User Manual of WebGestalt 2024

October 2023

The WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) is a suite of tools for functional enrichment analysis in various biological contexts. 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), followed by two updates: WebGestalt 2013, described in the paper "WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013" (Nucleic Acids Res. 2013 Jul 1;41(Web Server issue): W77-83), WebGestalt 2017 described in "WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit" (Nucleic Acids Res. 2017 Jul 3;45(Web Server issue): W130-137), and WebGestalt 2024. WebGestalt has been visited by 50,000 unique users on average per year from 144 countries and territories according to Google Analytics. The publications have also been cited in more than 5,000 scientific according to Google Scholar.

For WebGestalt 2024, the following major improvements are introduced compared to the WebGestalt 2019 version:

- The new version significantly increases the number of functional categories from 321,251 to 620,000, which includes metabolomics
- Ability to add multiple lists for meta-analysis
    - Can be a mix of analyte types, or multiple lists of the same analyte type
- New pathway maps for multi-list analysis
    - Different colors indicate the effects from the different lists
- New user interface
    - Allows for easy entry of multiple lists, as well as naming
    - Less options necessary to start your analysis
- Ease of use features
    - IDs are now auto-detected.
    - Databases are automatically filtered after list input to show valid options
- Increase of performance by over 10 times
    - GSEA should now take under 15 seconds, compared to 3 minutes

At the top of the WebGestalt interface (blue box in Figure 1), the user can find the link about the sample run, external examples, the manual, citation information and user forum. Users are encouraged to post questions or problems they encounter in the user forum. Clicking the “ORA Sample Run”, “GSEA Sample Run”, “NTA Sample run” or “Phosphosite Sample Run” can automatically fill out all parameters in the interface related to each method or data type for a test run. “External Examples” link provides some examples that can guide users to prepare their own data for analysis. The right part of the page also includes the introduction of the WebGestalt 2024, Data sources and News. The following sections will introduce the detailed information about the input parameters and the output report of WebGestalt.

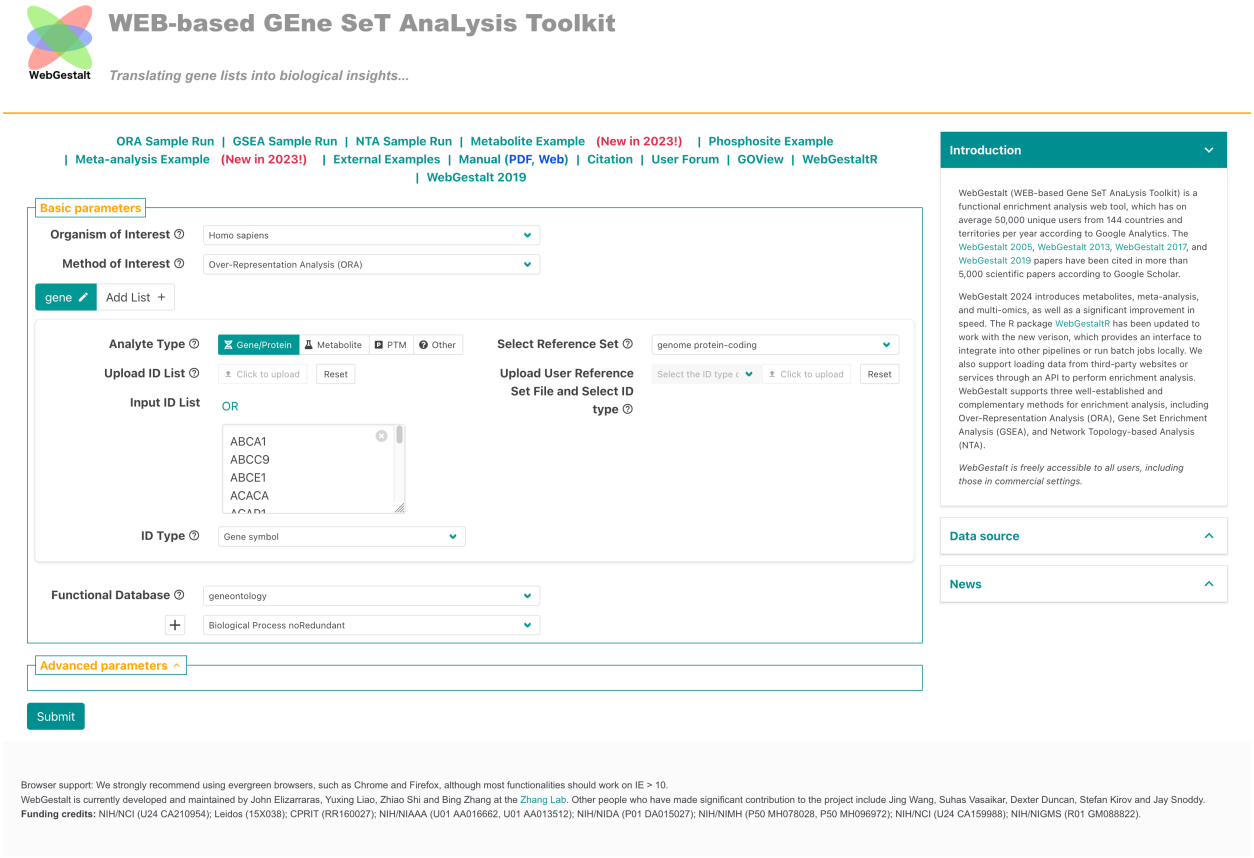


Figure 1. WebGestalt submission interface

1. Select an interesting organism

A user needs to select an organism of interest from the drop-down menu that includes 12 organisms plus an “others” option. We will introduce “others” in the “Select a functional database” section.

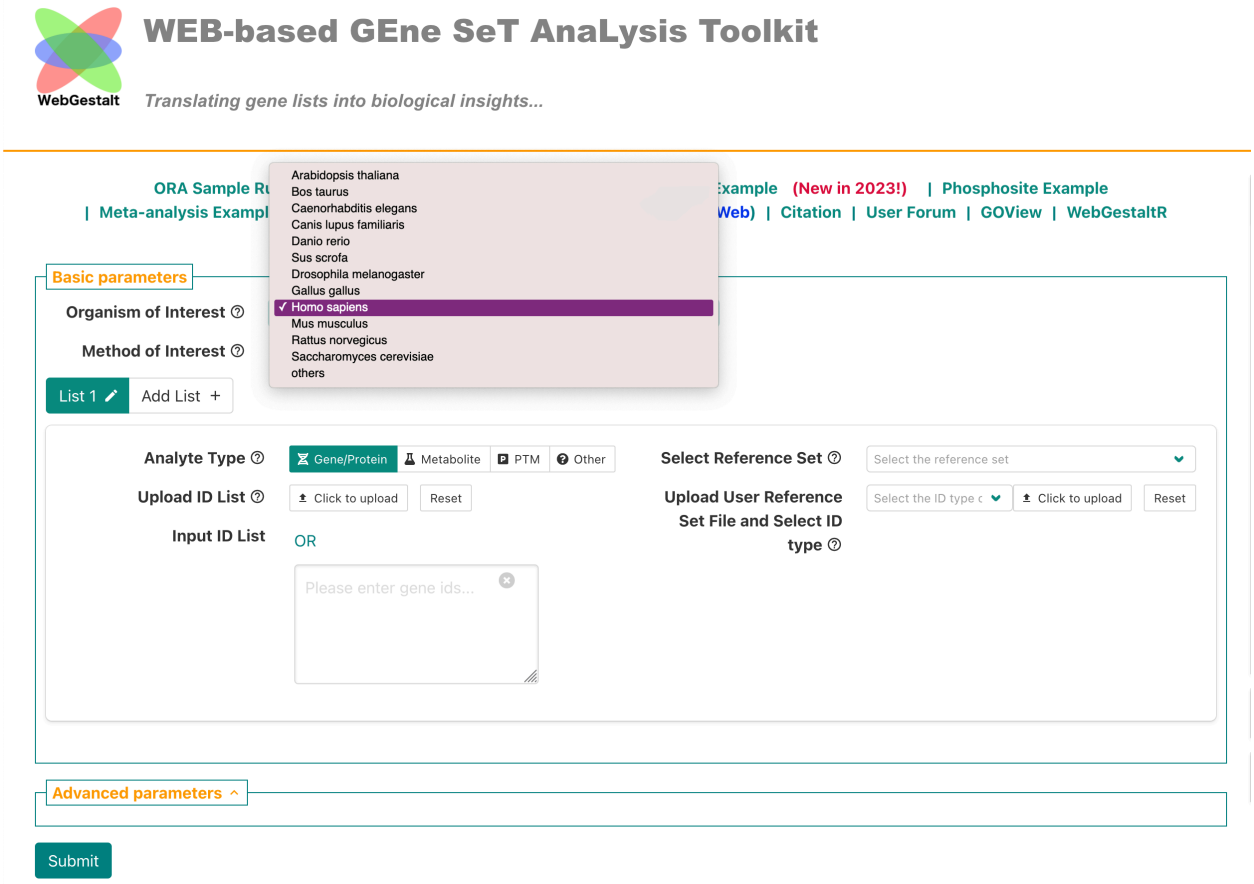


Figure 2. Selection of an interesting organism.

2. Select an interesting method

A user needs to select a method of interest from the dropdown. Different methods have different parameter inputs.

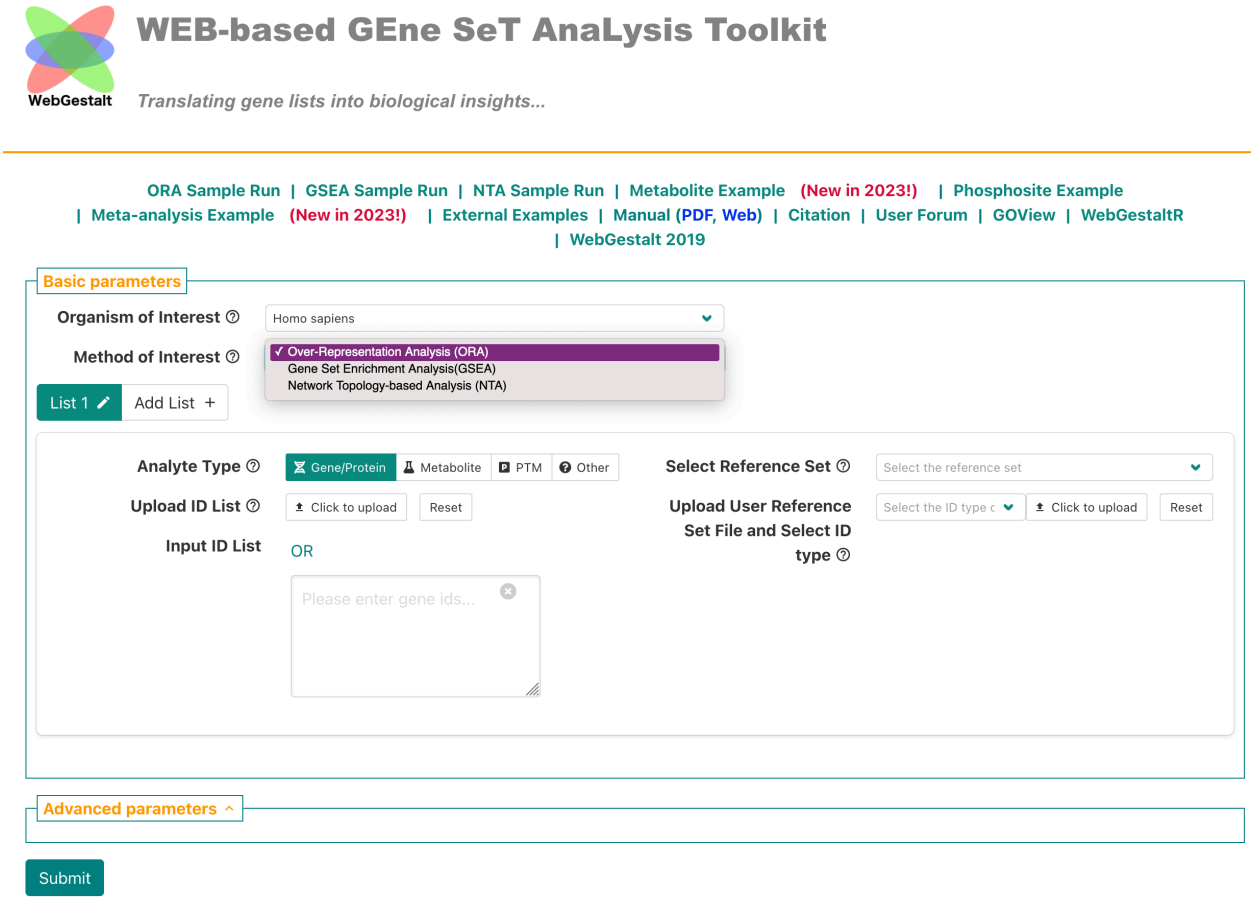


Figure 3. Selection of an interesting method

3. NEW Upload gene list

Before selecting or uploading the functional database, the user needs to upload or paste a gene list. As shown in Figure 8, the user should first select the analyte type of the analyte list from the list (red box). Then, if the user selects the “ORA” or “NTA” method, the user can upload a “txt” file with only one column or paste a gene list to the text box (blue box in Figure 4). If the user selects the “GSEA” method, the user should upload a “rnk” file with two columns: gene IDs and the scores separated by tab (see Table 3) or paste gene IDs and the scores separated by tab in the text box. After inputting your list, WebGestalt will automatically determine your ID type (green box). If you receive an error message indicating that there are no matched ID types, please verify you have selected the correct analyte time. If so, we may not support the ID type you provided. If the message indicates there are multiple options, there are different ID types that could match, and you must select the ID type.

For multi-list inputs, fill out the first list completely, then click on the “Add List +” button (orange box). Then fill out the list. You can use the buttons on in the orange box to rename the lists, and remove lists. List names can be a maximum of 10 characters long, and we only support up to 10 lists.

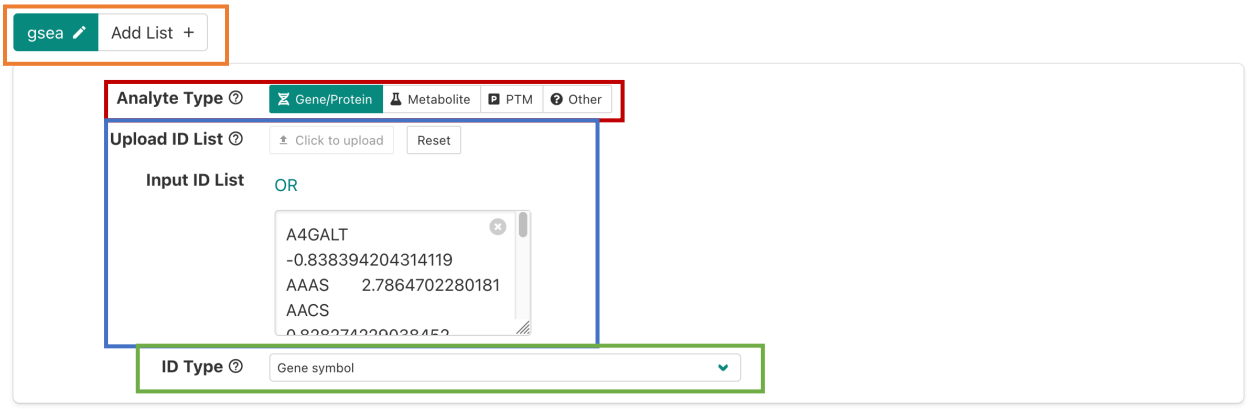


Figure 4. Upload Gene list and select ID type.

Table 3. An example of “rnk” file

Gene1	1.2
Gene2	-1.1
Gene3	-3.3
Gene4	5.5

4. Upload the reference gene list

For the "**ORA**" method, the user needs to select a reference gene list from a dropdown menu (see Figure 5) for example genome or many microarray platforms or select ID type and upload a custom "txt" file. The "GSEA" and "NTA" methods do not need a reference gene list. If you do not see the option to set your reference set, verify that you have selected the correct method (step 2). If you upload your list, the ID type will also be automatically detected, similar to the ID list.

Note: For metabolomics it is *strongly* recommended you provide your own reference set.

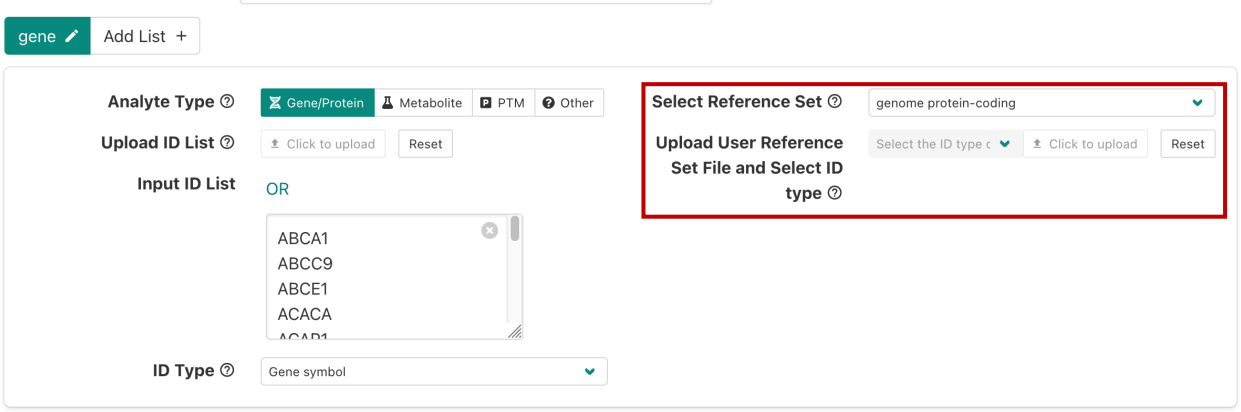


Figure 5. Selection of the reference gene list

5. Select a functional database
5.1 For “others” organism

If a user selects “others” from the “Select Organism of Interest” menu, the user can analyze any kind of data not currently provided by WebGestalt. The user needs to upload the functional database, genes of interest, and reference genes (for the ORA method). Because WebGestalt will not perform any ID mapping for the uploaded data, “others, the user needs to make sure that the ID types in all uploaded data are the same. For this case, select “other” as the analyte type in step 3 (red box).

The uploaded functional database file should have an extension of “**gmt**”. As shown in Table 1, the first column of the file is the gene set ID, the second column is the external link to the gene set and other columns are the gene IDs annotated to this gene set. The file should be tab-delimited. If each gene set ID also has a description (e.g. the name of the gene set ID), the user can also upload a “**des**” file (this is **optional**). As shown in Table 2, the first column is the gene set ID which should be identical to the ID in the gmt file, and the second column is the description for each gene set. All columns should be separated by tab.

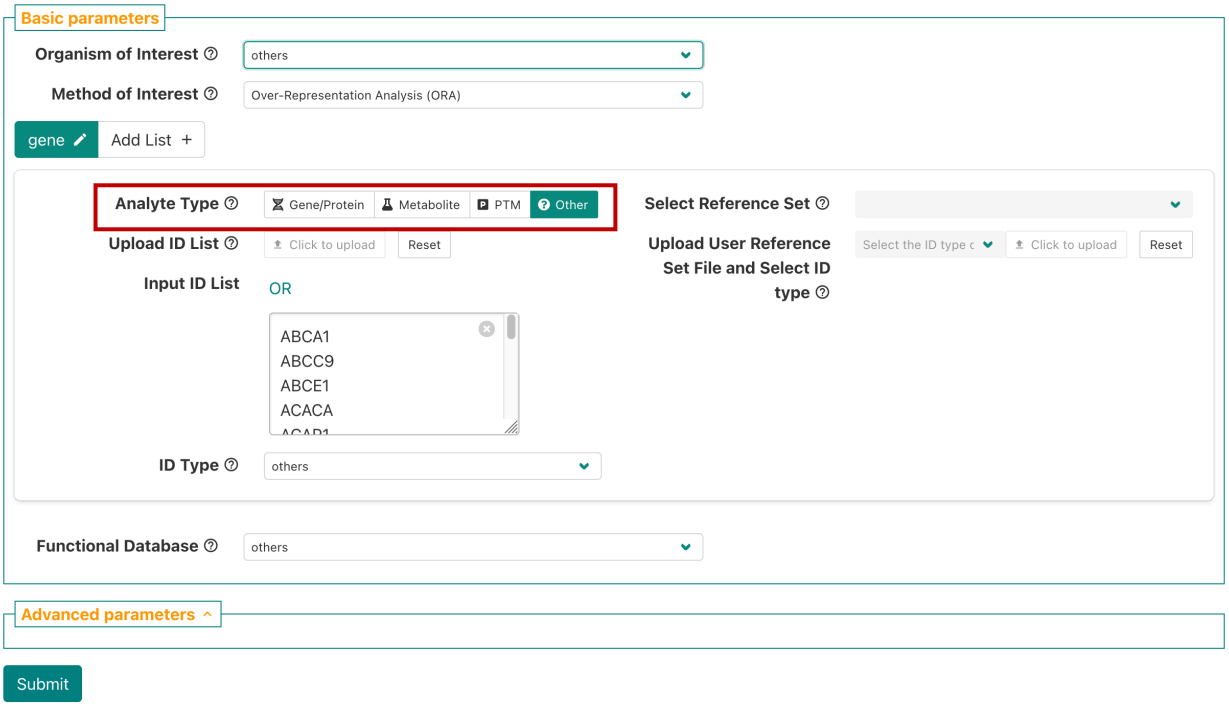


Figure 6. The “others” option for “Select Organism of Interest”.

Table1. An example of the “gmt” file

Geneset1	http://www.webgestalt.org/Geneset1	Gene11	Gene12	Gene13
Geneset2	http://www.webgestalt.org/Geneset2	Gene21	Gene22	
Geneset3	http://www.webgestalt.org/Geneset3	Gene31	Gene32	

Table 2. An example of the “des” file

Geneset1	The description of Geneset1
Geneset2	The description of Geneset2
Geneset3	The description of Geneset3

5.2 For supported organisms

If the user selects one of the 12 organisms and inputs their ID lists, there is a dropdown menu to show eight categories: geneontology, pathway, network, phenotype, disease, drug, chromosomal location and "**others**" option. **Except "others"**, after selecting one of the other seven classes, the detailed database name in the class will be shown in another dropdown menu (see Figure 7). For gene ontology categories, we used an algorithm to remove redundant terms and create three more databases only containing non-redundant terms. Due to running time consideration, the user can only use the non-redundant version for the **GSEA** GO enrichment analysis.

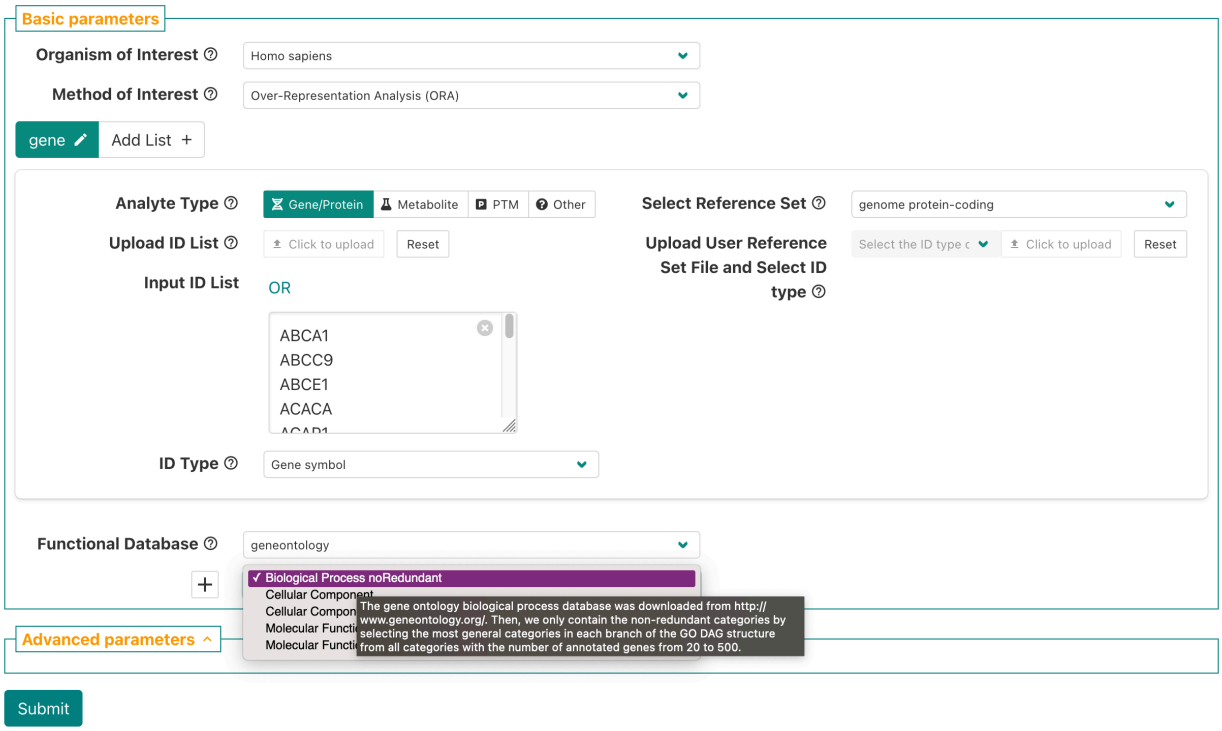


Figure 7. Selection of the functional database

If the user selects the “others” category, the user can upload a functional database not included in WebGestalt (see blue box in Figure 8). Because the user has selected one of 12 organisms instead of the “others” organism, WebGestalt will perform the ID mapping for all uploaded files (all ID type related options are still active (red box in Figure 8)), which means it is not necessary to upload the files with the same ID types. Thus, the user also needs to select the ID type of the uploaded functional database file.

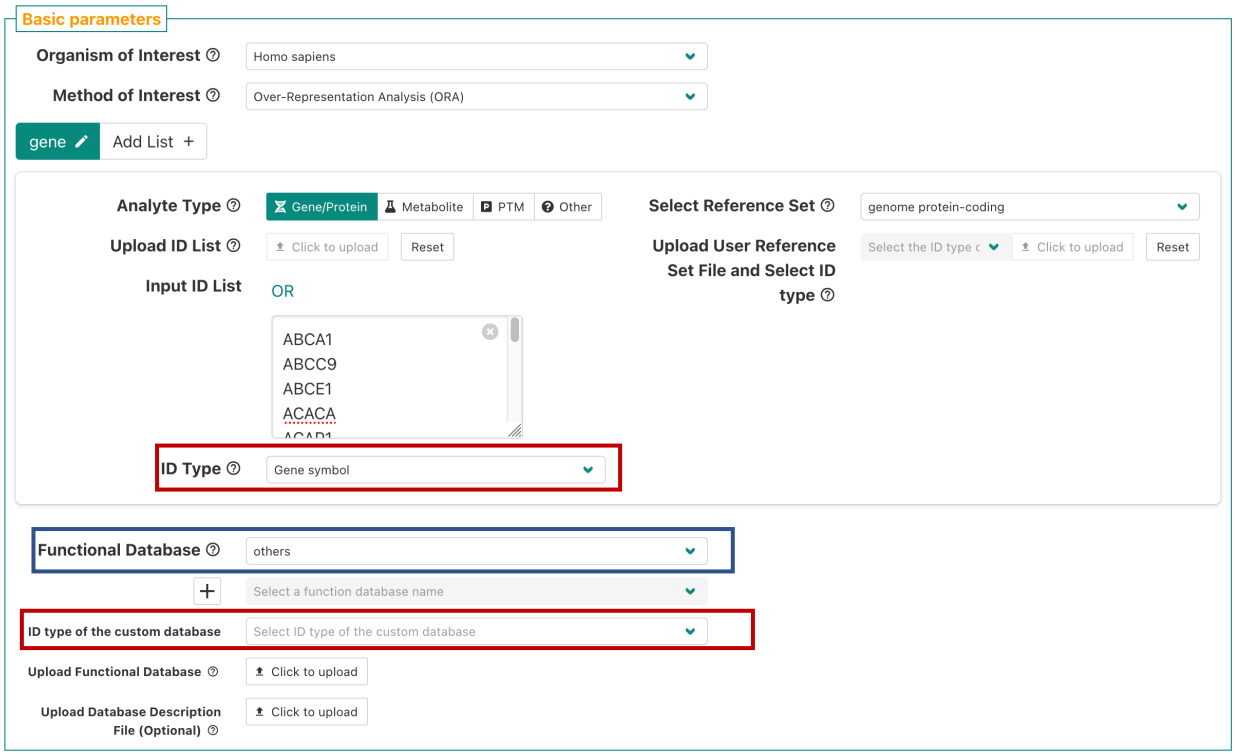


Figure 8. The “others” class selection of functional database

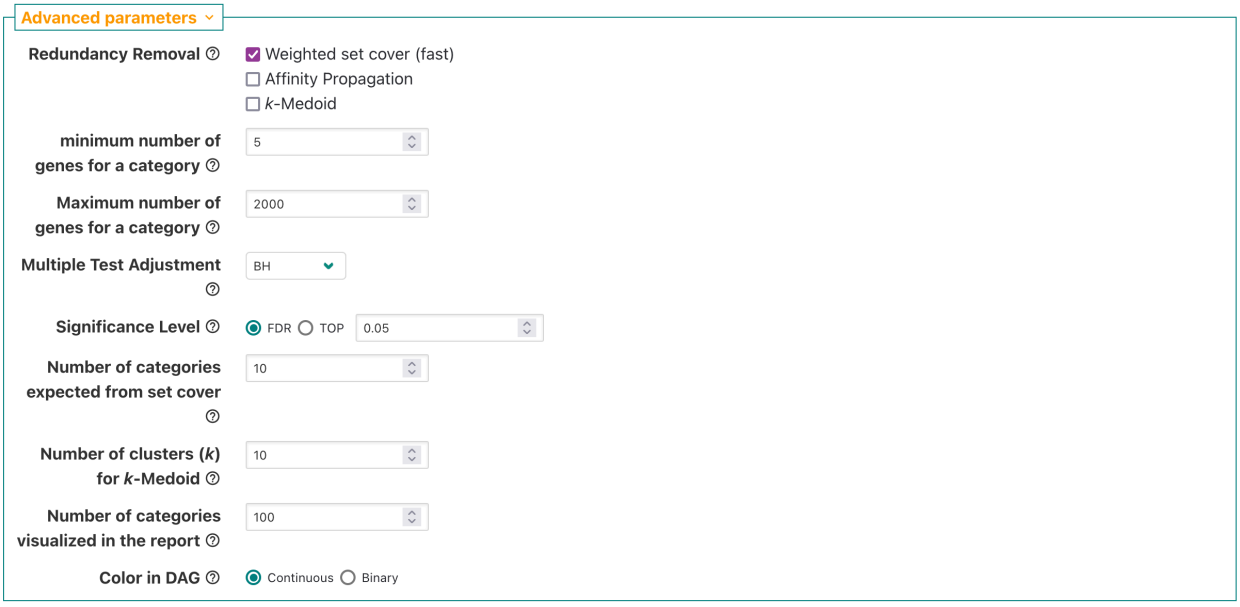
For the NTA method, users only need to select the networks from the dropdown menu.

6. Advanced parameters

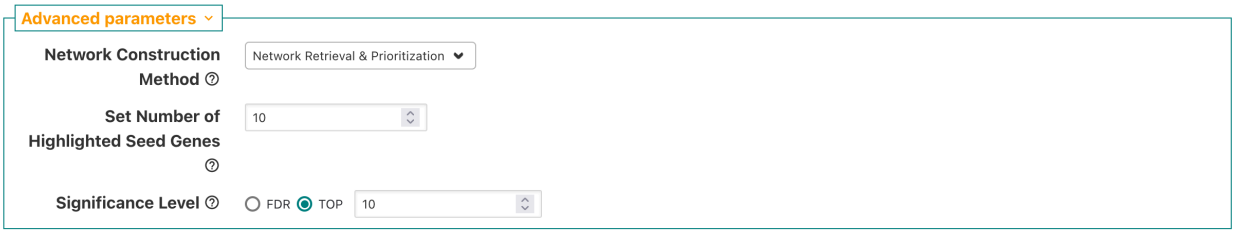
The user can also set some advanced parameters for different methods. The first panel in Figure 10 shows the advanced parameters for the ORA and GSEA methods.

- Setting the “Minimum Number of Genes for a Category” will remove the categories with sizes less than this number. The category size is calculated based on the number of overlapping genes between the annotated genes in the category and the reference gene list for the “ORA” method (or ranked gene list for the “GSEA” method).
- Setting the “Maximum Number of Genes for a Category” will remove the categories with sizes greater than this number.
- “Multiple Test Adjustment” will set the FDR method from “BH”, “BY”, “bonferroni", “Holm” and “Hommel”. This option is only for the “ORA” method.
- The “Significance Level” parameter has two options. “FDR” means the enriched categories will be identified based on the FDR threshold and “TOP” means the categories will be first ranked based on the FDR and then the top N most significant categories will be selected. **For the GSEA method, the “TOP” method will select the top N most significant categories from positive and negative related categories separately.**
  - For meta-analysis GSEA, only FDR is allowed.
- “Number of Permutations” and “Collapse Method” parameters are just for the “GSEA” method. These indicate how many permutations to perform and which
  method will be used for collapsing any duplicate IDs.
- “Number of categories expected from set cover” indicates the number of the expected reduced sets of the weighted set cover algorithm for redundancy reduction in the report. The algorithm stops either covering all the genes or reaching the number.
- “Number of categories visualized in the report” represents how many significant categories will be shown in the report. All the categories passing the significance level will can be obtained in the download result. **For the GSEA method, this number represents how many positive related categories or how many negative related categories will be visualized. If this number is 40, the report can at most contain 80 significant categories.**
- If “Color in DAG” is continuous, the categories of the DAG structure in the output report will be colored in gradient based on the FDR. Otherwise, WebGestalt will use steel blue color for ORA method or steel blue and dark orange color for GSEA method to color the significant categories.
- Redundancy Removal allows you to specify which methods you would like to run. More methods will require more computation time. By default, only weighted set cover is enabled.

ORA and GSEA



Network Retrieval and Prioritization



Network Expansion

Figure 9. Advanced parameters for all the methods

Figure 9 also shows the advanced parameters for the two NTA methods.

- NTA includes two network construction method: “Network Retrieval & Prioritization” and “Network Expansion”. “Network Retrieval & Prioritization” first uses random walk analysis to calculate random walk probability for the input

seeds, then identifies the relationships among the seeds in the selected network and returns a retrieval sub-network. The seeds with the top random walk probability are highlighted in the sub-network. “Network Expansion” first uses random walk analysis to rank all genes in the selected network based on their network proximity to the input seeds and then returns an expanded sub-network in which nodes are the input seeds and their top ranking neighbors and edges represent their relationships.

- For “Network Retrieval & Prioritization”, users need to set how many seeds will be highlighted in the retrieved sub-network. For “Network Expansion” method, users need to set how many top-ranking neighbors will be included in the expanded sub-network and also to set which type of genes will be highlighted in the sub-network (seeds or neighbors).
- These two methods finally will perform the enrichment analysis for the identified sub-networks based on the GO Biology Process ontology. Thus, users need to select the criteria for identifying the significant categories.

7. Output report for ORA or GSEA methods

7.1 Overview

if the ID type of the uploaded data is from one of 12 organisms, the output report will contain two major sections: “Summary” and “Enrichment Results”. Summary includes two folded sections of job parameters used in the analysis and the GO Slim summary, which contains three bar plots illustrating the number of genes in the uploaded gene list that overlap with the annotated genes in the GO Slim terms from biological process (red bar plot), cellular component (blue barplot) and molecular function (green bar plot) ontologies, respectively (see Figure 11). Clicking the “Result Download” link will download a zip file with the HTML report and text file of all the results.

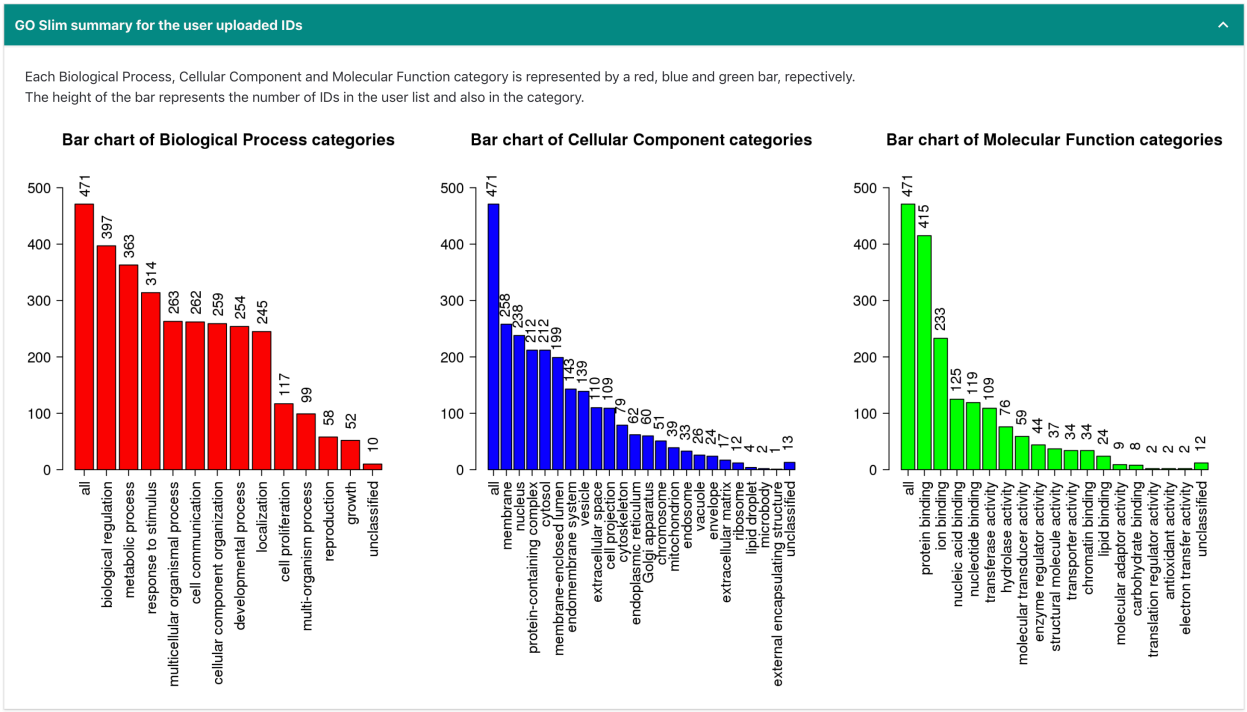


Figure 11. GO Slim summary of input genes

7.2 Enrichment Results

7.2.1 Result visualization

To illustrate the new visualizations and features in the 2019 version, we will use the **ORA sample run** directly available from the homepage as an example. For this sample run, we set the FDR threshold to 0.05 and allowed to visualize up to 100 enriched sets. Note the ORA sample run in the 2017 version only visualized the top 10 enriched sets.

The “Enrichment Results” section first has tabs of different visualizations of the significant enrichment result, followed by the detailed information of the current viewing gene set with scoring statistics, gene table etc. The visualizations include table summarization, bar chart and volcano plot (see Figure 12).

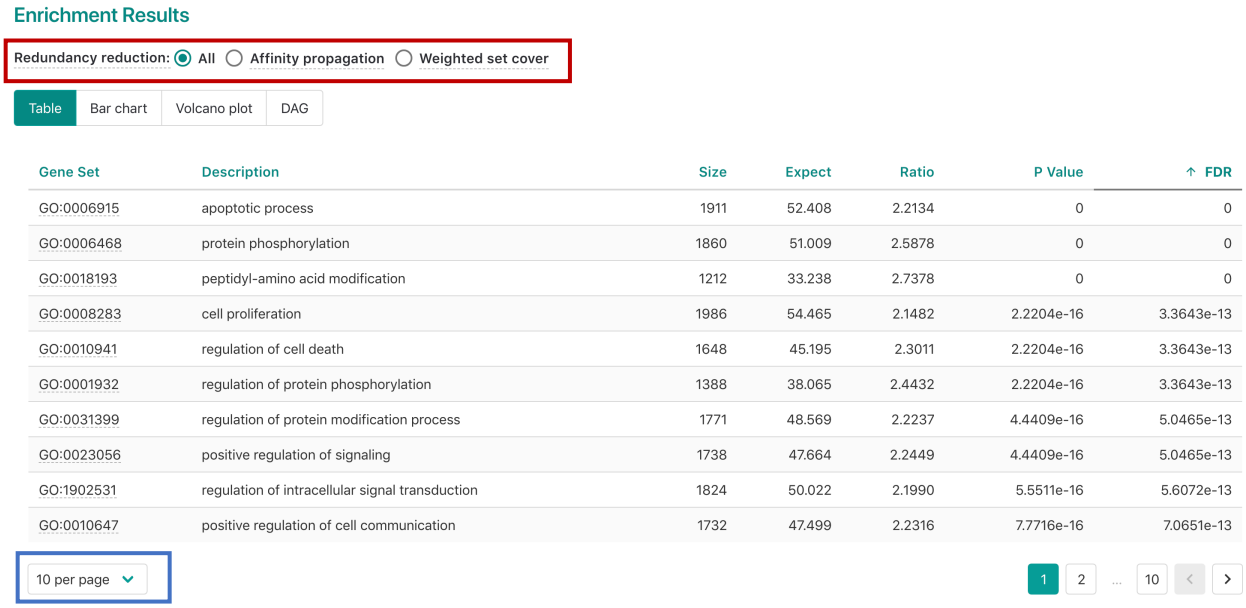


Figure 12. Visualizations, redundancy reduction and table of results in enrichment results section

The **table** concisely summarize the enriched functional categories with their statistics. It is similar to the returned value with the WebGestaltR package. The table can be sorted by the scores and statistics by clicking on the headers (Figure 12). The number of rows per page can be adjusted (blue box in Figure 12). Clicking the gene set name will pull up the detailed information about the category in the bottom section and scroll to it.

The **bar chart** plots the enrichment results vertically with the bar width equal to enrichment ratio in ORA and NES in GSEA (Figure 13). If there are negatively related categories in GSEA result, the chart will go in two directions with different colors. When the FDR for the categories is smaller than or equal to 0.05, the colors of the bar are darker, while the color for categories with FDR larger than 0.05 is in a lighter shade. This is useful when significant level is chosen as “Top”.

As for the ORA example, the bar chart can adjust its height, but it is still a very long graph (Figure 13). Another bar chart from a different GSEA run shows a two-sided bar chart (Figure 14). Right clicking on the plot will show download buttons to save it in SVG and PNG formats (Figure 14).

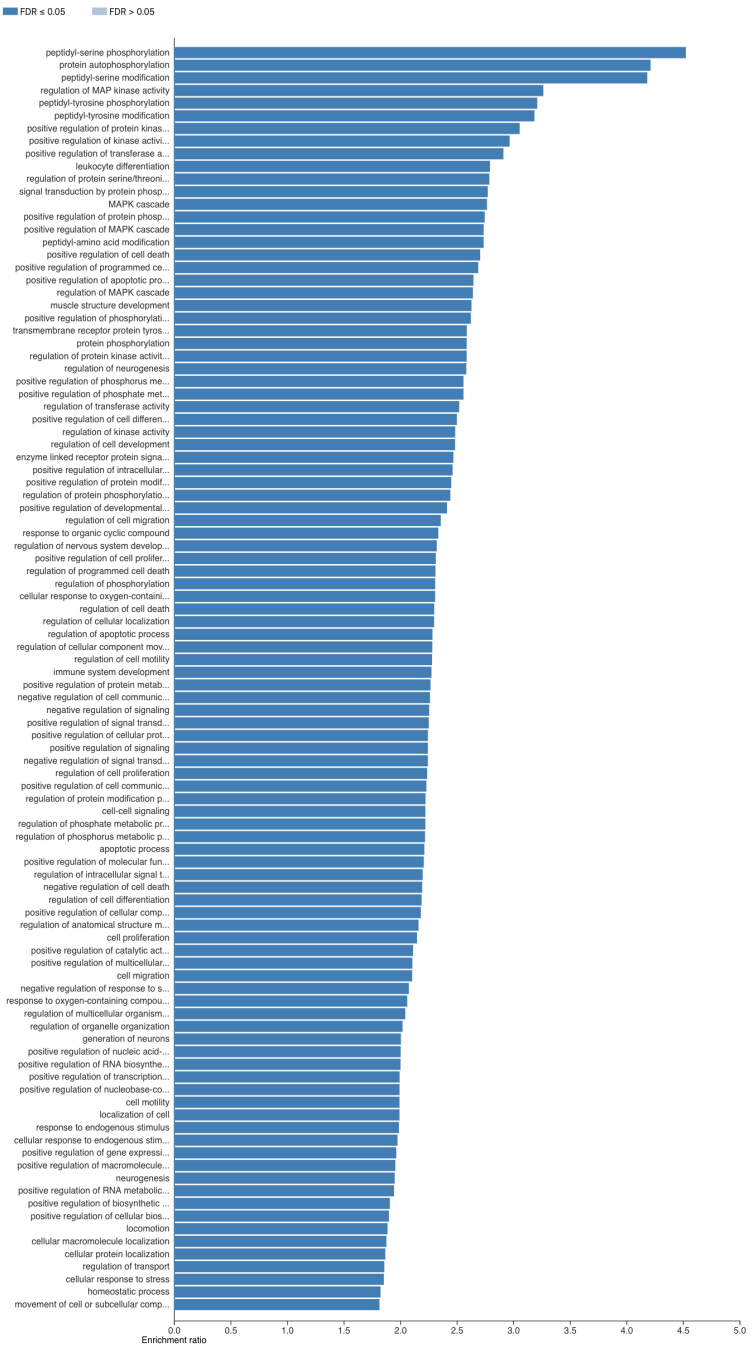


Figure 13. The bar chart for the ORA sample run with a lot of results

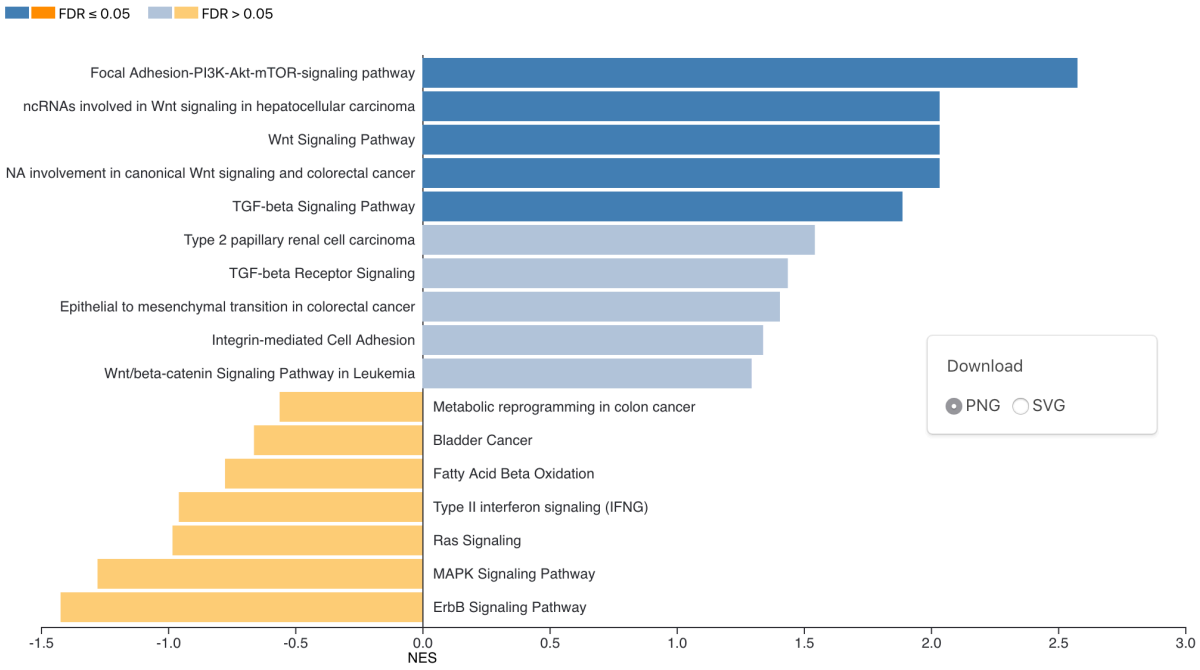


Figure 14. Another example of the Bar chart for a GSEA run

The **volcano plot** shows log of FDR against enrichment ratio or NES for all the categories in the search database (Figure 15). Significant categories will be near the upper corners. The size and color of the dot is proportional to the size of the category. Hovering over a dot will show some information about it and clicking on it will update the detailed information section. User can pan and zoom the plot after activating a switch button. The enriched categories are labeled, and the positions of the labels can be adjusted manually with mouse, since no perfect automatic positioning can always be achieved. The label can be changed to the gene set name and a linking line to the dot can be added with a button (Details about the customization are further discussed later in 7.2.3). User can download the plot in SVG or PNG formats after fine tuning in the bottom toolbox (blue box in Figure 15).

For the ORA sample run, all functional categories with FDR <= 0.05 are labelled, heavily overlapping with each other, and not suitable for publication.

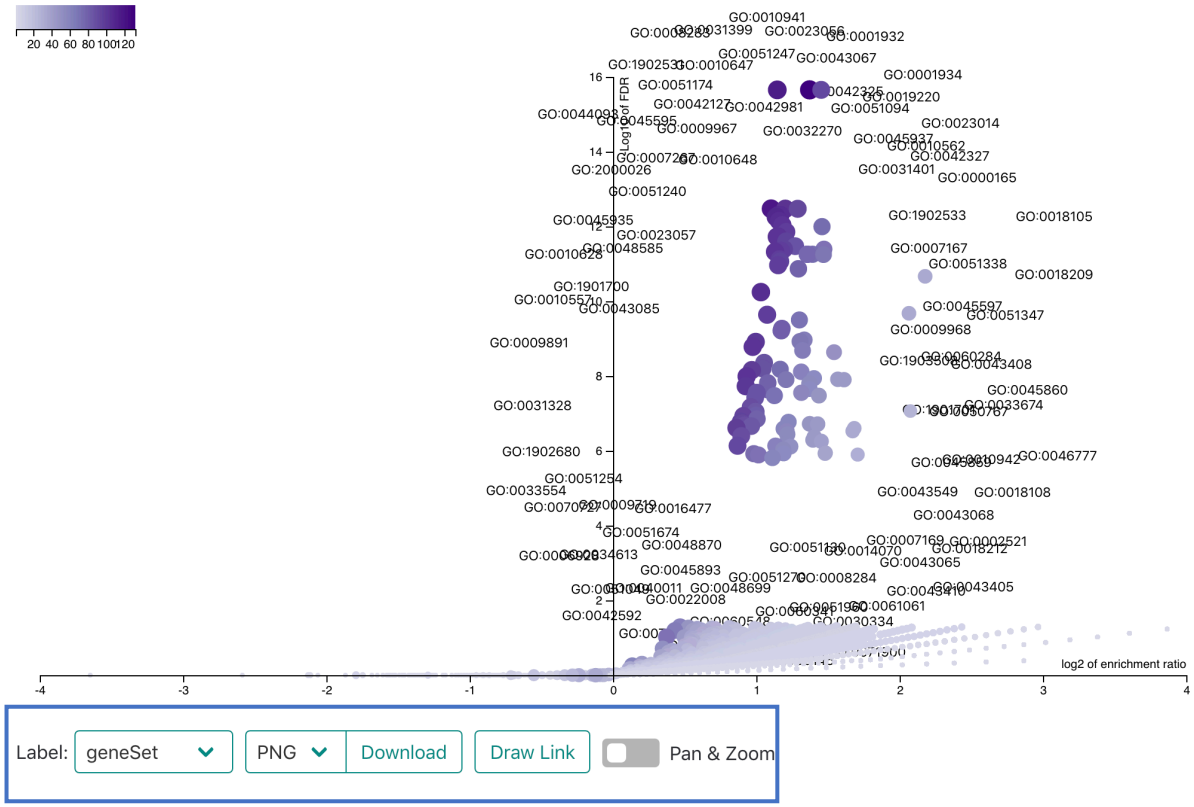


Figure 15. The volcano plot for the ORA sample run labelling categories with FDR <= 0.05

If the functional database contains the **DAG** or tree structure, such as GO terms, the structure will be visualized in another tab (Figure 16). If the “Color in DAG” is continuous, the significant categories will be colored by gradient based on the FDR. Otherwise, the categories will be colored by steel blue. For the GSEA method, the positively related categories will be colored by blue gradient while the negatively related categories will be colored by orange gradient if the “Color in DAG” is continuous. Otherwise, the positively and negatively related categories will be colored by steel blue and dark orange, respectively. The user can pan and zoom with mouse scroll or touchpad to navigate the DAG. Right clicking on the plot shows a menu with the options to resize and save the current view to a high-resolution PNG file. Clicking on the colored nodes (significant categories) will select the category to view the detailed information.

The sample ORA run is a typical situation when many diverse GO terms are identified and make the graph very wide and difficult to browse and present. The 2019 version utilizes full browser window width compared with old version, which only used half of the window width, but may still fall short in such cases.

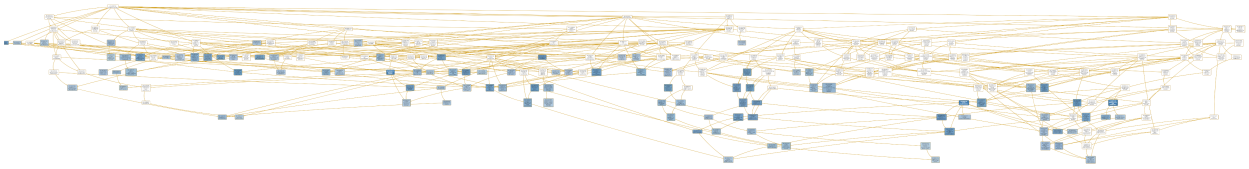


Figure 16. Diagram for functional categories with DAG structure

7.2.2 Redundancy reduction

The **redundancy reduction** selection above the visualization tabs (See red box in Figure 12) give users a reduced set of results to inspect, which is especially meaningful when the database is redundant, or the number of results is overwhelming. Two methods of redundancy reduction, affinity propagation and weighted set cover are run as a post-processing step to identify the most representative sets with low redundancy for visualization (manuscript under revision).

In short, affinity propagation clusters gene sets with Jaccard index as similarity measurement and automatically identify an “exemplar” or representative for each cluster with priority for set with significant P-value. Weighted set cover finds a minimum subset of gene sets that can cover all the genes from the enriched sets, while the weight or cost of adding a set is associated with its P-value. Weighted set cover may stop early before convergence when the input parameter limiting the expected set number is reached. Details of the algorithms could be found at the end.

Selecting either method will result in a subset of categories appearing in the visualization tabs. Detailed information about the clustering can be found in the download result, such as members in the affinity propagation cluster and gene coverage of the set cover.

Here we then show the corresponding visualizations of the subset from affinity propagation for the ORA sample run results below. There are much less gene sets in the reduced result, making it more manageable. Meanwhile, important biological themes are all covered with these selected gene sets. More information about gene sets relationship in affinity propagation and gene coverage of set cover can be found in the download result.

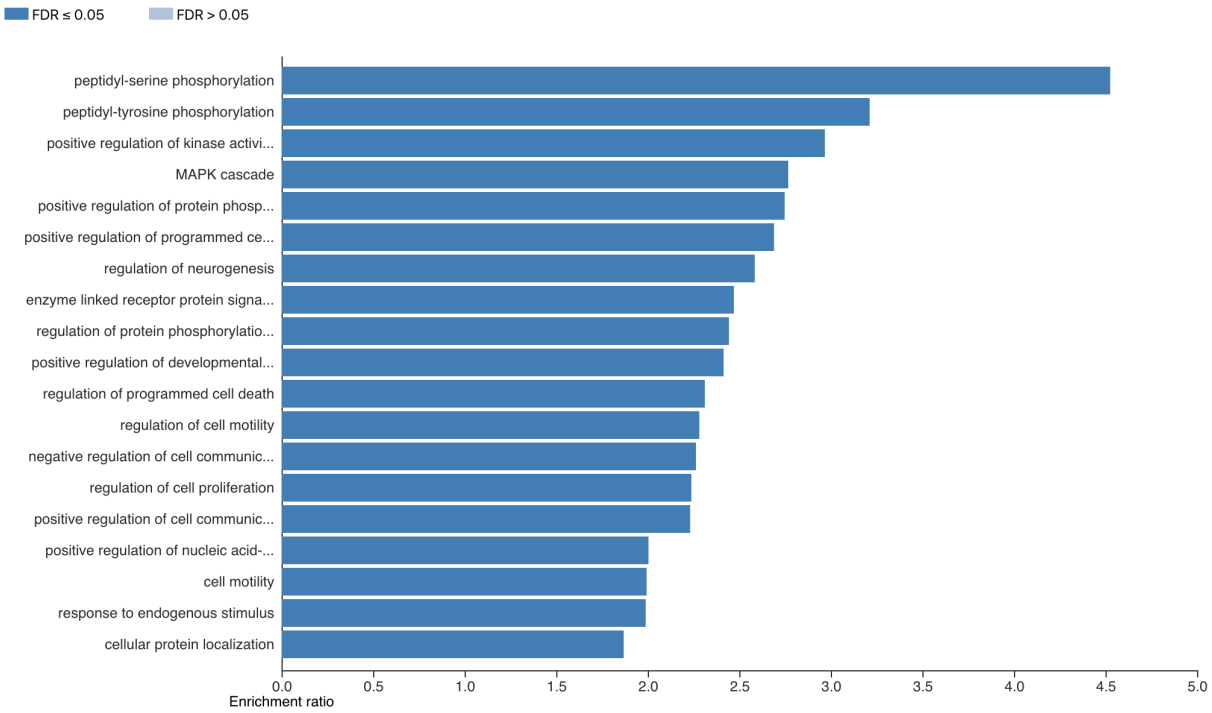


Figure 17. Bar chart of reduced sets of the ORA sample run from affinity propagation

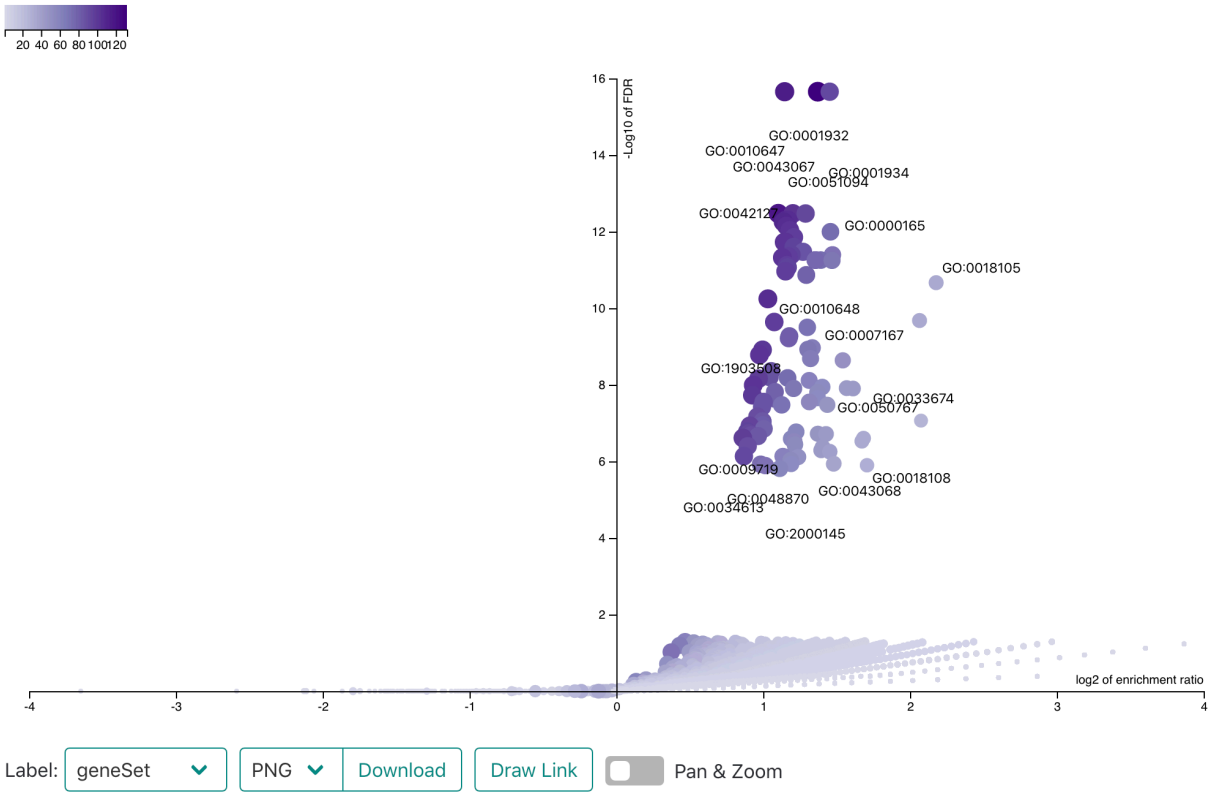


Figure 18. Volcano plot of reduced sets of the ORA sample run from affinity propagation

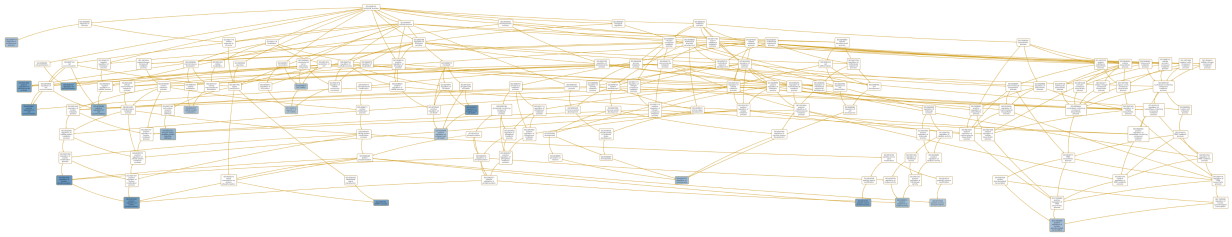


Figure 19. DAG of reduced sets of the ORA sample run from affinity propagation

7.2.3 Volcano plot customization

Volcano plot is implemented with interactivity and customization in mind. On startup, the labels of enrich sets are positioned automatically by a force field-based algorithm. User could further adjust label position by dragging the label around with mouse. The label text can be switched between set ID or the description, which is usually more meaningful but also could be too verbose sometimes. A line linking the label and its data point could be drawn by clicking the button in the bottom toolbox. The plot can be moved and zoomed in after activating the “Pan & zoom” switch in the toolbox. After fine adjustment, the plot can be saved in SVG or PNG formats with the download button.

Figure 20 shows the label positions after manual adjustment for the affinity propagation subset of the ORA sample run. The inset compares it with the default positions from Figure 18. Figure 21 demonstrates further changing label text to gene set description, and a high-resolution plot can be downloaded and used directly for presentation or publication.

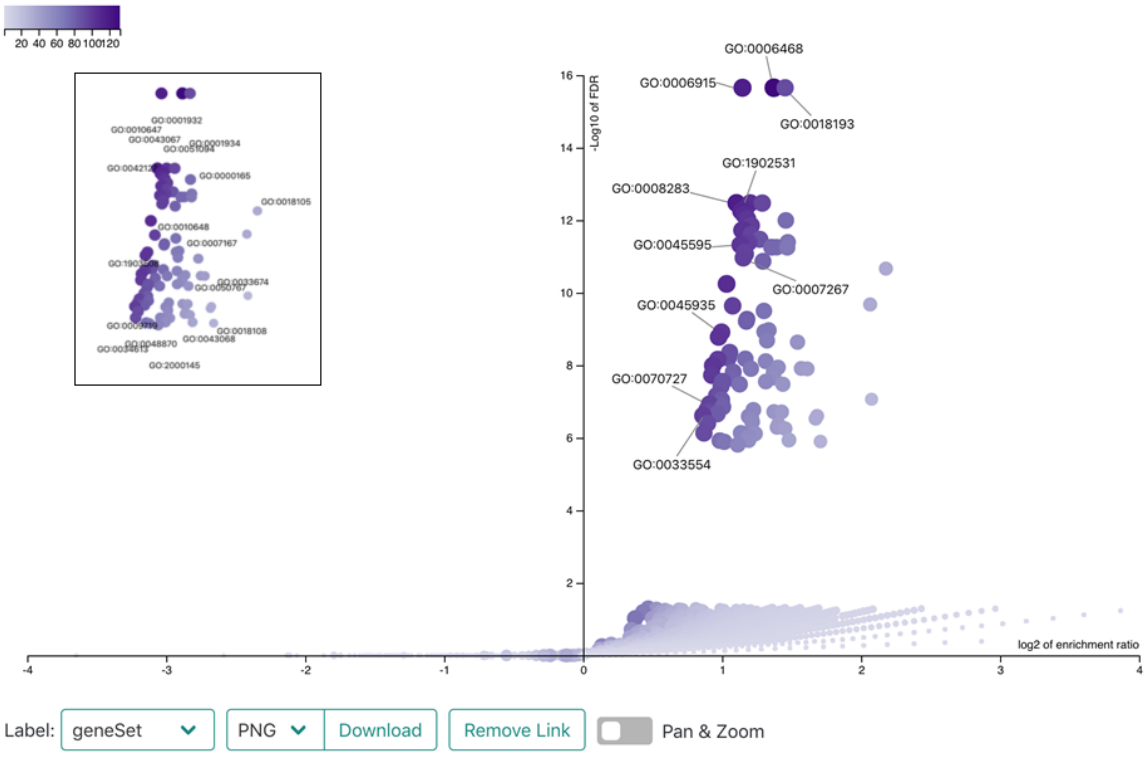


Figure 20. Manual adjustment of volcano plot label position

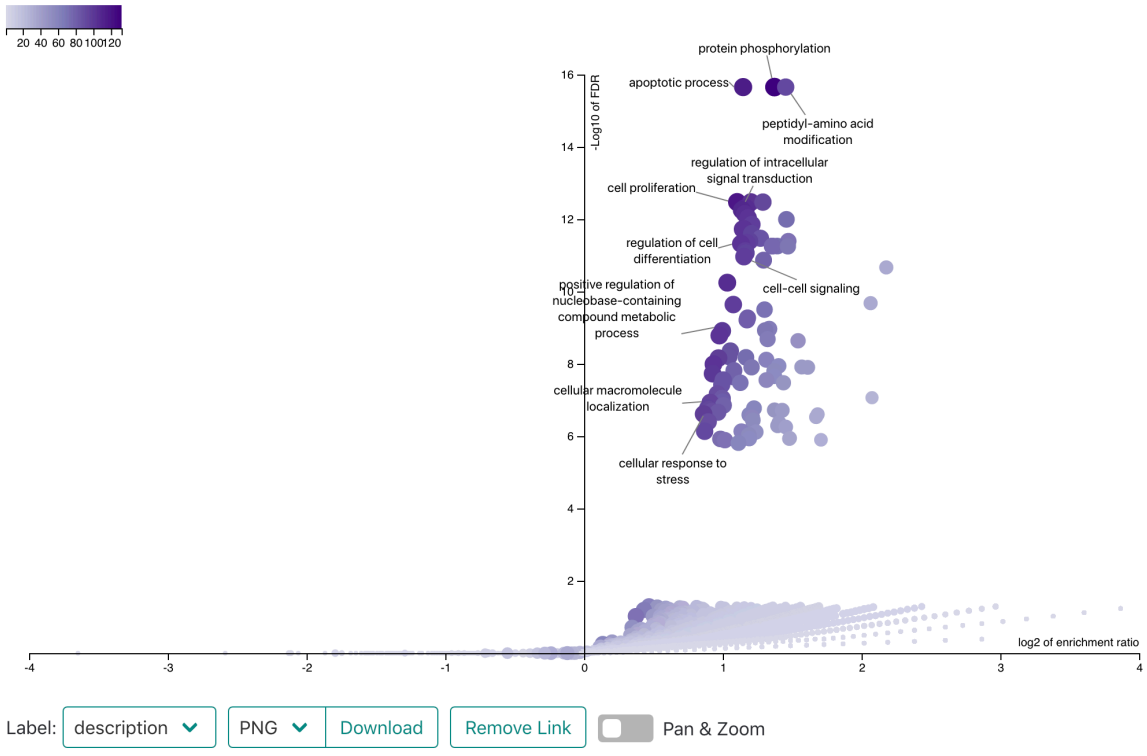


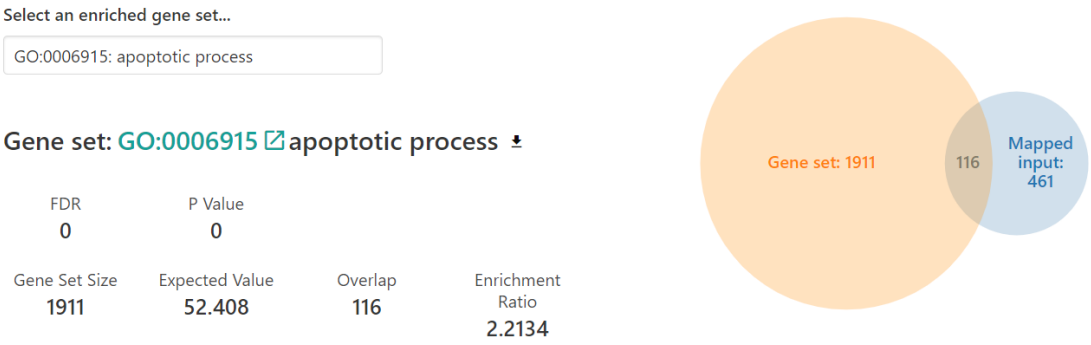
Figure 21. Switching label text to gene set description in volcano plot

7.3 Detailed information for individual enriched gene sets

The **detailed information** section contains scoring statistics and links to external database and downloading the gene table. The section can be updated to select category either by clicking on corresponding elements in the plots or directly typing or selecting through the select box (Blue box in Figure 22). The gene table lists the overlapping or leading edge genes with the gene symbol, name and links to NCBI and can be sorted by clicking headers. The number of rows per page can be adjusted. For ORA, there is a Venn diagram showing the overlapping between the genes in the input and those in the database (Figure 23). For GSEA, it is replaced by an enrichment plot showing the rank distribution, running sum, and where its peak is (Figure 22).

User ID	Score	Gene Symbol	Gene Name	Entrez Gene ID
14871	-0.655177561451468	Gstt1	glutathione S-transferase, theta 1	14871
14651	-0.661135105068991	Hagh	hydroxyacyl glutathione hydrolase	14651
12908	-0.70317323838164	Crat	carnitine acetyltransferase	12908
75475	-0.706364140164144	Oplah	5-oxoprolinase (ATP-hydrolysing)	75475
100039707	-0.726878668478243	Mthfsl	5, 10-methenyltetrahydrofolate synthetase-like	100039707

Figure 22. Detailed information section for GSEA results



User ID	Gene Symbol	Gene Name	Entrez Gene ID
ACIN1	ACIN1	apoptotic chromatin condensation inducer 1	22985
ADORA1	ADORA1	adenosine A1 receptor	134
AHI1	AHI1	Abelson helper integration site 1	54806
APC	APC	APC, WNT signaling pathway regulator	324

Figure 23. Detailed information section for ORA results

When the gene set has a network relationship between genes, such as TCGA co-expression network modules, there is also a button to show the network of the genes (Figure 24). This is reimplemented with Cytoscape.js library in the new version, removing the requirement of the obsolete Flash plugin.

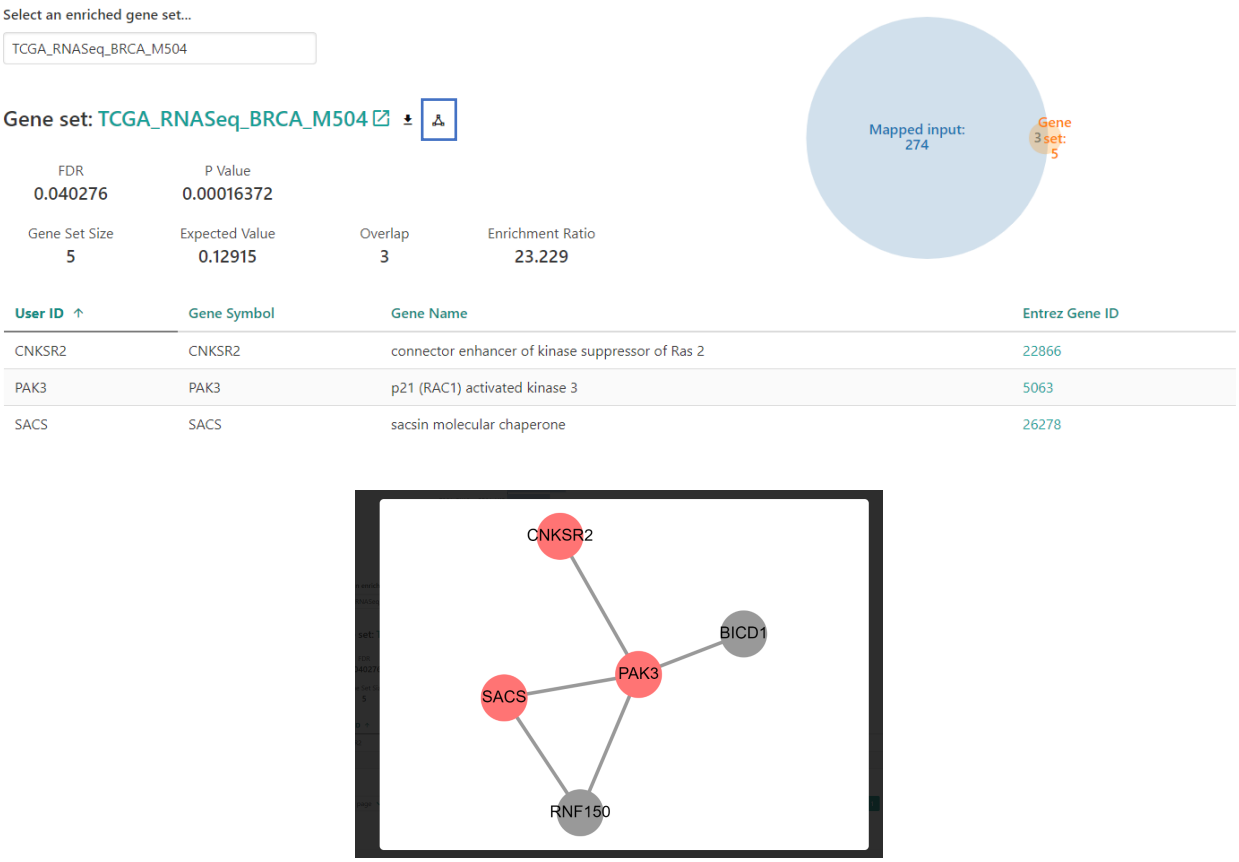


Figure 24. Gene network of some functional categories

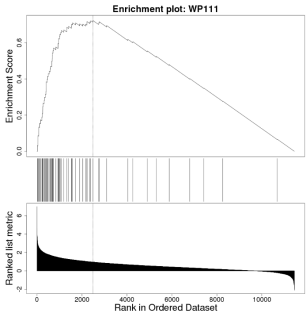
For searching against WikiPathway and KEGG databases, external links lead to their viewer of the pathway with overlapping/leading edge genes highlighted. For GSEA runs against WikiPathway, the scores for the genes are also used to color genes in gradient (Figure 25).

Select an enriched gene set...

WP111: Electron Transport Chain (OXPHOS sy

Gene set: WP111

Electron Transport Chain (OXPHOS system in mitochondria)



FDR	P Value
0	0

Size	Number of leading edge IDs	Enrichment Score	Normalized Enrichment Score
69	57	0.72396	1.9662

User ID	Score	Gene Symbol	Gene Name	Entrez Gene ID
NDUFA1	3.41383266545354	NDUFA1	NADH:ubiquinone oxidoreductase subunit A1	4694
NDUFS5	2.92884762596056	NDUFS5	NADH:ubiquinone oxidoreductase subunit S5	4725
SDHB	2.92884762596056	SDHB	succinate dehydrogenase complex iron sulfur subunit B	6390
COX6C	2.62419632911226	COX6C	cytochrome c oxidase subunit 6C	1345
NDUFB8	2.62419632911226	NDUFB8	NADH:ubiquinone oxidoreductase subunit B8	4714

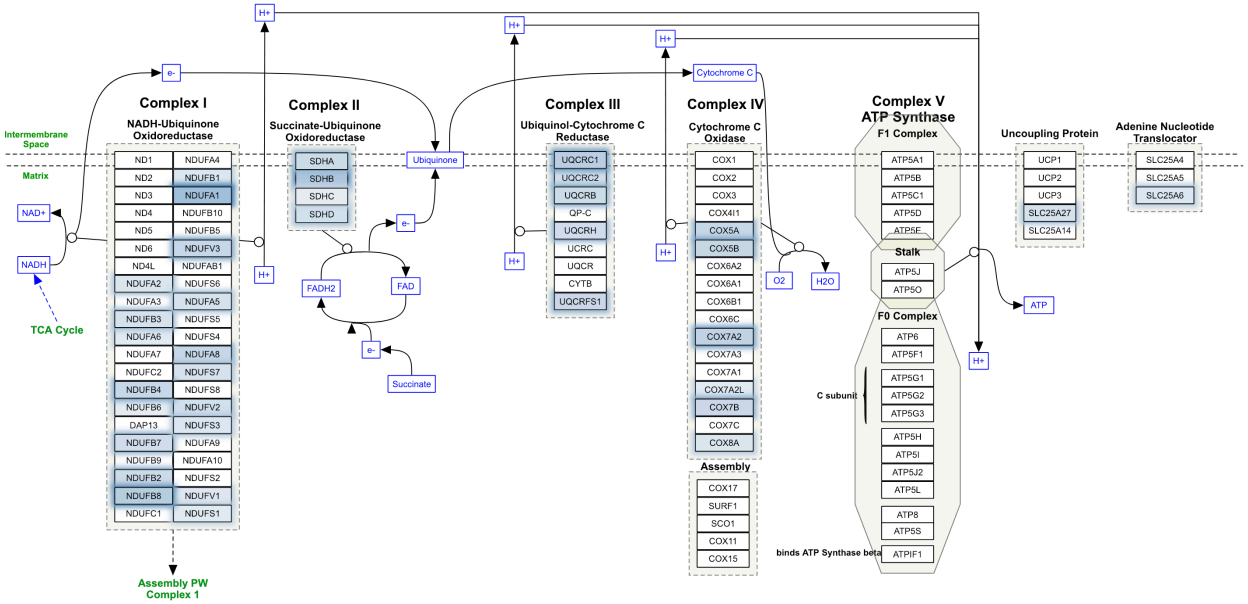
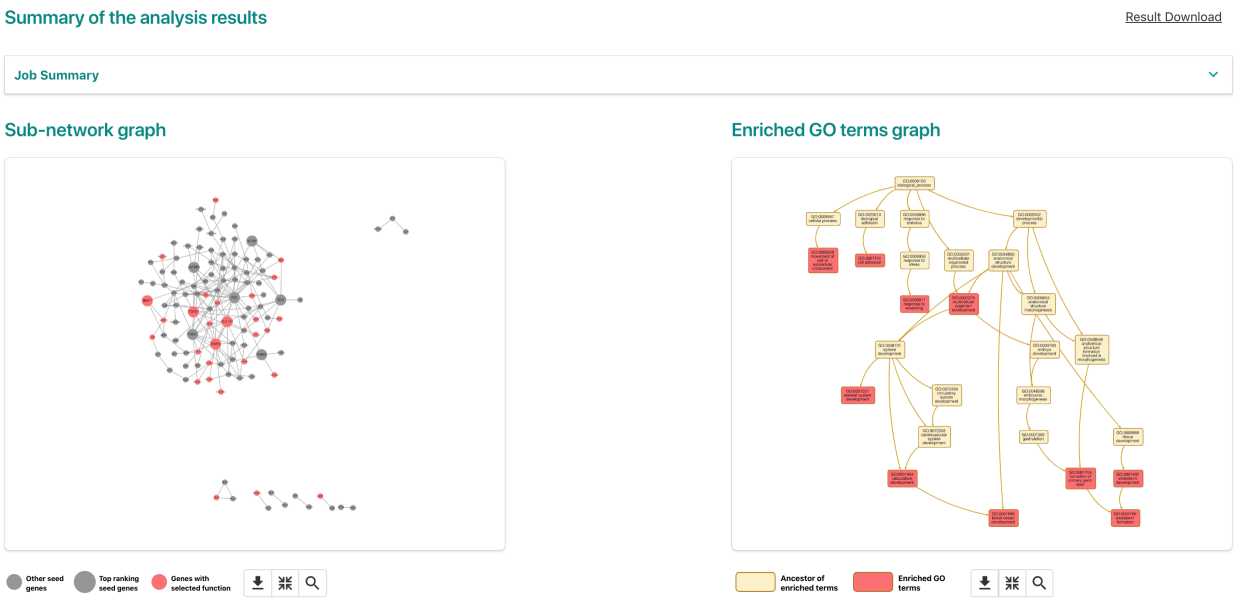


Figure 25. Link to WikiPathway viewer highlights genes with color gradient

For “other” organism, the result report will not have the GO Slim summary and gene table in detailed information section will only list user IDs for input genes. But bar char and volcano plot are still available, while the 2017 version just had a plain table of results.

8 Output report for NTA method

The NTA result page similarly starts with the job summary (Figure 26). The two side-by-side graphs show the retrieved or expanded network on the left and DAG for GO Biological Process enrichment results. Below are several tables of seeds and information about enriched GO terms. Clicking on the red box in the DAG or in the flag icon in the table will zoom to the DAG node and color the corresponding genes in the network. Genes in the enriched GO category can be viewed by clicking the list icon. The result can be downloaded through the “Result Download” link including the HTML report and text files. The current view of the network and DAG can be saved as PNG files.



Detailed Information

Seeds in the sub-network

Gene Symbol
FN1
MMP2
TGFB1
THBS1
ELN
COL1A1
VCAN
TGM2
TIMP1
IGFBP3

Ranked Seeds

Gene Symbol	Random Walk Probability
FN1	4.99E-03
MMP2	3.68E-03
TGFB1	3.49E-03
THBS1	3.48E-03
ELN	3.46E-03
COL1A1	3.42E-03
VCAN	3.34E-03
TGM2	3.24E-03
TIMP1	3.22E-03
IGFBP3	3.14E-03

Enriched GO categories (Top 10 categories)

GO ID	GO Name	C	O	Raw P-value	Adjusted P-value	Seed & Neighbors
GO:0001501	skeletal system development	459	28	0	0	
GO:0001568	blood vessel development	609	39	0	0	
GO:0001704	formation of primary germ layer	109	17	0	0	
GO:0001706	endoderm formation	48	13	0	0	
GO:0001944	vasculature development	635	40	0	0	
GO:0006928	movement of cell or subcellular component	1802	55	0	0	
GO:0007155	cell adhesion	1239	54	0	0	

Figure 26. NTA result page

9 API for third party applications

9.1 Web API

WebGestalt 2024 provides an API for forwarding data from other websites for analysis. The parameters below can be sent via GET/POST methods to www.webgestalt.org/option.php for loading data in the submission interface. It may be used to load some data like gene list and then let users choose other parameters for example functional database. Similarly, all required parameters can be sent via POST to www.webgestalt.org/process.php to perform an analysis. This is basically the programmatic way of the same form submission through the web interface.

- **Organism**: athaliana, btaurus, celegans, cfamiliaris, dmelanogaster, drerio, ggallus, hsapiens, mmusculus, rnorvegicus, scerevisiae, sscrofa, others
- **enrich_method**: ORA, GSEA, NTA
- **enriched_database_category**: geneontology, pathway, network, disease, drug, phenotype, chromosomalLocation, community-contributed, others
- **enriched_database_name**: Depends on *enriched_database_category*
- **id_type**: ID type of the gene list
- **gene_list**
- **ref_set**: reference set (only supported now)
- **min_num**: minimum number of genes for a category
- **max_num**: maximum number of genes for a category
- **fdr_method**
- **sig_method**: top or fdr
- **sig_value**: Example: Top 10 or FDR 0.05
- **per_num**: permutation number for GSEA
- **collapse_method**: method to collapse duplicated genes for GSEA
- **set_cover_num**: number of sets when set cover algorithm stops
- **report_num**: Number of categories visualized in the report
- **color_scheme**: binary or continuous

Example:
http://www.webgestalt.org/option.php?organism=hsapiens&enrich_method=ORA&fdr_method=BY&enriched_database_category=geneontology&enriched_database_name=Biological_Process_noRedundant&sig_method=top&sig_value=0.01&max_num=200&id_type=entrezgene&gene_list=BMP2%0AAPC&ref_set=genome

9.2 R API

WebGestalt 2024 also comes with a matched WebGestaltR package, which is available on CRAN. The package provides an R interface for running large batch of jobs and integrating into other pipelines. Below lists the documentation of the main function for enrichment analysis.

WebGestaltR: The R interface for enrichment analysis with WebGestalt.

Description: Main function for enrichment analysis

Usage: WebGestaltR(enrichMethod = "ORA", organism = "hsapiens", enrichDatabase = "geneontology_Biological_Process", enrichDatabaseFile = NULL, enrichDatabaseType = NULL, enrichDatabaseDescriptionFile = NULL, interestGeneFile = NULL, interestGene = NULL, interestGeneType = NULL, collapseMethod = "mean", referenceGeneFile = NULL, referenceGene = NULL, referenceGeneType = NULL, referenceSet = NULL, minNum = 10, maxNum = 500, sigMethod = "fdr", fdrMethod = "BH", fdrThr = 0.05, topThr = 10, reportNum = 20, perNum = 1000, isOutput = TRUE, outputDirectory = getwd(), projectName = NULL, dagColor = "continuous", setCoverNum = 10, networkConstructionMethod = NULL, neighborNum = 10, highlightType = "Seeds", highlightSeedNum = 10, nThreads = 1, hostName = "http://www.webgestalt.org/", ...) WebGestaltRBatch(interestGeneFolder = NULL, enrichMethod = "ORA", isParallel = FALSE, nThreads = 3, ...)

Arguments
enrichMethod Enrichment methods: ORA, GSEA or NTA.
organism Currently, WebGestaltR supports 12 organisms. Users can use the function listOrganism to check available organisms. Users can also input others to perform the enrichment analysis for other organisms not supported by WebGestaltR. For other organisms, users need to provide the functional categories, interesting list and reference list (for ORA method). Because WebGestaltR does not perform the ID mapping for the other organisms, the above data should have the same ID type.
enrichDatabase The functional categories for the enrichment analysis. Users can use the function listGeneSet to check the available functional databases for the selected organism. Users can also input others to provide a custom functional databases not supported by WebGestaltR for the selected organism.
enrichDatabaseFile If users set organism as others or set enrichDatabase as others, users need to provide a GMT file as the functional category for enrichment analysis. The WebGestaltR 13 extension of the file should be gmt and the first column of the file is the category ID, the second one is the external link for the category. Genes annotated to the category are from the third column. All columns are separated by tabs.
enrichDatabaseType If users set enrichDatabase as others, WebGestaltR will also perform ID mapping for the supplied GMT file. Thus, users need to set the ID type of the genes in the enrichDatabaseFile. If users set organism as others, users do not need to set this ID type because WebGestaltR will not perform ID mapping for other organisms. The supported ID types of WebGestaltR for the selected organism can be found by the function listIdType.
enrichDatabaseDescriptionFile Users can also provide a description file for the custom enrichDatabaseFile. The extension of the description file should be des. The description file contains two columns: the first column is the category ID that should be exactly the same as the category ID in the custom enrichDatabaseFile and the second column is the description of the category. All columns are separated by tabs.
interestGeneFile If enrichMethod is ORA or NTA, the extension of the interestGeneFile should be txt and the file can only contain one column: the interesting gene list. If enrichMethod is GSEA, the extension of the interestGeneFile should be rnk and the file should contain two columns separated by tab: the gene list and the corresponding scores. interestGene Users can also use an R object as the input. If enrichMethod is ORA or NTA, interestGene should be an R vector object containing the interesting gene list. If enrichMethod is GSEA, interestGene should be an R data.frame object containing two columns: the gene list and the corresponding scores.
interestGeneType The ID type of the interesting gene list. The supported ID types of WebGestaltR for the selected organism can be found by the function listIdType. If the organism is others, users do not need to set this parameter. collapseMethod The

method to collapse duplicate IDs with scores. mean, median, min and max represent the mean, median, minimum and maximum of scores for the duplicate IDs.
referenceGeneFile For the ORA method, the users need to upload the reference gene list. The extension of the referenceGeneFile should be txt and the file can only contain one column: the reference gene list.
referenceGene For the ORA method, users can also use an R object as the reference gene list. referenceGene should be an R vector object containing the reference gene list.
referenceGeneType The ID type of the reference gene list. The supported ID types of WebGestaltR for the selected organism can be found by the function listIdType. If the organism is others, users do not need to set this parameter.
referenceSet Users can directly select the reference set from existing platforms in WebGestaltR and do not need to provide the reference set through referenceGeneFile. All 14 WebGestaltR existing platforms supported in WebGestaltR can be found by the function listReferenceSet. If referenceGeneFile and refereneceGene are NULL, WebGestaltR will use the referenceSet as the reference gene set. Otherwise, WebGestaltR will use the user supplied reference set for enrichment analysis.
minNum WebGestaltR will exclude the categories with the number of annotated genes less than minNum for enrichment analysis. The default is 10.
maxNum WebGestaltR will exclude the categories with the number of annotated genes larger than maxNum for enrichment analysis. The default is 500.
sigMethod Two methods of significance are available in WebGestaltR: fdr and top. fdr means the enriched categories are identified based on the FDR and top means all categories are ranked based on FDR and then select top categories as the enriched categories. The default is fdr.
fdrMethod For the ORA method, WebGestaltR supports five FDR methods: holm, hochberg, hommel, bonferroni, BH and BY. The default is BH.
fdrThr The significant threshold for the fdr method. The default is 0.05.
topThr The threshold for the top method. The default is 10.
reportNum The number of enriched categories visualized in the final report. The default is 20. A larger reportNum may be slow to render in the report.
perNum The number of permutations for the GSEA method. The default is 1000.
isOutput If isOutput is TRUE, WebGestaltR will create a folder named by the projectName and save the results in the folder. Otherwise, WebGestaltR will only return an R data.frame object containing the enrichment results. If hundreds of gene list need to be analyzed simultaneously, it is better to set isOutput to FALSE. The default is TRUE.
outputDirectory The output directory for the results.
projectName The name of the project. If projectName is NULL, WebGestaltR will use time stamp as the project name.
dagColor If dagColor is binary, the significant terms in the DAG structure will be colored by steel blue for ORA method or steel blue (positive related) and dark orange (negative related) for GSEA method. If dagColor is continous, the significant terms in the DAG structure will be colored by the color gradient based on corresponding FDRs.
setCoverNum The number of expected gene sets after set cover to reduce redundancy. It could get fewer sets if the coverage reaches 100%. The default is 10.
networkConstructionMethod Netowrk construction method for NTA. Either Network_Retrieval_Prioritization or Network_Expansion. Network Retrieval & Prioritization first uses random walk analysis to calculate random walk probabilities for the input seeds, then identifies the relationships among the seeds in the selected network and returns a retrieval sub-network. The seeds with the top random walk probabilities are highlighted in the sub-network. Network Expansion first uses random walk analysis to rank all genes in the selected network based on their network proximity to the input seeds and then return an expanded sub-network in which nodes

are the input seeds and their top ranking neighbors and edges represent their relationships.
neighborNum The number of neighbors to include in NTA Network Expansion method.
highlightType The type of nodes to highlight in the NTA Network Expansion method, either Seeds or Neighbors.
highlightSeedNum The number of top input seeds to highlight in NTA Network Retrieval & Prioritizaiton method.
nThreads The number of cores to use for GSEA and set cover, and in batch function.
hostName The server URL for accessing data. Mostly for development purposes.

Details WebGestaltR function can perform three enrichment analyses: ORA (Over-Representation Analysis) and GSEA (Gene Set Enrichment Analysis).and NTA (Network Topology Analysis). Based on the user-uploaded gene list or gene list with scores, WebGestaltR function will first map the gene list to the entrez gene ids and then summarize the gene list based on the GO (Gene Ontology) Slim. After performing the enrichment analysis, WebGestaltR function also returns a user-friendly HTML report containing GO Slim summary and the enrichment analysis result. If functional categories have DAG (directed acyclic graph) structure or genes in the functional categories have network structure, those relationship can also be visualized in the report.

Value The WebGestaltR function returns a data frame containing the enrichment analysis result and also outputs an user-friendly HTML report if isOutput is TRUE. The WebGestaltRBatch function returns a list of enrichment results.

Examples

```
####### ORA example #########
geneFile <- system.file("extdata", "interestingGenes.txt", package="WebGestaltR")
refFile <- system.file("extdata", "referenceGenes.txt", package="WebGestaltR")
outputDirectory <- getwd()
enrichResult <- WebGestaltR(enrichMethod="ORA", organism="hsapiens",
enrichDatabase="pathway_KEGG", interestGeneFile=geneFile,
interestGeneType="genesymbol", referenceGeneFile=refFile,
referenceGeneType="genesymbol", isOutput=TRUE, outputDirectory=outputDirectory,
projectName=NULL)

####### GSEA example #########
rankFile <- system.file("extdata", "GeneRankList.rnk", package="WebGestaltR")
outputDirectory <- getwd()
enrichResult <- WebGestaltR(enrichMethod="GSEA", organism="hsapiens", 16
weightedSetCover enrichDatabase="pathway_KEGG", interestGeneFile=rankFile,
interestGeneType="genesymbol", sigMethod="top", topThr=10, minNum=5,
outputDirectory=outputDirectory)

####### NTA example #########
enrichResult <- WebGestaltR(enrichMethod="NTA", organism="hsapiens",
enrichDatabase="network_PPI_BIOGRID", interestGeneFile=geneFile,
interestGeneType="genesymbol", sigMethod="top", topThr=10, outputDirectory=getwd(),
highlightSeedNum=10, networkConstructionMethod="Network_Retrieval_Prioritization")
```

Appendix. Algorithm details of redundancy reduction

Affinity Propagation

Given a list of input gene sets, the affinity propagation algorithm (20) clusters similar gene sets into groups and identifies one exemplar that best represents each group. The affinity propagation algorithm splits the gene sets such that each partition is associated with its most representative gene set (called "exemplar"). At the same time, the sum of the user-defined similarity measures between gene sets and their exemplars and the exemplars' prior preference to be designated as exemplars is maximized. This is achieved by simultaneously considering all gene sets as potential exemplars and exchanging messages between gene sets until a satisfying set of exemplars and clusters emerges. The algorithm takes as input a real valued similarity matrix M where M_{ij} implies the appropriateness of selecting gene set j to be the exemplar for gene set i. We use the following formula to set M:

$$m_{ij} = \begin{cases} \text{Jaccard}(i,j) & \text{if Jaccard}(i,j) > 0 \\ -\infty & \text{if Jaccard}(i,j) = 0 \end{cases}$$

In other words, if a pair of gene sets i and j overlap, we set its Jaccard distance as its similarity. Otherwise, its similarity is set to $-\infty$. Intuitively, it is not appropriate for gene set i to represent gene set j if they do not overlap. Besides the similarity matrix, another important parameter of the algorithm is the input preference. This can be interpreted as the suitability of a gene set to become an exemplar. The input preference can either be customized for each gene set or it can be set to the same value shared among all gene sets. It has the effect of adjusting the graininess of the resulting clusters where large input preference values can result in a large number of clusters. We use the following procedure to set the preference values. Assume that the gene set enrichment significance levels, i.e., $-\log(p\text{-value})$, are in the range of $[p_{min}, p_{max}]$ and let m_{med} denote the median of all finite values in the similarity matrix M. We set the maximum of preference to m_{med} and the minimum to 0. For gene set i, its input preference is interpolated linearly as:

$$ip_i = \frac{m_{med}(x - p_{min})}{p_{max} - p_{min}}, \text{ where } x = -\log(p\text{-value}_i),$$

For gene sets with a large significance level, it is preferable to increase its tendency to be selected as an exemplar. The R package "apcluster" is used for the heavy lifting.

Weighted Set Cover

Given a universe of finite set U with $|U| = n$ and sets $C = \{C_1, \ldots, C_m\} \subseteq U$, a set cover is a collection S of some of the sets from C whose union is the entire universe. Here, C corresponds to all gene sets of interest, and U corresponds to the union of all genes

within these gene sets. We consider a generalized version of weighted set cover and maximum coverage called size-constrained weighted set cover where each set is also associated with a weight w_i, which is calculated as $-\log(p\text{-value})$. Therefore, higher weights are assigned to gene sets with smaller enrichment p values. The input to the problem further includes a size constraint k . The goal is to find S, a sub-collection of up to k sets, whose sum of weights is maximal and whose union covers as many elements as possible. Assume that one or more sets have been selected into S. We denote the marginal benefit set of a candidate set s given S, $B_m(s,S)$ as the set of elements from U covered by s but have not yet been covered by any set in S. In addition, we define the marginal gain of selecting s into S as $G_m(s,S) = |B_m(s,S)|w_s$. The algorithm starts with computing the marginal benefit set for all sets in C. Next, it selects the set with maximal marginal gain and adds it to the solution S. The algorithm then updates the marginal benefit set of the remaining candidate sets and removes those candidates with empty marginal benefit set before repeating the selection step. The algorithm returns as soon as it has covered all elements in U or after k iterations. It outputs the selected sets and fraction of coverage $\hat{s}$.