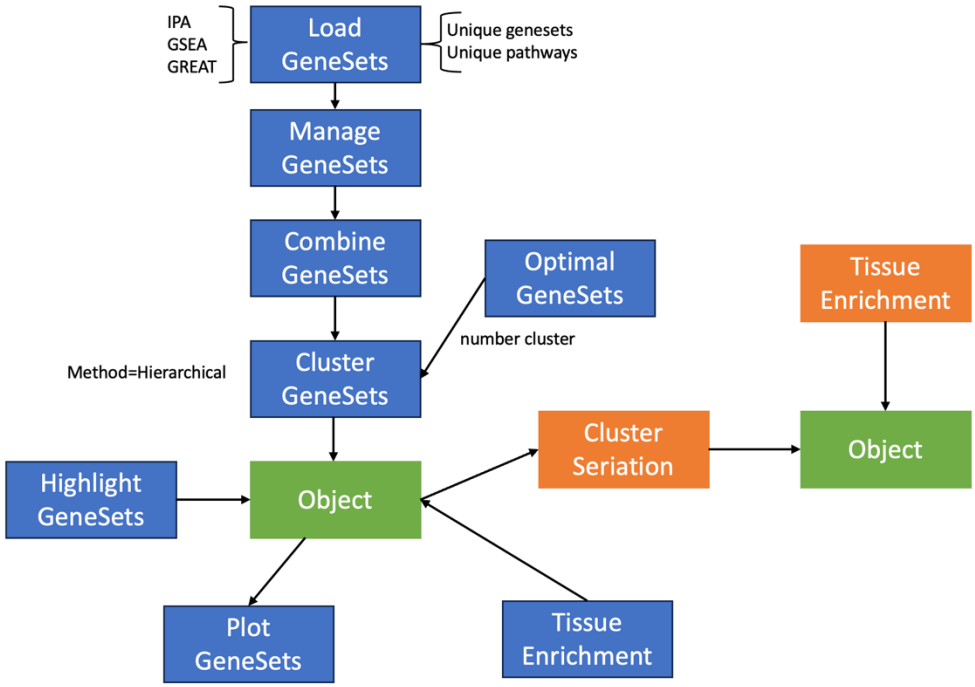
GeneSetCluster Shiny User Guide

By Alberto Maillo, Asier Ortega-Legarreta and Ewoud Ewing

1. General

This is a schematic representation of the workflow implemented by default in the shiny:

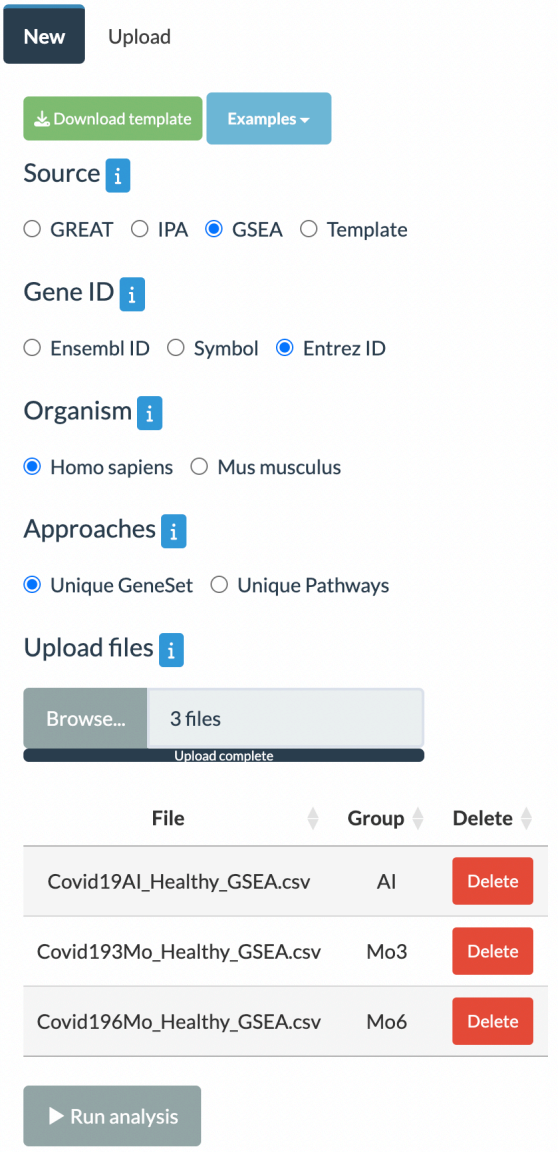


Main page of the web:

2. Inputs

2.1 Loading genesets file

All the following input fields are mandatory:



Source: Choose the data source for your analysis from three options: GREAT, IPA or GSEA. In the case of using another gene set enrichment tool, a template Excel can be downloaded (clicking *Download template* button), and uploaded into the app by selecting *Source* = "*Template"*. All the following template's fields (columns) have to be filled:

- *ID:* gene set id
- *Count:* number of your genes mapped in this gene set
- *GeneRatio:* "Count" column value divided by the total of genes of the gene set.
- *p.adjust:* adjusted p-value
- geneID: genes id (Ensembl ID, Symbol or Entrez ID) separated by comma.

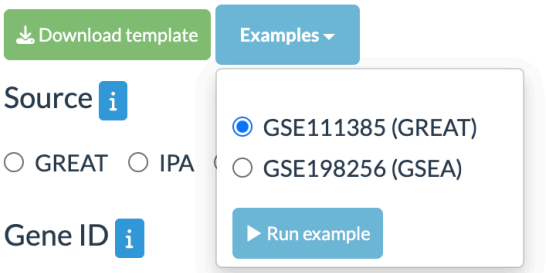
Gene ID: Specify the format of gene representation in your dataset.

Organism: Either Homo sapiens or Mus musculus.

Approaches: The "*Unique Pathways"* option combines duplicated gene sets by merging unique genes. "*Unique GeneSet"* treats each gene set as unique.

Upload files: Upload your data files, which must be in .txt, .csv, or .xls format. Once uploaded, a table will appear below. By default, each file represents a separate group (1, 2, etc.). You can modify the group names by double-clicking on the respective cell.

Run example: Show an example of the application using GSE111385 and GSE198256 dataset.



2.2 Loading from R data

- **Saved shiny session**: load the Rdata downloaded in previously.

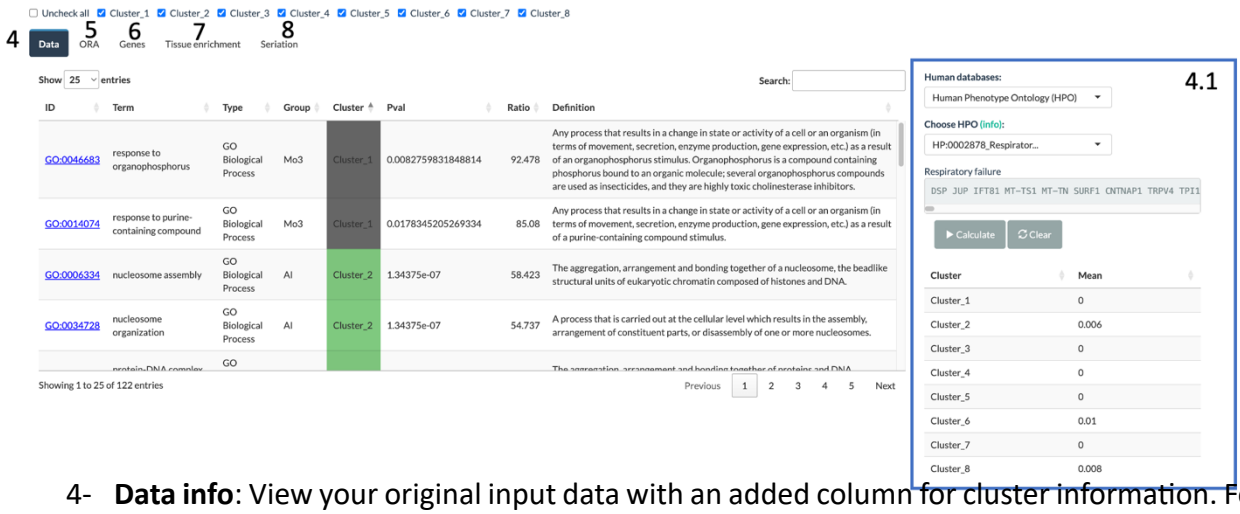
- **R object from R package GeneSetCluster:**
  - **Approaches**: "*Unique Geneset*" or "*Unique Pathways*".
  - **Object name**: The GeneSetCluster object name stored in the Rdata.
  - **File**: Rdata file containing the object.

3. Output

1- **Summary cluster**: Barplot summarizing the number of terms per selected cluster and their distribution among the Groups.

2- **Heatmap**: Visualize the RR matrix as a heatmap, annotated with cluster and group info. Additional annotations are available such as ORA info per cluster and wordcloud (only with GO ids).

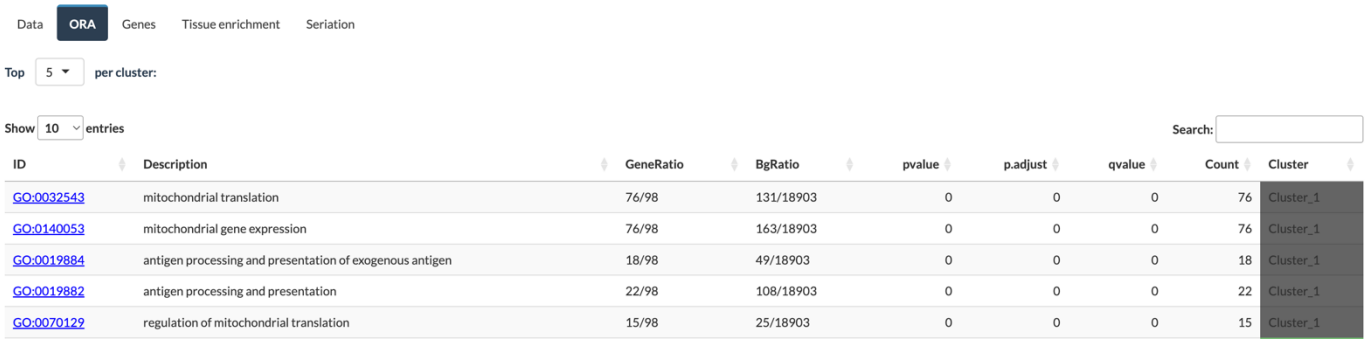
3- **Re/Subclustering:** Modify the default number of cluster set by the OptimalGeneSets function. All plots and data will update corresponding to the new value. Additionally, users can break up a cluster by selecting it and specifying the number of subclusters. The "*Automatic*" option uses the OptimalGeneSets function.



4- **Data info**: View your original input data with an added column for cluster information. For inputs from GREAT or GSEA, a "Description" column provides GO term descriptions. By clicking in “ID”, a hyperlink will direct you to the QuickGO database for further details.

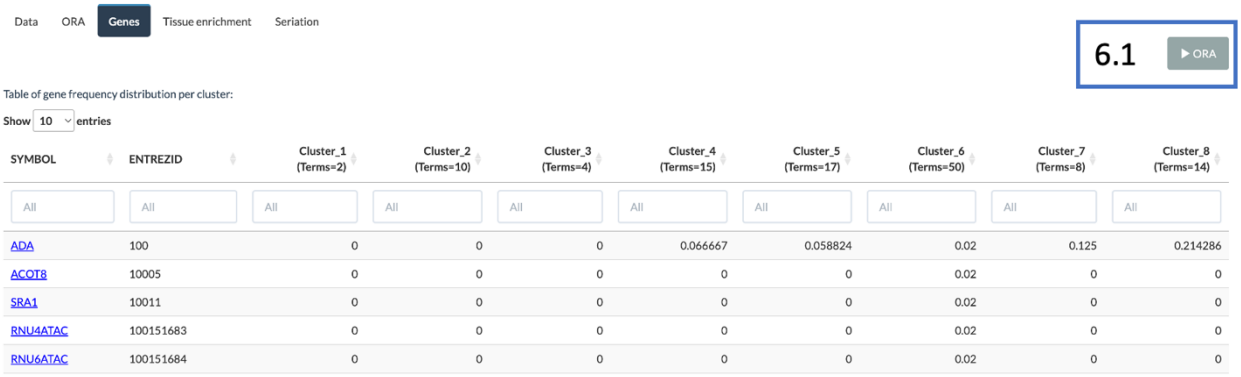
4.1- HighlightGeneSets: Select a gene of interest from the Mammalian Phenotype (Mus musculus), Human Phenotype Ontology (Homo sapiens) or *Customize* file. Ensure that the customized file contains only one column with gene symbols. View the corresponding scores per cluster in a table.

5- **ORA**



Explore ORA results per cluster, showing the top 5 enrichments by default (also top 10, 15, and 20 are allowed). Clicking on the “ID” column links to QuickGO.

6- **Genes**



Gene frequency distribution per cluster. For instance, if Cluster_1 has 10 pathways and a gene appears in only one pathway, its frequency is 1/10=0.1. Filter genes and frequencies as needed. Clicking on a gene name leads to GeneCards (https://www.genecards.org/) for additional information, in case of human data, otherwise to Mouse Genome Informatics (https://www.informatics.jax.org/).

6.1- **ORA:** After filtering the genes of interest, users can perform an ORA of the resulting genes. The ORA results will be automatically downloaded.

7- **Tissue enrichment**: (optional, for Organism=Homo sapiens): Calculate tissue enrichment analysis per cluster using the selected 15 representative tissues from https://gtexportal.org/home/. A plot will be generated with the results.

Tissue enrichment results per cluster (Cluster_number..number of genes):

Tissues:

Brain (Cortex), Liver, Lung, Pancreas, W ▶ Run

Show 10 entries Search:

Tissue	Cluster_1..173.	Cluster_2..26.	Cluster_3..133.	Cluster_4..16.	Cluster_5..1201.	Cluster_6..106.	Cluster_7..138.	Cluster_8..36.
Brain (Cortex)	1.91	1.516	1.626	0	2.216	1.385	1.601	1.479
Liver	2.054	1.793	1.932	1.4	2.238	1.754	1.935	1.72
Lung	2.383	2.213	2.158	1.535	2.396	1.898	2.176	1.749
Pancreas	2.019	1.644	1.787	0	2.22	1.541	1.782	1.775
Whole Blood	2.076	1.953	2.15	1.706	2.303	2.069	1.994	1.648

8- **Seriation-based analysis**: Optionally click to perform the seriation-based analysis

8.1- **Plots**: A barplot summarizing the cluster (seriation) information and the heatmap of the RR matrix with ORA annotations (similar to points 1 and 2).

8.2- **Data table:** Table result of the seriation-based analysis, showing which gene sets belong to each cluster.

8.3- **Other features**: ORA, gene, and tissue information, as explained in points 5, 6, 7.

9- **Downloads:**

9.1- **Plots**: downloads plots in png, jpg, or pdf format (in a zip).

9.2- **Data**: Raw data with cluster and RR information in .csv format.

9.3- **Results**: Save all the results in .Rdata format for import into the R package or re-upload in the Shiny app to continue analysis.