**GScluster (v1.1.6)**

* *User’s Manual -*

Author: Sora Yoon <[yoonsora1@unist.ac.kr](mailto:yoonsora1@unist.ac.kr)>

Maintainer: Jinhwan Kim <[kjh0530@unist.ac.kr](mailto:kjh0530@unist.ac.kr)>

Last updated: 2019. 2. 28.

1. **Introduction**

GScluster is a Shiny R package that performs network-weighted gene-set clustering by incorporating both gene-set overlap and protein-protein interaction (PPI) networks. Gene-set enrichment analysis of complex diseases often yields a long list of dysregulated pathways in disease. GScluster can cluster such pathways and help users summarize them. Built on Shiny package, it provides interactive user-interface so that users can easily navigate the gene-set and each cluster's gene network. GScluster provides PPI networks for ten species such as human, mouse, rat, arabidopsis, C.elegans, E.Coli, fly, rice, yeast and zebrafish. The installation and usage of GScluster is described below. PPI networks and gene networks are used interchangeably in this manual, because the network data for GScluster can be any kind of functional interactions between genes.

1. **Installation (Do only once)**
2. Open R program.
3. Type following commands in R console.

|  |
| --- |
| >> install.packages('devtools')  >> library(devtools)  >> install\_github('unistbig/GScluster') |

1. **Launch**
2. Load the GScluster R package by typing following line in R console

|  |
| --- |
| >> library(GScluster) |

1. Run GScluster using ‘GScluster’ function. To run the demo, enter GScluster(). Otherwise, modify following example source codes to fit your data.

|  |
| --- |
| ## example code ##  # 1) Read gene-set analysis result table.  >> GSAresult=read.delim('https://github.com/unistbig/GScluster/raw/master/sample\_geneset.txt', stringsAsFactors=FALSE)  # 2) Read gene score table.  >> GeneScores=read.delim('https://github.com/unistbig/GScluster/raw/master/sample\_genescore.txt', header=F)  # 3) Read PPI matrix.  >> download.file(url='https://github.com/unistbig/GScluster-Data/raw/master/human/string.RData', dest='humanPPI.RData')  >> load('humanPPI.Rdata') # It loads a PPI matrix named ‘string’  >> PPI=string  # 4) Run GScluster  >> GScluster(GSAresult = GSAresult, GeneScores = GeneScores, PPI = PPI, Species = 'H', alpha = 1, GsQCutoff = 0.25, GQCutoff = 0.05) |

As shown in the code above, *GScluster* function takes seven arguments as follows:

1. **GSAresult**: A data frame representing gene-set analysis result, consisting of three or four columns as follows:
   1. **Geneset name:** A character vector of gene-set names
   2. **Gene list:** A character vector of gene-set member lists. For each gene-set, genes should be separated by a space (Figure 1).

\* Any kind of gene IDs can be used for custom PPI data, but gene symbols should be used for the default STRING.

* 1. **q-values:** A numeric vector of gene-set q-values (or p-values).
  2. **Direction (optional):** A character vector indicating whether each gene-set is up-regulated (UP) or down-regulated (DN).

An example GSA result file is available at:

<https://github.com/unistbig/GScluster/raw/master/sample_geneset.txt>

|  |  |  |  |
| --- | --- | --- | --- |
| **Geneset Name** | **Gene list** | **q-values** | **Direction** |
| Geneset 1 | A B C | 0.1 | UP |
| Geneset 2 | A B D E F | 0.005 | DN |
| Geneset 3 | C D E G | 2.00E-03 | DN |

Figure 1. GSA result table format. A header is required.

1. **GeneScores:** A data frame of gene scores (e.g., p-value of q-value in differential expression analysis) consisting of two columns as follows:
   1. **Gene:** A character vector of gene names.
   2. **Scores:** A numeric vector of gene p-values (or q-values). Each value must be between 0 and 1. Exponential formats such as 1E-3 are allowed.

An example gene score file is available at:

<https://github.com/unistbig/GScluster/raw/master/sample_genescore.txt>

|  |  |
| --- | --- |
| **Gene** | **Q-value** |
| A | 1.00E-06 |
| B | 1.00E-06 |
| C | 0.0012 |
| D | 0,275 |
| E | 0.8324 |

Figure 2. Gene score table format. A header is required.

1. **PPI:** A matrix of protein-protein interaction scores consisting of numeric values between 0 and 1. If the ‘**Species**’parameteris specified as one of the ten species provided by GScluster, the user does not need to specify this option (In this case, STRING PPI network table is automatically loaded).

An example PPI matrix file (an object named ‘string’ in the .Rdata file) is available at:

[https://github.com/unistbig/GScluster-Data/raw/master/human/string.RData](https://github.com/unistbig/GScluster-Data/raw/master/human/string.RData?raw=true)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **A** | **B** | **C** |
| **A** | 0 | 0.1 | 0.76 |
| **B** | 0.1 | 0 | 0.324 |
| **C** | 0.76 | 0.324 | 0 |

Figure 3. PPI table format. It is a symmetric matrix with diagonal value of zero.

1. **Species:** One uppercase alphabet representing the species of input data. Default value is “H” to indicate ‘human’. Possible values are “A” (Arabidopsis), “C” (C.elegans), “E” (E.coli), “F” (Fly), “H” (Human), “I” (RIce), “M” (Mouse), “R” (Rat), “Y” (Yeast) and “Z” (Zebrafish).
2. **alpha:** A network weight for the pMM distance, in the range 0 to 1. The default value is 1. If Alpha is 0, pMM and MM have the same value.
3. **GsQCutoff:** A numeric value indicating gene-set q-value cutoff (0~1). The gene-sets with a q-value greater than the cutoff is excluded from clustering. Default value is 0.25.
4. **GQCutoff:** A numeric value indicating gene q-value cutoff (0~1). The genes with a q-value greater than this cutoff is excluded from the graph.

* **TIP**
  + We have developed gene-set analysis tools for

1. GWAS summary data

(**GSA-SNP2**, download site: <https://sourceforge.net/projects/gsasnp2/>) and,

1. Gene expression data (RNA-seq or microarray)

(**GSAseq**, available at <http://gsaseq.appex.kr/gsaseq/>)

The gene-set analysis result file obtained from GSA-SNP2 or GSAseq can be directly applied to *GScluster* function using **‘GetGSASNP2Data’** and **‘GetGSASeqData’** functions, respectively. Example codes are as follows:

|  |
| --- |
| **# Case 1: using GSA-SNP2 result file**  # 1) Transform GSA-SNP2 result file. ‘filename’ is specified as the location of the file.  >> data=GetGSASNP2Data(filename='https://github.com/unistbig/GScluster/raw/master/inst/GScluster/GSASNP2\_DIAGRAM.txt')  # Here, the ‘data’ object is a list of two tables including GSA result (‘GSAresult’) and gene score table (‘GeneScores’)  # 2) Get GSA result table.  >> GSAresult = data$GSAresult  # 3) Get Gene score table  >> GeneScores = data$GeneScores  # 4) Run GScluster function  >> GScluster(GSAresult = GSAresult, GeneScores = GeneScores, Species = 'H', alpha = 1, GsQCutoff = 0.25, GQCutoff = 0.05) |

The GSAseq result file can be applied in a similar way.

|  |
| --- |
| **# Case 2: using GSAseq result file**  # 1) Transform the GSAseq result file.  >> data=GetGSASeqData(filename=https://github.com/unistbig/GScluster/blob/master/inst/GScluster/GSAseq\_GSE4107.txt?raw=true')  # 2) Get GSA result table.  >> GSAresult = data$GSAresult  # 3) Get Gene score table  >> GeneScores = data$GeneScores  # 4) Run GScluster function  >> GScluster(GSAresult = GSAresult, GeneScores = GeneScores, Species = 'H', alpha = 1, GsQCutoff = 0.01, GQCutoff = 0.01) |

1. **Exploring Networks**

After running the GScluster function, the Shiny app will run in a few seconds or minutes and display the gene-set networks (Figure 4). Since the program was developed under Chrome environment, we recommend that the user set Chrome as default browser.

The default options for gene-set clustering are

1. Distance type: pMM (PPI-weighted Meet/Min distance)
2. Distance cutoff: Top % pMM value. (: Percentage value corresponding to MM = 0.5)
3. Seed cluster size: 3 (required for fuzzy clustering)

For more information on pMM, please refer to the manuscript.

스크린샷이(가) 표시된 사진

자동 생성된 설명

**Figure 4.** Gene-set networks. The functions or detailed information are described below (numbers 1 – 10).

**Useful Functions in GScluster**

It provides a number of functions to explore both gene-set and gene networks.

1. **Graph control panel:** By controlling mouse or graph control panel (① in Fig. 4), the user can zoom in and out of the graph, adjust it to fit the screen, or move it.

Using mouse drag, the user can 1) select multiple nodes on the graph or 2) move the entire graph by holding down mouse button for 2 seconds.

1. **Node (Gene-set or gene) search:** To locate a particular node (gene-set or gene) on the graph, type the keyword in the upper left box (② in Fig. 4) and select the item you want. Then, the node will be highlighted and blink for a few seconds in magenta color on the graph. In case of gene-set node, corresponding cluster number will also appear (Fig. 5).

**스크린샷이(가) 표시된 사진

자동 생성된 설명**

Figure 5. Gene-set node search

1. **Clustering Method:**
   1. The distance type (meet/min distance (MM), PPI-weighted meet/min distance (pMM) or Cohen’s Kappa distance (Kappa))
   2. minimum seed size (= minimum gene-set cluster size in fuzzy clustering) and
   3. maximum distance between connected gene-sets

can be set in ‘Clustering Method’ (③ in fig 4). After setting these parameters, click on **‘APPLY’ button**.

스크린샷이(가) 표시된 사진

자동 생성된 설명

Figure 6. Clustering options

1. **Label size:** It adjusts the node label size (④ in Figure 4, Figure 7).

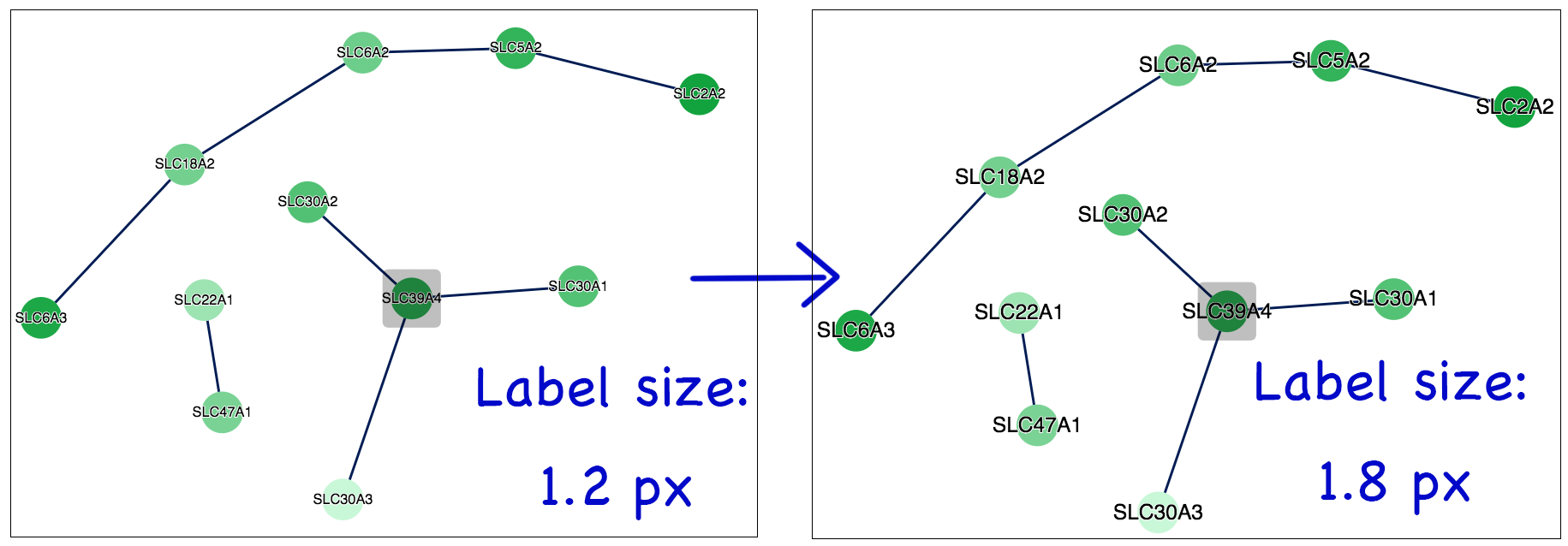
****

Figure 7. Node label size adjustment

1. **Graph Layout:** Click on the 'Graph layout' button (⑤ in Figure 4) to change the graph layout between 'circle' and 'cola' (Figure 8).

**\*TIP.** If the clusters do not appear to be fully isolated, click on ‘Graph Layout’ and ‘Cola’ button.

지도, 실내이(가) 표시된 사진

자동 생성된 설명

Figure 8. Graph layout. (A) Circle and (B) Cola layout

1. **Download Graph:** The user can download a vector image file (.SVG) for current plot by clicking on the ‘Download Graph’ button (⑥ in Figure 4).
2. **Highlight gene-set cluster**

The user can highlight a gene-sets cluster by selecting a cluster number from ‘Select Gene-set Cluster’ menu (⑦ in Figure 4). Selecting ‘Unselected’ turns off highlighting.

**스크린샷, 모니터이(가) 표시된 사진

자동 생성된 설명**

Figure 9. Highlighting gene-set cluster. (A) A gene-set cluster is highlighted. (B) Highlighting is turned off by selecting ‘Unselect’.

1. **Gene network**

GScluster also visualize the gene networks of each gene-set cluster based on STRING network (PPI) data (Custom network data is also applicable). For example, the user can see the gene networks in cluster 7, as follows:

* 1. Choose the cluster number (‘7’ in this case) from **‘Select Gene-set Cluster’** box.
  2. Set the **PPI cutoff** (default=700, value is in the range 0 - 1000 where 1000 is the strongest).
  3. Click on **‘Show PPI’** button (⑧ in Figure 4). Then, it will show the network for genes in cluster 7. To go back to the gene-set network mode, please click onthe **‘Go back’** button.
  4. Each gene node is hyperlinked to the corresponding *GeneCards* site (<https://www.genecards.org>).
  5. In gene networks, two additional functions are provided (**PPI Evidence** (① in Fig 10, Figure 11) and **WordCloud** (② in Fig 10, Figure 12, human only))

스크린샷, 실내이(가) 표시된 사진

자동 생성된 설명

Figure 10. Gene network for a specific cluster ( cluster 7 )

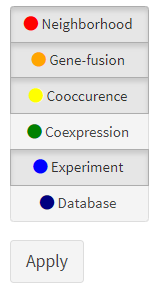
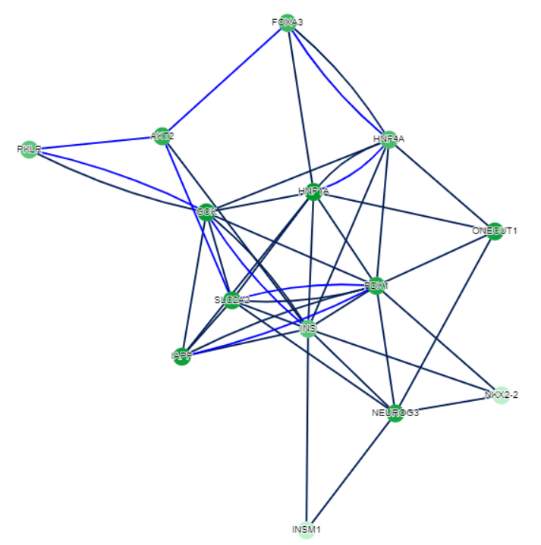
* 1. PPI Evidence

Figure 11. Six different networks from STRING are visualized. The text-mining evidence is excluded.

PPI Evidence shows detailed network information**.** Six edge types such as Neighborhood, Gene fusion, Co-occurrence, Co-expression, Experiments, and Databases are visualized. Detailed explanation for each PPI evidence type is described in STRING web page ( <https://string-db.org/cgi/help.pl> ). Default edges represent the combined scores.

1. Wordcloud: GScluster supports disease information obtained from DisGeNET ([www.disgenet.org](http://www.disgenet.org)), which is available only for human. It shows wordcloud of any common disease (frequency > 1) for the input genes (Figure 12).

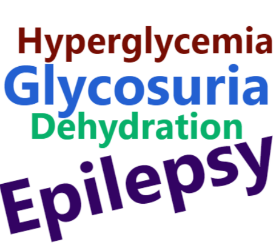
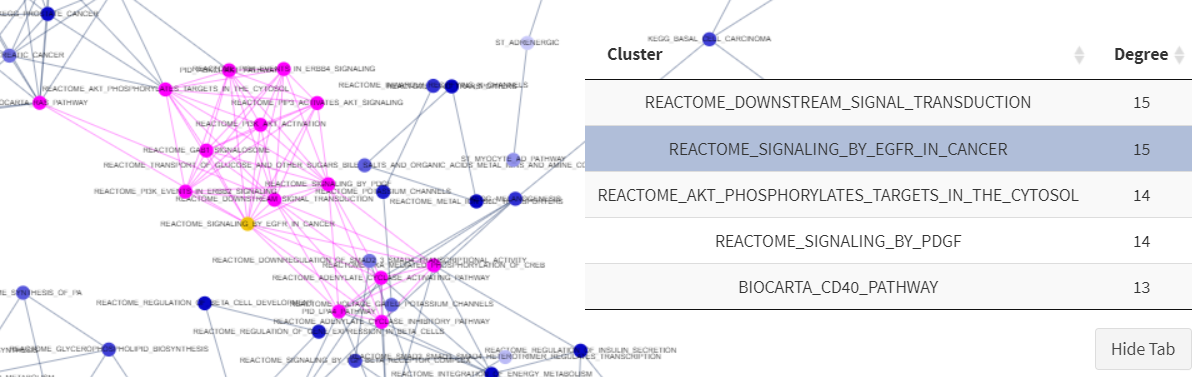


Figure 12. Wordcloud generated from a specific cluster

1. **Hub**

GScluster supports three types of hubs. (⑨ in Fig. 4. Hub means highly connected node).

* 1. ‘Geneset Hub’ shows top 5 hub gene-sets in Gene-set network. By clicking on them in table, that hub (yellow) and its neighbors (magenta) will be highlighted (Figure 13). Clicking on the network background the highlighting is canceled. It is available from the gene-set network mode.

Figure 13. Gene-set hub and its neighbor nodes are highlighted

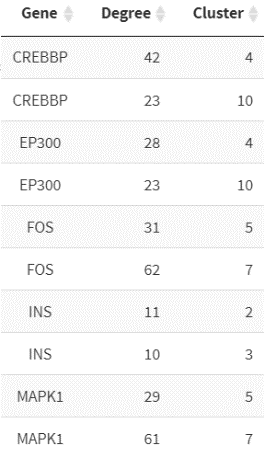
* 1. ‘Gene hubs’ are visualized in the same way (Available from the gene network mode).
  2. ‘Multi-cluster Hub’ shows gene hubs included in at least two clusters. For example, CREBBP is a hub gene in both cluster 4 and cluster 10 with 23 and 42 neighbors, respectively (Available from the gene-set network mode, Figure 14).

Figure 14. Example of Multi-Cluster Hubs

1. **Clustering Results.**

스크린샷이(가) 표시된 사진

자동 생성된 설명

Figure 15. Clustering result table

Clicking on the **‘Clustering Results’** button (10 in Figure 4) will show the table for gene-set clustering results. The result table can be filtered based on cluster number, Gene-set name or Gene-set q-value. The user can copy, download (as .csv file) and print the table.