

## 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):

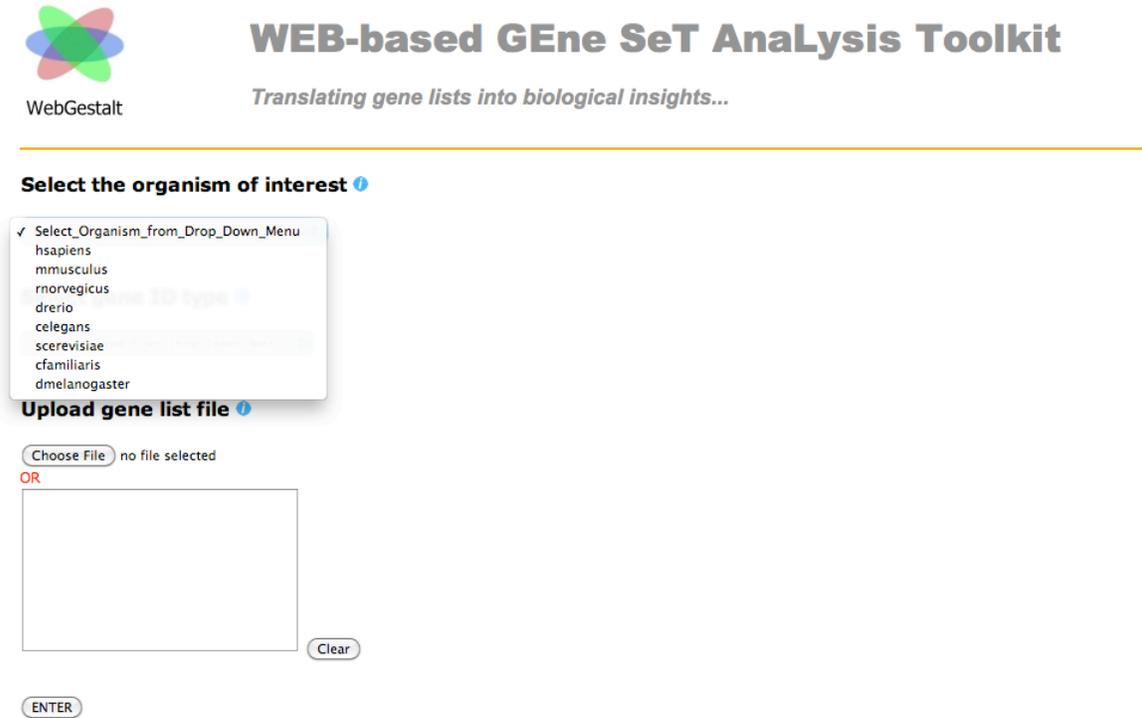
- To simplify the analysis process, the new version requires no user account and log in.
- The new version supports eight organisms including human, mouse, rat, worm, fly, yeast, dog, and zebrafish.
- The new version supports 196 gene identifiers as input, including gene and protein IDs from major public 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



The screenshot displays the WebGestalt website interface. At the top left is the WebGestalt logo, a stylized four-lobed shape in green, red, and blue. To its right is the title "WEB-based GENE SeT AnaLysis Toolkit" in a bold, sans-serif font, with the tagline "Translating gene lists into biological insights..." below it. A horizontal orange line separates the header from the main content area.

The main content area is titled "Select the organism of interest" with a blue information icon. Below this title is a dropdown menu that is currently open, showing a list of organism identifiers: "hsapiens", "mmusculus", "rnorvegicus", "drerio", "celegans", "scerevisiae", "cfamiliaris", and "dmelanogaster". The "celegans" option is highlighted with a light blue background. Above the list, there is a checkmark and the text "Select\_Organism\_from\_Drop\_Down\_Menu". To the right of the list, there is a link "Change ID type" with a blue information icon.

Below the dropdown menu is the section "Upload gene list file" with a blue information icon. It contains a "Choose File" button followed by the text "no file selected". Below this is the word "OR" in red. Underneath is a large empty rectangular box for file upload, and to its right is a "Clear" button. At the bottom of this section is an "ENTER" button.

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

## 2. Selecting ID type



WebGestalt

# WEB-based GENE SeT AnaLysis Toolkit

*Translating gene lists into biological insights...*

### Select the organism of interest ?

hsapiens

### Select gene ID type ?

✓ Select\_Id\_Type\_from\_Drop\_Down\_Menu

hsapiens__affy_100K_Xba240_Hind240_SNP_probe	SNP_A-1645328
hsapiens__affy_100K_Xba240_Hind240_SNP_rsId	SNP_A-1717301
hsapiens__affy_10K_Xba142_SNP_probe	SNP_A-1657566
hsapiens__affy_10K_Xba142_SNP_rsId	SNP_A-1756117
hsapiens__affy_250K_Nsp_SNP_probe	SNP_A-1676647
hsapiens__affy_250K_Nsp_SNP_rsId	
hsapiens__affy_250K_Sty_SNP_probe	
hsapiens__affy_250K_Sty_SNP_rsId	
hsapiens__affy_500K_Nsp_Sty_SNP_probe	
hsapiens__affy_500K_Nsp_Sty_SNP_rsId	
hsapiens__affy_50K_Hind240_SNP_probe	
hsapiens__affy_50K_Hind240_SNP_rsId	
hsapiens__affy_50K_Xba240_SNP_probe	
hsapiens__affy_50K_Xba240_SNP_rsId	
hsapiens__affy_GenomeWideSNP_5_probe	
hsapiens__affy_GenomeWideSNP_5_rsId	
hsapiens__affy_GenomeWideSNP_6_probe	
hsapiens__affy_GenomeWideSNP_6_rsId	
hsapiens__affy_hc_g110	
hsapiens__affy_hg_focus	

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

## WEB-based GENE SeT ANALYSIS Toolkit

*Translating gene lists into biological insights...*

#### Select the organism of interest

Select\_Organism\_from\_Drop\_Down\_Menu 

#### Select gene ID type

Select\_Id\_Type\_from\_Drop\_Down\_Menu 

#### Upload gene list

no file selected

OR

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



WebGestalt

## WEB-based GENE SeT ANALYSIS Toolkit

*Translating gene lists into biological insights...*

### Select the organism of interest ⓘ

Select\_Organism\_from\_Drop\_Down\_Menu ▾

### Select gene ID type ⓘ

Select\_Id\_Type\_from\_Drop\_Down\_...

### Upload gene list ⓘ

Choose File no file selected

OR

Clear

ENTER

**Information** ×

Select an ID Type from the drop-down menu that corresponds to the gene list uploaded. These IDs will be converted to Entrez ids for analysis. When putting the mouse on one option, some id type examples for this option will be shown.

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.



WEB-based GENE SET ANALYSIS Toolkit

Translating gene lists into biological insights...

User Gene List Uploaded

User Data: network\_487.txt. Total number of User IDs: 487. 484 user IDs can unambiguously map to 484 unique Entrez Gene IDs. 3 user IDs were mapped to multiple Entrez Gene IDs or could not be mapped to any Entrez Gene ID. The Enrichment Analysis and GO Slim Classification will be based upon the 484 unique Entrez Gene IDs.

Click here for new analysis

Enrichment Analysis

- GO Analysis
KEGG Analysis
Wikipathways Analysis
Pathway Commons Analysis
Transcription Factor Target Analysis
MicroRNA Target Analysis
Protein Interaction Network Module Analysis
Cytogenetic Band Analysis
Disease Association Analysis
Drug Association Analysis
Phenotype Analysis

GO Slim Classification

- Biological Process
Molecular Function
Cellular Component

or Enrichment Ar Menu

Set File and Select ID Type Menu

Multiple Test Adjustment

BH

Significance Level

Top10

Minimum Number of Genes for a Category

2

Run Enrichment Analysis

User ID information table
Mapped User IDs back
VDR 7421 ENSG00000111424 VDR vitamin D (1,25- dihydroxyvitamin D3) receptor
POLA1 5422 ENSG00000101868 POLA1 polymerase (DNA directed), alpha 1, catalytic subunit
RPS11 6205 ENSG00000142534 RPS11 ribosomal protein S11
COL4A6 1288 ENSG00000197565 COL4A6 collagen, type IV, alpha 6
PFKL 5211 ENSG00000141959 PFKL phosphofructokinase, liver
MITD1 129531 ENSG00000158411 MITD1 MIT, microtubule interacting and transport, domain containing 1
ACTB 60 ENSG00000075624 ACTB actin, beta
MMP2 4313 ENSG00000087245 MMP2 matrix metalloproteinase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)
PSMA2 5683 ENSG00000106588, ENSG00000256646 PSMA2 proteasome (prosome, macropain) subunit, alpha type, 2
ANXA2 302 ENSG00000182718 ANXA2 annexin A2

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

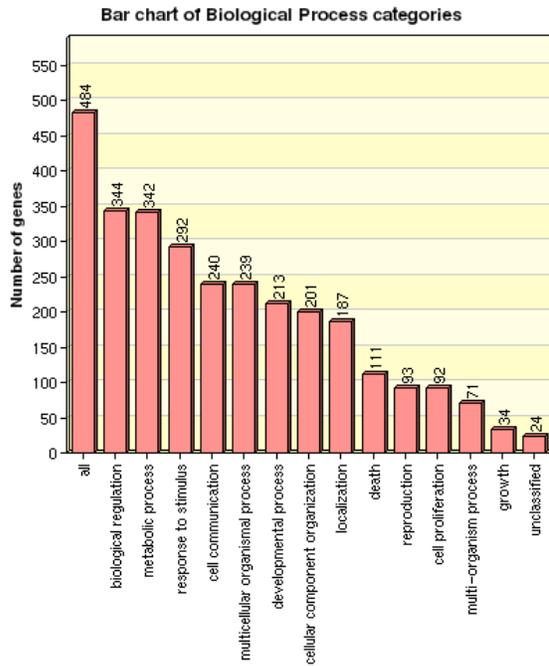
# WEB-based GENE SeT ANALYSIS TOOLKIT

*Translating gene lists into biological insights...*

Biological Process classification for gene set **network\_487.txt**.

Each Biological Process category is represented by a bar.

The height of the bar represents the number of user list genes observed in the category.



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



# WEB-based GENE SET ANALYSIS TOOLKIT

WebGestalt *Translating gene lists into biological insights...*

### User Gene List Uploaded

User Data: network\_487.txt. Total number of User IDs: 487. 484 user IDs can unambiguously map to 484 unique Entrez Gene IDs. 3 user IDs were mapped to multiple Entrez Gene IDs or could not be mapped to any Entrez Gene ID. The Enrichment Analysis and GO Slim Classification will be based upon the 484 unique Entrez Gene IDs.

[Click here for new analysis](#)

#### Enrichment Analysis ?

Enrichment Analysis

#### GO Slim Classification ?

GO Slim Classification

#### Select Reference Set for Enrichment Analysis ?

Select\_Id\_Type\_from\_Drop\_Down\_Menu

OR

#### Upload User Reference Set File and Select ID Type ?

Choose File No file chosen

Select\_Id\_Type\_from\_Drop\_Down\_Menu

#### Statistical Method ?

Hypergeometric

#### Multiple Test Adjustment ?

BH

#### Significance Level ?

Top10

#### Minimum Number of Genes for a Category ?

2

Run Enrichment Analysis

#### User ID information table

Mapped User IDs back			
VDR	7421	ENSG00000111424	VDR vitamin D (1,25- dihydroxyvitamin D3) receptor
POLA1	5422	ENSG00000101868	POLA1 polymerase (DNA directed), alpha 1, catalytic subunit
RPS11	6205	ENSG00000142534	RPS11 ribosomal protein S11
COL4A6	1288	ENSG00000197565	COL4A6 collagen, type IV, alpha 6
PFKL	5211	ENSG00000141959	PFKL phosphofructokinase, liver
MITD1	129531	ENSG00000158411	MITD1 MIT, microtubule interacting and transport, domain containing 1
ACTB	60	ENSG00000075624	ACTB actin, beta
MMP2	4313	ENSG00000087245	MMP2 matrix metalloproteinase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)
PSMA2	5683	ENSG00000106588,ENSG00000256646	PSMA2 proteasome (prosome, macropain) subunit, alpha type, 2
ANXA2	302	ENSG00000182718	ANXA2 annexin A2
ARHGEF10L	55160	ENSG00000074964	ARHGEF10L Rho guanine nucleotide exchange factor (GEF) 10-like

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” button of each section to get description.

## 8. Retrieve results.



WebGestalt

# WEB-based GENE SET ANALYSIS TOOLKIT

Translating gene lists into biological insights...

Your analysis is complete. Thank you for waiting.

**Analysis parameters:** Data: network\_487.txt, Organism: hsapiens, Id Type: gene\_symbol, Ref Set: human\_ppi\_2010\_connected\_all\_symbol.txt, Statistic: Hypergeometric, Significance Level: Top10, MTC: BH, Minimum: 2

### View results

Click on this button to visualize significantly enriched GO categories under Biological Process, Molecular Function, and Cellular Component with three separate Directed Acyclic Graphs (DAGs) in one page. Each GO category is a node in the DAG. GO categories in red are the enriched GO categories while the black ones are their non-enriched parents. If the "Top10" option is selected, GO categories in the top 10 that also have a p value < 0.05 are colored red. GO categories in the top 10 that have a p value > 0.05 are colored brown, and the black ones are the parents of the top 10 categories. The DAG groups related enriched GO categories together and helps the user identify important biological areas that are worth further study. Each node shows the name of the GO category, number of genes in the category and the adjusted p value indicating the significance of enrichment. Clicking on an enriched node will open a table in a new window showing the genes included in the GO category. The table will also provide the 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 databases by clicking on the Entrez IDs or the Ensembl Gene Stable IDs

### Export TSV Only

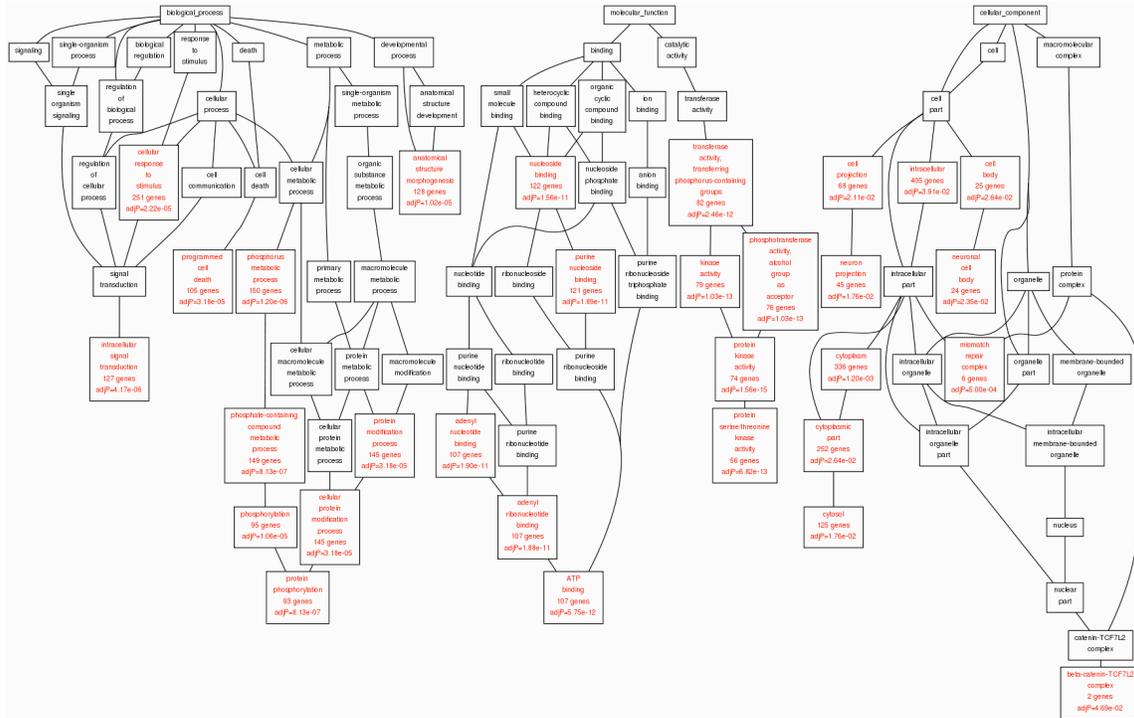
Click on this button to download or view a tab separated list of significant GO categories with corresponding User Uploaded IDs, Entrez IDs, Ensembl Gene Stable IDs, Gene Symbols, and descriptions for the genes.

### Export Complete Results Package

Click on this button to download a .zip file to your desktop. This file can be unzipped and one will find in the directory a html file prefixed with .DAG.. Opening this file in a browser will show you the same results that you can get from the "View results" button. This function is particularly useful for saving analysis results for future reference and sharing results with colleagues.

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.

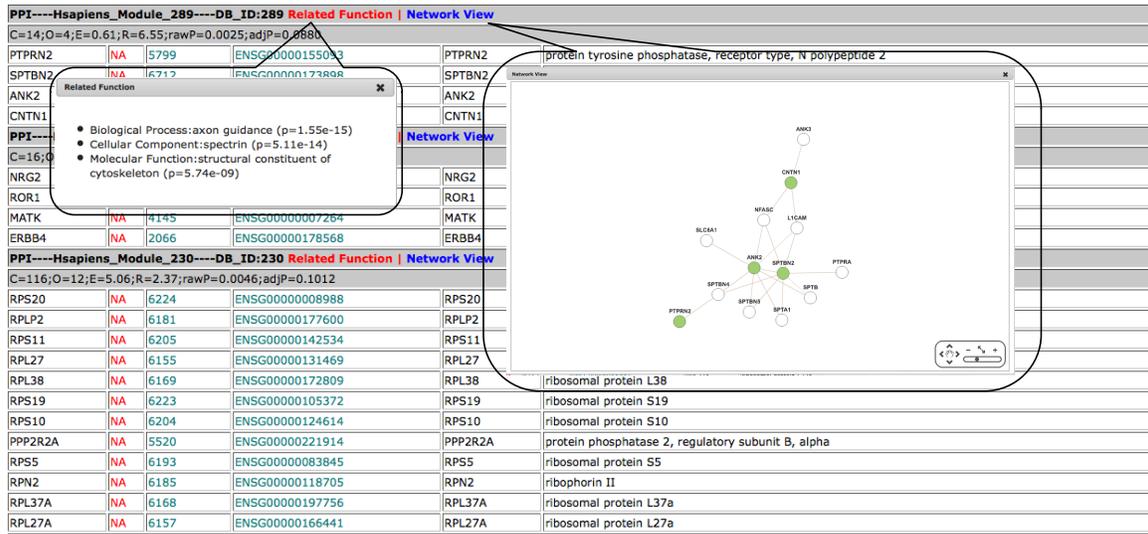
**User data and parameters:** User data: network\_487.txt, Organism: hsapiens, Id Type: gene\_symbol, Ref Set: human\_ppi\_2010\_connected\_all\_symbol.txt, Significance Level: Top10, Statistics Test: Hypergeometric, MTC: BH, Minimum: 2

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 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, 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

**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

This table lists the enriched KEGG pathways, number of Entrez IDs in your user data set for the pathway, the corresponding Entrez IDs, and the statistics for the enrichment of the pathway. The statistic column lists the 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, the pathway name is linked to KEGG where the user ids are highlighted, the number of user gene ids is linked to a table with information about the user ids, and the Entrez IDs are linked to Entrez Gene

Pathways in cancer	40	2033 673 2308 5979 7157 5900 1147 324 4292 7048 3914 650 2261 2737 999 1499 4313 5601 4436 5290 10342 3815 2956 5915 4893 3915 8322 5468 1630 862 3845 6934 1613 355 51176 4087 369 5579 4254 1288	C=315;O=40;E=13.63;R=2.93;rawP=7.68e-10;adjP=5.15e-08
Colorectal cancer	1	2033 673 2308 5979 7157 5900 1147 324 4292 7048 3914 650 2261 2737 999 1499 4313 5601 4436 5290 10342 3815 2956 5915 4893 3915 8322 5468 1630 862 3845 6934 1613 355 51176 4087 369 5579 4254 1288	C=62;O=17;E=2.68;R=6.34;rawP=6.04e-10;adjP=5.15e-08
Thyroid cancer	1	2033 673 2308 5979 7157 5900 1147 324 4292 7048 3914 650 2261 2737 999 1499 4313 5601 4436 5290 10342 3815 2956 5915 4893 3915 8322 5468 1630 862 3845 6934 1613 355 51176 4087 369 5579 4254 1288	C=29;O=11;E=1.26;R=8.76;rawP=1.52e-08;adjP=6.79e-07
ErbB signaling pathway	1	2033 673 2308 5979 7157 5900 1147 324 4292 7048 3914 650 2261 2737 999 1499 4313 5601 4436 5290 10342 3815 2956 5915 4893 3915 8322 5468 1630 862 3845 6934 1613 355 51176 4087 369 5579 4254 1288	C=87;O=16;E=3.77;R=4.25;rawP=8.17e-07;adjP=2.74e-05
Endometrial cancer	1	2033 673 2308 5979 7157 5900 1147 324 4292 7048 3914 650 2261 2737 999 1499 4313 5601 4436 5290 10342 3815 2956 5915 4893 3915 8322 5468 1630 862 3845 6934 1613 355 51176 4087 369 5579 4254 1288	C=52;O=12;E=2.25;R=5.33;rawP=1.60e-06;adjP=4.29e-05
Melanogenesis	1	2033 673 2308 5979 7157 5900 1147 324 4292 7048 3914 650 2261 2737 999 1499 4313 5601 4436 5290 10342 3815 2956 5915 4893 3915 8322 5468 1630 862 3845 6934 1613 355 51176 4087 369 5579 4254 1288	C=91;O=15;E=3.94;R=3.81;rawP=7.44e-06;adjP=0.0001
Acute myeloid leukemia	1	2033 673 2308 5979 7157 5900 1147 324 4292 7048 3914 650 2261 2737 999 1499 4313 5601 4436 5290 10342 3815 2956 5915 4893 3915 8322 5468 1630 862 3845 6934 1613 355 51176 4087 369 5579 4254 1288	C=57;O=12;E=2.47;R=4.86;rawP=4.50e-06;adjP=0.0001
Long-term potentiation	1	2033 673 2308 5979 7157 5900 1147 324 4292 7048 3914 650 2261 2737 999 1499 4313 5601 4436 5290 10342 3815 2956 5915 4893 3915 8322 5468 1630 862 3845 6934 1613 355 51176 4087 369 5579 4254 1288	C=67;O=12;E=2.90;R=4.14;rawP=2.58e-05;adjP=0.0004
Long-term depression	1	2033 673 2308 5979 7157 5900 1147 324 4292 7048 3914 650 2261 2737 999 1499 4313 5601 4436 5290 10342 3815 2956 5915 4893 3915 8322 5468 1630 862 3845 6934 1613 355 51176 4087 369 5579 4254 1288	C=58;O=11;E=2.51;R=4.38;rawP=3.20e-05;adjP=0.0005
Bladder cancer	6	2033 673 2308 5979 7157 5900 1147 324 4292 7048 3914 650 2261 2737 999 1499 4313 5601 4436 5290 10342 3815 2956 5915 4893 3915 8322 5468 1630 862 3845 6934 1613 355 51176 4087 369 5579 4254 1288	C=41;O=9;E=1.77;R=5.07;rawP=5.02e-05;adjP=0.0007

**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

This table lists the enriched Gene Sets, number of Entrez IDs in your user data set for the pathway, the corresponding Entrez IDs, and the statistics for the enrichment of the pathway. The statistic column lists the 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, gene set is linked to the wikipathway graph, which is generated dynamically from bioinfo.vanderbilt.edu/wg\_gsat and is not available yet in the exported zip file. An error message will be printed to your browser from the downloaded zip. The other links do work however. The number of user gene ids is linked to a table with information about the user ids, and the Entrez IDs are linked to Entrez Gene

Signaling Pathways in Glioblastoma	2033 4893 673 2308 5609 4763 7157 5588 3845 5606 5583 369 5290 5590 472 6416	C=83;O=17;E=3.59;R=4.73;rawP=7.16e-06;adjP=0.02e-06
Myometrial Relaxation and Contraction	2033 4893 673 2308 5609 4763 7157 5588 3845 5606 5583 369 5290 5590 472 6416	C=20;O=2;E=6.92;R=3.18;rawP=1.44e-05;adjP=7.48e-05
Wnt Signaling Pathway and Pluripotency	2033 4893 673 2308 5609 4763 7157 5588 3845 5606 5583 369 5290 5590 472 6416	C=16;O=3;E=3.98;R=4.02;rawP=1.78e-05;adjP=7.48e-05
G Protein Signaling Pathways	2033 4893 673 2308 5609 4763 7157 5588 3845 5606 5583 369 5290 5590 472 6416	C=15;O=3;E=3.81;R=3.94;rawP=4.85e-06;adjP=0.0002
Integrated Pancreatic Cancer Pathway	2033 4893 673 2308 5609 4763 7157 5588 3845 5606 5583 369 5290 5590 472 6416	C=21;O=5;E=7.57;R=2.77;rawP=2.16e-05;adjP=0.0005
MAPK Cascade	2033 4893 673 2308 5609 4763 7157 5588 3845 5606 5583 369 5290 5590 472 6416	C=8;O=1;E=1.30;R=6.16;rawP=2.93e-05;adjP=0.0006
Wnt Signaling Pathway	2033 4893 673 2308 5609 4763 7157 5588 3845 5606 5583 369 5290 5590 472 6416	C=10;O=2;E=2.16;R=4.62;rawP=4.53e-05;adjP=0.0008
Cytoplasmic Ribosomal Proteins	2033 4893 673 2308 5609 4763 7157 5588 3845 5606 5583 369 5290 5590 472 6416	C=12;O=2;E=3.81;R=3.15;rawP=0.0004;adjP=0.0056
AGE-RAGE pathway	2033 4893 673 2308 5609 4763 7157 5588 3845 5606 5583 369 5290 5590 472 6416	C=11;O=1;E=3.29;R=3.34;rawP=0.0004;adjP=0.0056
Physiological and Pathological Hypertroph Heart	2033 4893 673 2308 5609 4763 7157 5588 3845 5606 5583 369 5290 5590 472 6416	C=6;O=1;E=1.08;R=5.55;rawP=0.0006;adjP=0.0069

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 drawn, 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  
genbank  
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  
genbank  
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\_rsids  
affy\_10K\_Xba142\_SNP\_probe  
affy\_10K\_Xba142\_SNP\_rsids  
affy\_250K\_Nsp\_SNP\_probe  
affy\_250K\_Nsp\_SNP\_rsids  
affy\_250K\_Sty\_SNP\_probe  
affy\_250K\_Sty\_SNP\_rsids  
affy\_500K\_Nsp\_Sty\_SNP\_probe  
affy\_500K\_Nsp\_Sty\_SNP\_rsids  
affy\_50K\_Hind240\_SNP\_probe  
affy\_50K\_Hind240\_SNP\_rsids  
affy\_50K\_Xba240\_SNP\_probe  
affy\_50K\_Xba240\_SNP\_rsids  
affy\_GenomeWideSNP\_5\_probe  
affy\_GenomeWideSNP\_5\_rsids  
affy\_GenomeWideSNP\_6\_probe  
affy\_GenomeWideSNP\_6\_rsids  
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\_rm\_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