank
refseq_dna
refseq_dna_all
refseq_peptide
refseq_peptide_all
unigene
uniprot_swissprot_accession

D. melanogaster

affy_drosgenome1
affy_drosophila_2
ensembl_gene_stable_id
ensembl_peptide_id
entrezgene
flybasename_gene
genebank
refseq_dna
refseq_dna_all
refseq_peptide
refseq_peptide_all
unigene
uniprot_swissprot_accession

D. rerio

ZFIN_ID
affy_zebrafish
agilent_g2518a
agilent_g2519f
ensembl_gene_stable_id
ensembl_peptide_id
entrezgene
genebank
ipi
leiden_leiden2
leiden_leiden3
refseq_dna
refseq_dna_all
refseq_peptide
refseq_peptide_all
unigene
uniprot_swissprot_accession
zfin_symbol

H. sapiens

GN_GPL2700
GN_GPL4372
GN_GPL564
affy_100K_Xba240_Hind240_SNP_probe
affy_100K_Xba240_Hind240_SNP_rsid
affy_10K_Xba142_SNP_probe
affy_10K_Xba142_SNP_rsid
affy_250K_Nsp_SNP_probe
affy_250K_Nsp_SNP_rsid
affy_250K_Sty_SNP_probe
affy_250K_Sty_SNP_rsid
affy_500K_Nsp_Sty_SNP_probe
affy_500K_Nsp_Sty_SNP_rsid
affy_50K_Hind240_SNP_probe
affy_50K_Hind240_SNP_rsid
affy_50K_Xba240_SNP_probe
affy_50K_Xba240_SNP_rsid
affy_GenomeWideSNP_5_probe
affy_GenomeWideSNP_5_rsid
affy_GenomeWideSNP_6_probe
affy_GenomeWideSNP_6_rsid
affy_hc_g110
affy_hg_focus
affy_hg_u133_plus_2
affy_hg_u133a
affy_hg_u133a_2
affy_hg_u133b
affy_hg_u95a
affy_hg_u95av2
affy_hg_u95b
affy_hg_u95c
affy_hg_u95d
affy_hg_u95e
affy_huex_1_0_st_v2
affy_hugene_1_0_st_v1
affy_hugene_1_1_st
affy_hugenefl
affy_u133_x3p
agilent_G4112A
agilent_cgh_44b
agilent_wholegenome_4x44k_v1
agilent_wholegenome_4x44k_v2

codelink
dbSNP
ensembl_gene_stable_id
ensembl_peptide_id
entrezgene
gene_symbol
genebank
illumina_Omni1-Quad_SNP
illumina_OmniExpress_SNP
illumina_humanwg_6_v1
illumina_humanwg_6_v2
illumina_humanwg_6_v3
ipi
refseq_dna
refseq_dna_all
refseq_peptide
refseq_peptide_all
unigene
uniprot_swissprot_accession

M. musculus

GN_GPL2510
GN_GPL339_340
GN_GPL6105
MGI_ID
MOUSEDIVm520650_SNP_probe_set
MOUSEDIVm520650_rs_number
affy_mg_u74a
affy_mg_u74av2
affy_mg_u74b
affy_mg_u74bv2
affy_mg_u74c
affy_mg_u74cv2
affy_moe430a
affy_moe430b
affy_moex_1_0_st_v1
affy_mogene_1_0_st_v1
affy_mouse430_2
affy_mouse430a_2
affy_mu11ksuba
affy_mu11ksubb
agilent_G4121A

agilent_G4121B
agilent_G4122A
agilent_G4122F
agilent_SurePrintG3
agilent_wholegenome_4x44k_v1
agilent_wholegenome_4x44k_v2
codelink
ensembl_gene_stable_id
ensembl_peptide_id
entrezgene
gene_symbol
genebank
illumina_MouseRef-8_v2
illumina_mousewg_6_v1
illumina_mousewg_6_v2
ipi
refseq_dna
refseq_dna_all
refseq_peptide
refseq_peptide_all
unigene
uniprot_swissprot_accession

R. norvegicus

affy_rae230a
affy_rae230b
affy_raex_1_0_st_v1
affy_ragene_1_0_st_v1
affy_rat230_2
affy_rg_u34a
affy_rg_u34b
affy_rg_u34c
affy_rn_u34
affy_rt_u34
agilent_G4131A
agilent_G4131F
agilent_wholegenome_4x44k_v1
agilent_wholegenome_4x44k_v3
codelink
ensembl_gene_stable_id
ensembl_peptide_id
entrezgene

gene_symbol
genebank
ipi
refseq_dna
refseq_dna_all
refseq_peptide
refseq_peptide_all
unigene
uniprot_swissprot_accession

S. cerevisiae

SGD_ID
affy_yeast_2
affy_yg_s98
ensembl_gene_stable_id
ensembl_peptide_id
entrezgene
gene_symbol
genebank
refseq_dna
refseq_dna_all
refseq_peptide
refseq_peptide_all
unigene
uniprot_swissprot_accession