

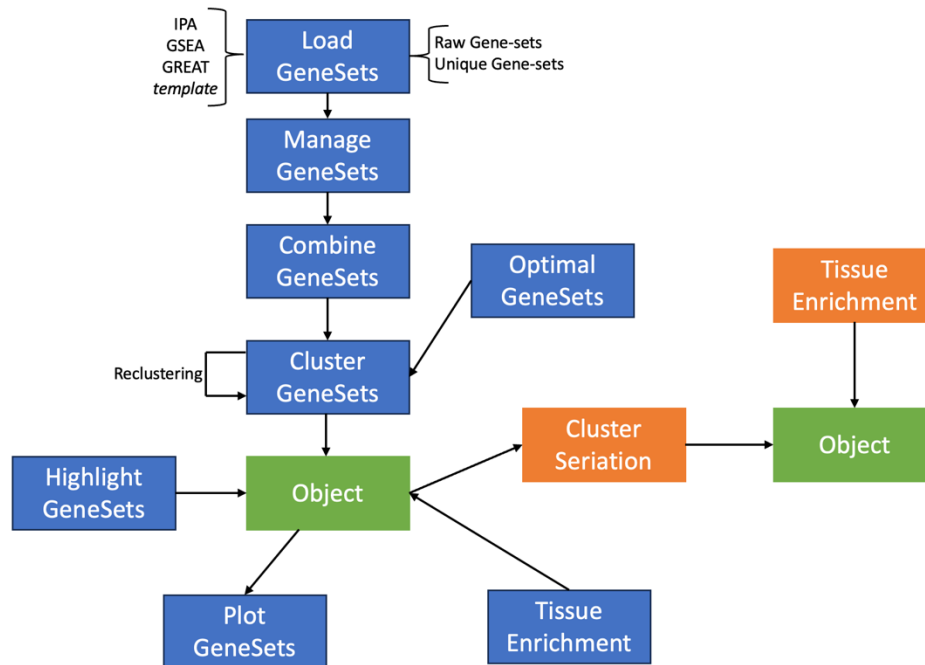
GeneSetCluster Shiny User Guide

1. General	2
2. Inputs	3
2.1 Loading gene-sets file	3
2.2 Loading from <i>RData</i>	4
3. Output	4
1- Summary cluster	4
2- Heatmap	4
3- Re/Subclustering:	4
4- Data info	5
5- Over-representation Analysis (ORA)	5
6- Genes	5
7- Tissue enrichment	6
8- Seriation-based analysis	7
9- Downloads	7

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 application:

GeneSet Cluster Main About

New Upload

[Download template](#) [Examples](#)

Source [i](#)

☐ GREAT ☐ IPA ☐ GSEA ☐ Template

Gene ID [i](#)

☐ Ensembl ID ☐ Symbol ☐ Entrez ID

Organism [i](#)

☐ Homo sapiens ☐ Mus musculus

Approaches [i](#)

☐ Raw Gene-sets ☐ Unique Gene-sets

Upload files [i](#)



Gene Set Cluster

Summarizing and integrating genesets results

2. Inputs

2.1 Loading gene-sets 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 “*Raw Gene-sets*” option includes all gene-sets independently. “*Unique Gene-sets*” combines gene-sets with the same label.

Upload files: Upload your data files, which must be in .txt, .csv, or .xls format. Once uploaded, a table will appear below. Each file represents a separate group (1, 2, etc.) by default. 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 datasets.

The screenshot shows the application's input interface. At the top, there are tabs for 'New', 'Upload', and 'Results'. Below these are buttons for 'Download template' and 'Examples'. The 'Source' field has a dropdown menu with options: GREAT, IPA, GSEA (selected), and Template. The 'Gene ID' field has a dropdown menu with options: Ensembl ID, Symbol, and Entrez ID (selected). The 'Organism' field has a dropdown menu with options: Homo sapiens (selected) and Mus musculus. The 'Approaches' field has a dropdown menu with options: Raw Gene-sets and Unique Gene-sets (selected). The 'Upload files' section has a 'Browse...' button and a 'No file selected' message. Below this is a table with columns 'File' and 'Group'. The table contains three rows: 'Covid19AI_Healthy.csv' with group 'AI', 'Covid193Mo_Healthy.csv' with group 'Mo3', and 'Covid196Mo_Healthy.csv' with group 'Mo6'. At the bottom right, there is a 'Run analysis' button.

File	Group
Covid19AI_Healthy.csv	AI
Covid193Mo_Healthy.csv	Mo3
Covid196Mo_Healthy.csv	Mo6

This screenshot shows a close-up of the 'Source' dropdown menu. The menu is open, showing two options: 'GSE111385 (GREAT)' (selected) and 'GSE198256 (GSEA)'. Below the menu, there is a 'Run example' button.

2.2 Loading from RData

- **Saved shiny session:** load the Rdata downloaded in previously.
- **R object from R package GeneSetCluster:**
 - **Approaches:** “Raw Gene-sets” or “Unique Gene-sets”.
 - **Object name:** The GeneSetCluster object name stored in the RData file.
 - **File:** RData file containing the object.

New Upload

From:

☐ Shiny ☒ R package

Approaches

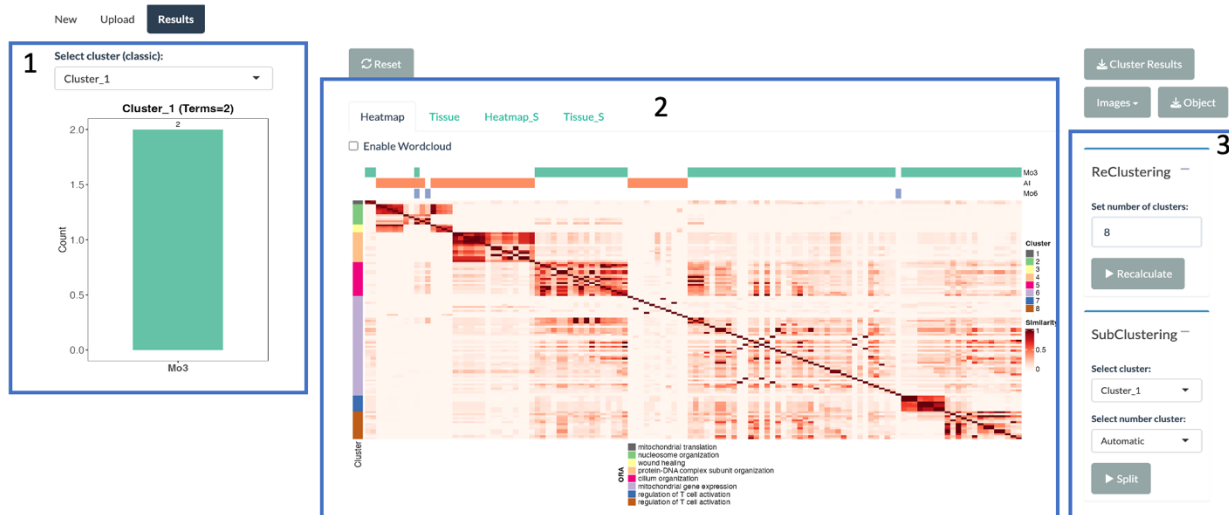
☐ Raw Gene-sets ☒ Unique Gene-sets

Object name

Genesetcluster_ObjectName

Choose a file

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 over-representation analysis (ORA) info per cluster and wordcloud (only with GO ids).
- 3- **Re/Subclustering:** Modify the default number of clusters set by the *OptimalGeneSets* function. All plots and data will be updated according 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.

☐ Uncheck all
 ☒ Cluster_1
 ☒ Cluster_2
 ☒ Cluster_3
 ☒ Cluster_4
 ☒ Cluster_5
 ☒ Cluster_6
 ☒ Cluster_7
 ☒ Cluster_8

4 Data
 5 ORA
 6 Genes
 7 Tissue enrichment
 8 Seriation

Show 25 entries
 Search:

ID	Term	Type	Group	Cluster	Pval	Ratio	Definition
GO:0046683	response to organophosphorus	GO Biological Process	Mo3	Cluster_1	0.0082759831848814	92.478	Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of an organophosphorus stimulus. Organophosphorus is a compound containing phosphorus bound to an organic molecule; several organophosphorus compounds are used as insecticides, and they are highly toxic cholinesterase inhibitors.
GO:0014074	response to purine-containing compound	GO Biological Process	Mo3	Cluster_1	0.0178345205269334	85.08	Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a purine-containing compound stimulus.
GO:0006334	nucleosome assembly	GO Biological Process	AI	Cluster_2	1.34375e-07	58.423	The aggregation, arrangement and bonding together of a nucleosome, the beadlike structural units of eukaryotic chromatin composed of histones and DNA.
GO:0034728	nucleosome organization	GO Biological Process	AI	Cluster_2	1.34375e-07	54.737	A process that is carried out at the cellular level which results in the assembly, arrangement of constituent parts, or disassembly of one or more nucleosomes.
GO:0000000	contains DNA, composed	GO					The interaction, arrangement and location of histones and DNA

Showing 1 to 25 of 122 entries
 Previous 1 2 3 4 5 Next

Human databases:
 Human Phenotype Ontology (HPO)
 Choose HPO (Info):
 HP:0002878_Respirator...
 Respiratory failure
 DSP JUP IFT81 MT-TS1 MT-TN SURF1 CNTNAP1 TRPV4 TPST1
 Calculate Clear
 Cluster Mean
 Cluster_1 0
 Cluster_2 0.006
 Cluster_3 0
 Cluster_4 0
 Cluster_5 0
 Cluster_6 0.01
 Cluster_7 0
 Cluster_8 0.008

- 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- Over-representation Analysis (ORA)

Data
 ORA
 Genes
 Tissue enrichment
 Seriation

Top 5 per cluster:

Show 10 entries
 Search:

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	Count	Cluster
GO:0032543	mitochondrial translation	76/98	131/18903	0	0	0	76	Cluster_1
GO:0140053	mitochondrial gene expression	76/98	163/18903	0	0	0	76	Cluster_1
GO:0019884	antigen processing and presentation of exogenous antigen	18/98	49/18903	0	0	0	18	Cluster_1
GO:0019882	antigen processing and presentation	22/98	108/18903	0	0	0	22	Cluster_1
GO:0070129	regulation of mitochondrial translation	15/98	25/18903	0	0	0	15	Cluster_1

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

Data
 ORA
 Genes
 Tissue enrichment
 Seriation

Table of gene frequency distribution per cluster:

Show 10 entries

SYMBOL	ENTREZID	Cluster_1 (Terms=2)	Cluster_2 (Terms=10)	Cluster_3 (Terms=4)	Cluster_4 (Terms=15)	Cluster_5 (Terms=17)	Cluster_6 (Terms=50)	Cluster_7 (Terms=8)	Cluster_8 (Terms=14)
ADA	100	0	0	0	0.066667	0.058824	0.02	0.125	0.214286
ACOT8	10005	0	0	0	0	0	0.02	0	0
SRA1	10011	0	0	0	0	0	0.02	0	0
RNU4ATAC	100151683	0	0	0	0	0	0.02	0	0
RNU6ATAC	100151684	0	0	0	0	0	0.02	0	0

Gene frequency distribution per cluster. For instance, if Cluster_1 has 10 gene-sets and a gene appears in only one gene-set, 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.

☐ Uncheck all ☒ Cluster_1 ☒ Cluster_2 ☒ Cluster_3 ☒ Cluster_4 ☒ Cluster_5 ☒ Cluster_6 ☒ Cluster_7 ☒ Cluster_8

Data ORA Genes **Tissue enrichment** Seriation

Choose specific tissues for performing enrichment analysis (the tissues were selected from [here](#)):

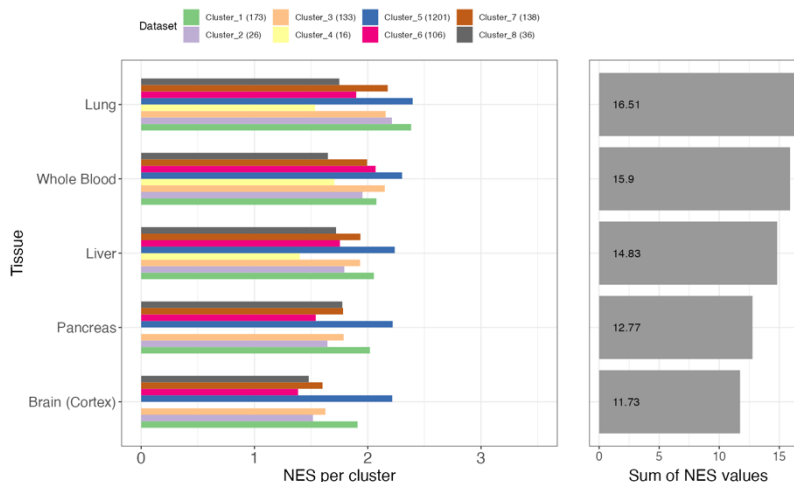
Tissues:

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

Run

Heatmap

Tissue



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-based) 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, and 7 respectively.

9- Downloads:

9.1- Cluster Results: Results of the clustering including the RR information in .csv format.

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

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