**GScluster User’s Manual**

Contact: Jinhwan Kim [kjh0530@unist.ac.kr](mailto:kjh0530@unist.ac.kr)

2019-01-05

1. **Install and Launch**

GScluster is a standalone R package, available at Github ( [**github.com/unistbig/GScluster**](https://github.com/unistbig/GScluster) ).

In addition to the package, data required to run GScluster are also downloadable from Github ([**github.com/unistbig/GScluster-Data/**](https://github.com/unistbig/GScluster-Data/)) or by using the R function **DownloadData**.

To install and launch GScluster. Type the following lines on the R console.

**> library(devtools) # To use install\_github command**

**> install\_github("unistbig/GScluster ") # install GScluster**

**> library(GScluster)**

**> GScluster()**

\* Note that this manual demonstrates analysis for human GWAS summary data (DIAGRAM consortium) and STRING network data.

1. **How to use**

GScluster takes 7 input parameters. Gene names (denoted by A, B, C here) should be gene symbols when using the default STRING data, but other kinds of gene names can be used if the corresponding customized data (GSAresult and network data) are used.

1. GSAresult: the main object of gene-set analysis (GSA) result file, consisting of 3 or 4 columns.

Geneset Name: names for gene-sets

Genelist: member genes of each gene-set

Qvalues: Score of each gene-set

Direction (optional): indicates whether gene-set is up-regulated (UP) or down-regulated (DN)

|  |  |  |  |
| --- | --- | --- | --- |
| **GS** | **GeneList** | **Qvalues** | **Direction** |
| GenesetA | A B C | 0.1 | UP |
| GenesetB | A B D E F | 0.005 | DN |
| GenesetC | C D E G | 2.00E-03 | DN |
| GenesetD | A D G H | 0.72 | UP |

Figure 1. Example for gsFile. A header exists and GeneList should be separated by white spaces.

1. GeneScores: objects of DE analysis result consisting of 2 columns, genes and their scores.

Gene: gene name.

Score: numeric value between 0 and 1. Exponential format like 1E-3 is acceptable.

|  |  |
| --- | --- |
| A | 1.00E-06 |
| B | 1.00E-06 |
| C | 0.0012 |
| D | 0,275 |
| E | 0.8324 |

Figure 2. Example of gene score input file. There is no header.

1. PPI: a matrix object of protein-protein interaction networks, consisting of numeric values 0 – 1.

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

Figure 3. Example for network input data. It is symmetric matrix and diagonal values are 0.

1. Species: single character indicating the species of data. Default is H.

Possible value is (A)rabidopsis, (C)elegans, (E).coli, (F)ly, (H)uman, R(I)ce (M)ouse, (R)at, (Y)east, (Z)ebrafish

1. Alpha: the network weighting value for pMM distance between 0 and 1. Default is 1. If Alpha is 0, pMM and MM have the same values.
2. GsQCutoff: numeric value between 0 and 1. Default is 0.25

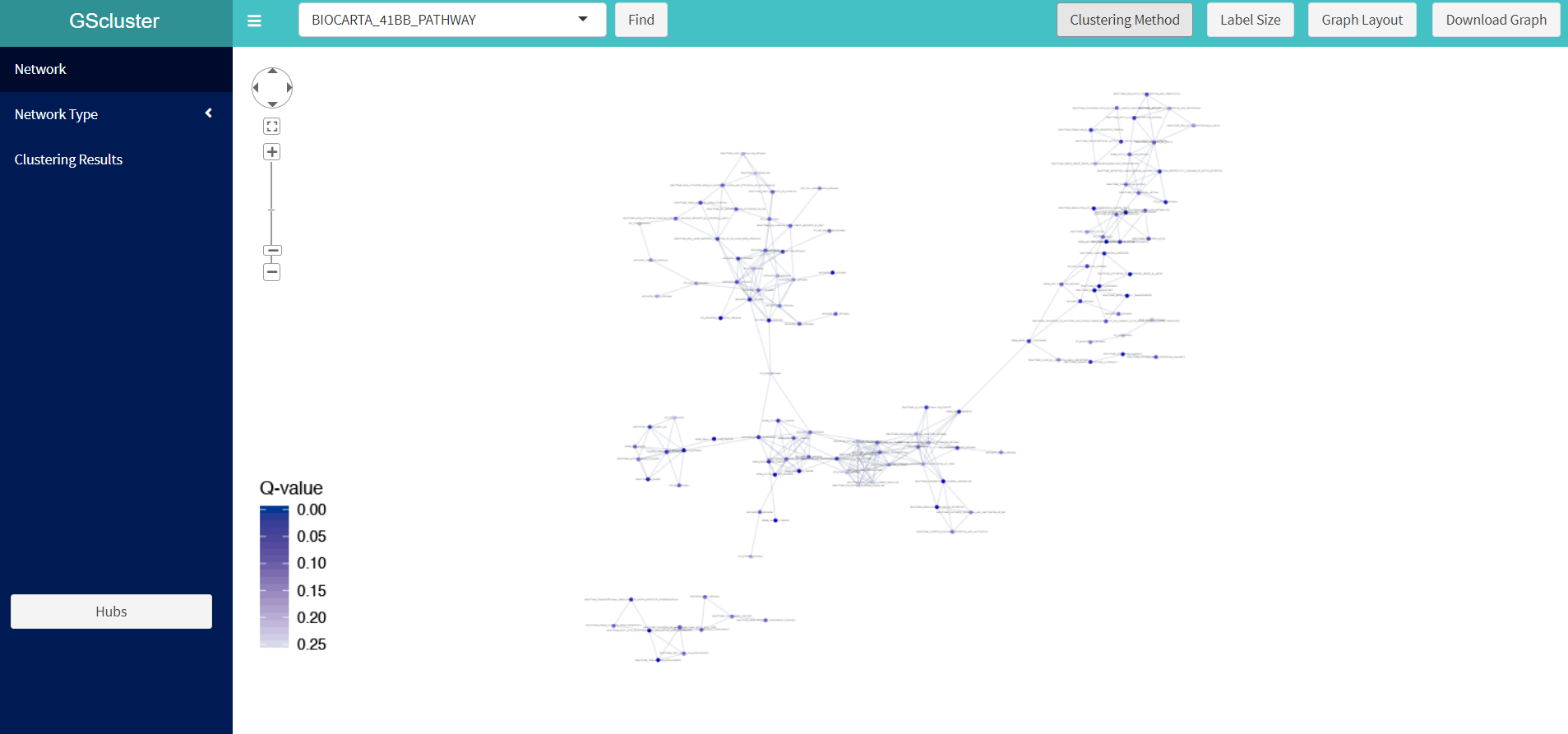
This parameter stands for ‘gene set q-value cutoff’. Gene-sets with a q-value larger than this cutoff will be excluded.

1. GQCutoff : numeric value between 0 and1. Default is 0.25

This parameter stands for genes core q-value cutoff’. Genes with a q-value larger than this cutoff will be excluded.

**3. Exploring the Network**

After clustering gene-sets, the gene-set networks will be displayed. The default gene-set clustering uses pMM distance, Seed size: 3, Distance cutoff: the same percentage value that corresponds to 0.5 of MM distance. The detailed clustering results are represented in **Clustering Results** tab and users can download the clustered result (⑨ in Fig 4) as csv file.



**←①**

**↑②**

**③→**

**↑④**

**↑⑤**

**↑⑥**

**←⑦**

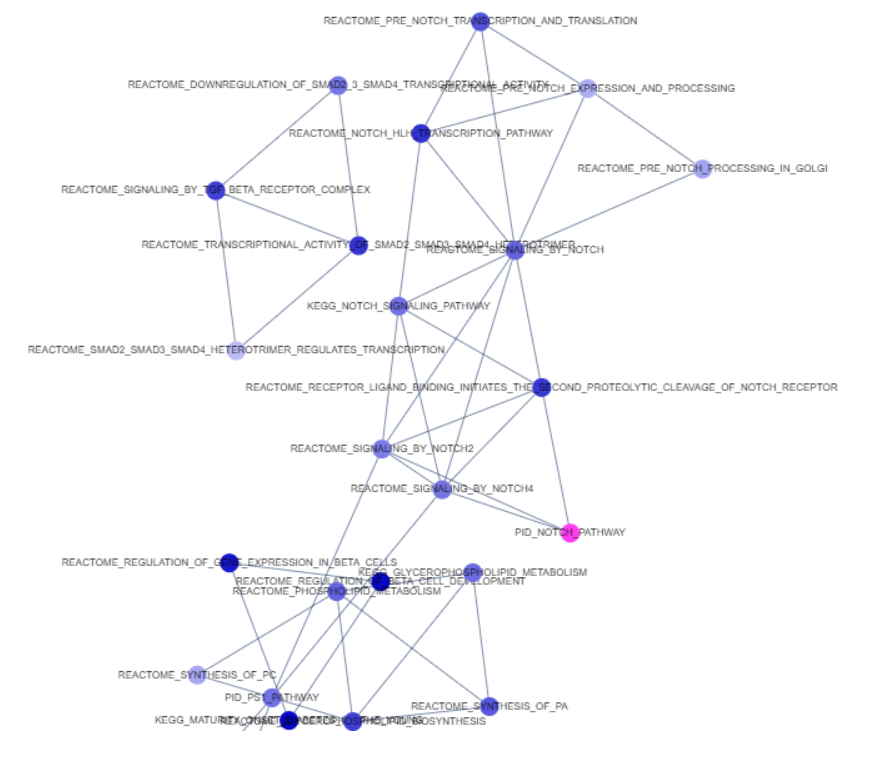
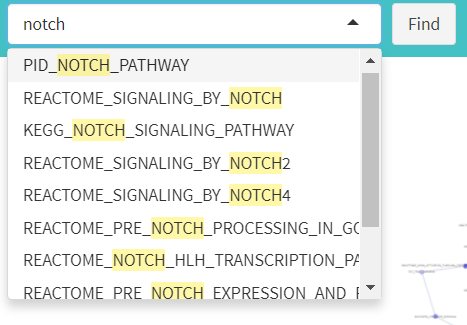
**↓⑧**

**↑⑨**

Figure 4. Gene-set networks. The functions or detailed information are described below (numbers 1 – 8).

**GScluster functions**

1. **Graph control panel:** zoom in/out, fit or move the graph by simple mouse control or using graph control panel (① in Fig 4). Drag network will perform 2 actions, 1) select multiple Nodes and 2) move whole network role can be changed by clicking background ( white empty network space ) with right-click.
2. **Gene-set search:** To find a specific node ( Gene-set or Gene), type a keyword in box(② in fig 4), select target’s name, and click ‘Find’ button. Then corresponding node will be Highlighted with magenta color in a few seconds. (Fig. 5).



**1. Type a keyword**

**2. Select target & Find**

**3. Node will Highlighted**

Figure 5. Search for a gene-set node

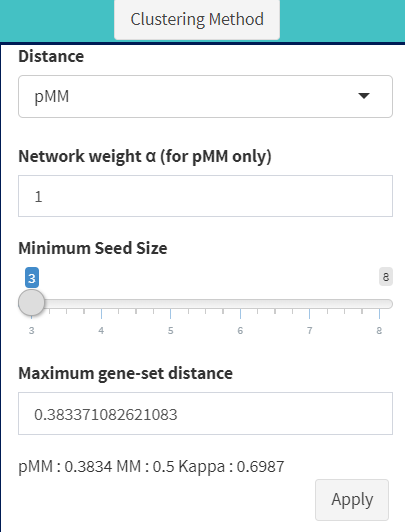
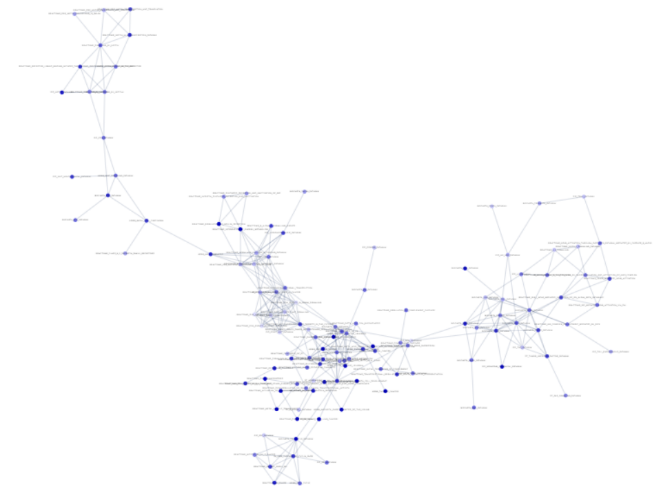
1. **Clustering Method:** The distance type, minimum seed size and maximum distance allowed between gene-sets can be set in ‘Clustering Method’ (③ in fig 4). After setting these parameters, ****click **‘APPLY’**.

Figure 6. Clustering Method

1. **Label size:** label for network node can be adjusted.
2. **Graph Layout:** Click ‘Graph Layout’ button (⑤ in Fig 4) and choose circle or cola layout (Fig. 7).

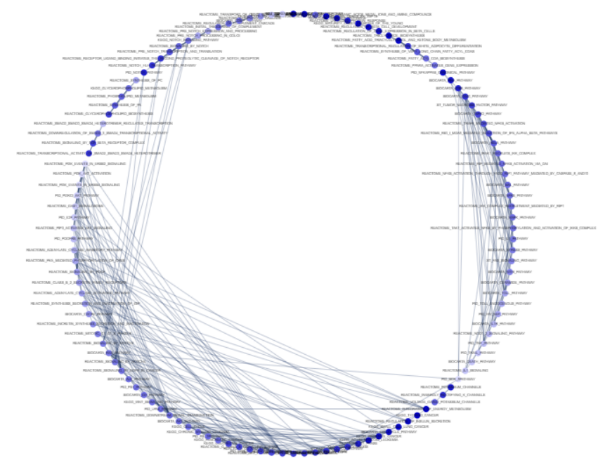
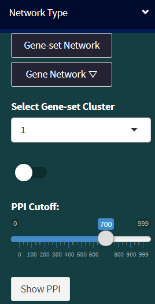


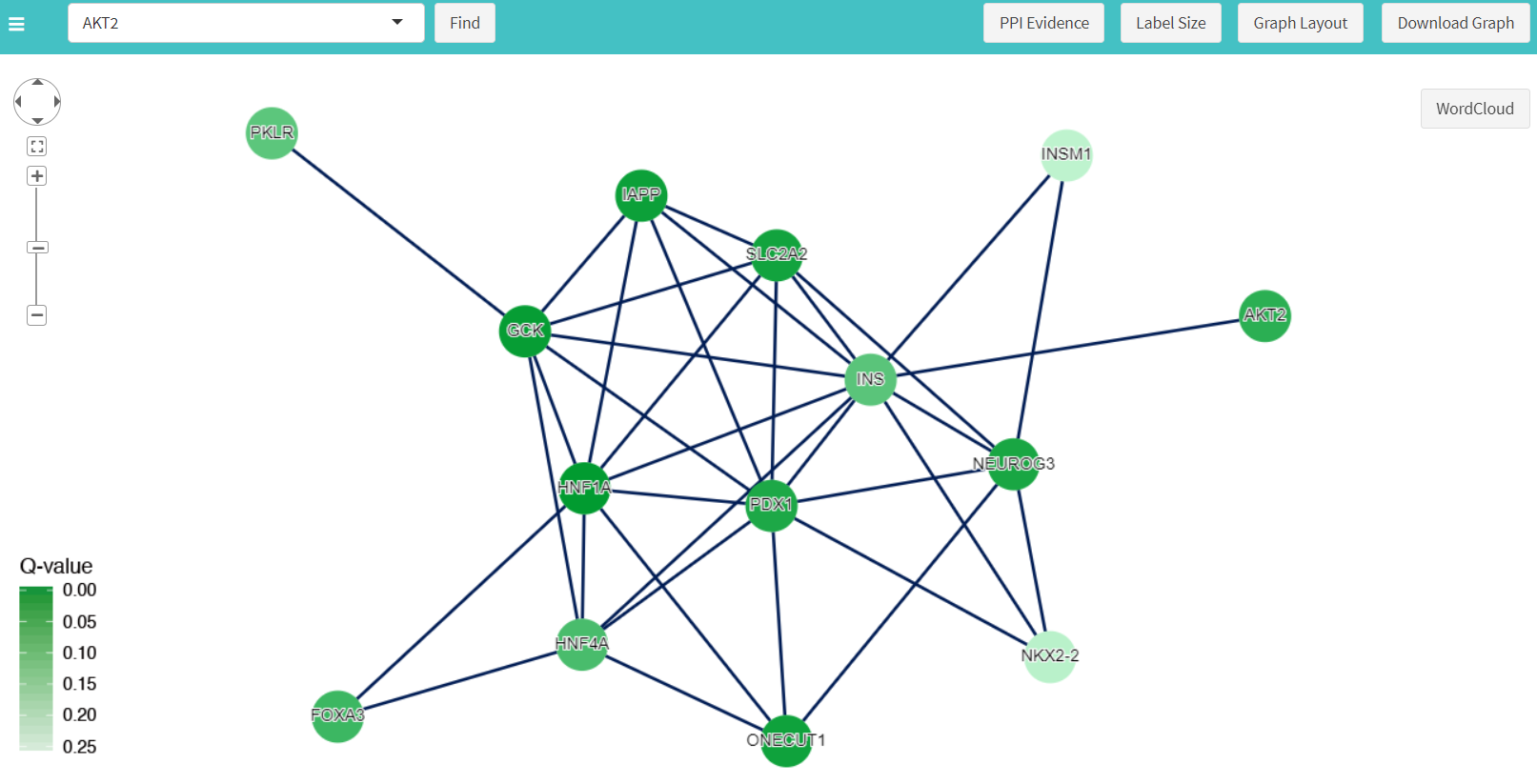
Figure 7. Circle (left) and Cola (right) layout

1. **Download Graph:** The user can download a vector image file (.SVG) for current plot by clicking on ‘Download Graph’ button (⑥ in fig 5).
2. **Network Type :**

**←①**

Figure 8. Network Type

1. Gene Network: GScluster also provides gene network plot of each gene-set cluster based on STRING human network (PPI) data. For example, the user can see the gene network in cluster 1, as follows:
   1. Click the **‘Gene Network’** button.
   2. Choose the cluster number (‘1’ in this case) from **‘Select Gene-set Cluster’** box.
   3. Set the **PPI cutoff** (default=700, value is 0 - 1000 where 1000 is the most strongest )
   4. Click **‘Show PPI’** button. Then it will show the network for genes in cluster 1.
   5. In the phase of gene networks, two additional functions are provided (WordCloud ① in Fig 9 and PPI Evidence in ② Fig 10)



**↑①**

**↑②**

Figure 9. Gene network for a specific cluster ( cluster 3 )

* 1. Wordcloud: GScluster supports disease information gathered from DisGeNET ([www.disgenet.org](http://www.disgenet.org)), note that this function is available for only human. It shows wordcloud of any common disease (frequent > 1) for the selected (significant) genes

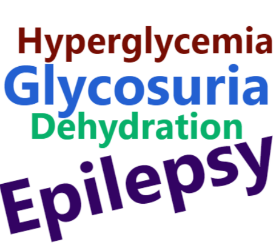


Figure 10. Wordcloud generated from a specific cluster

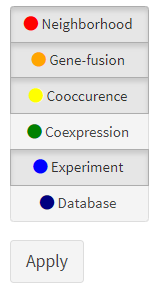
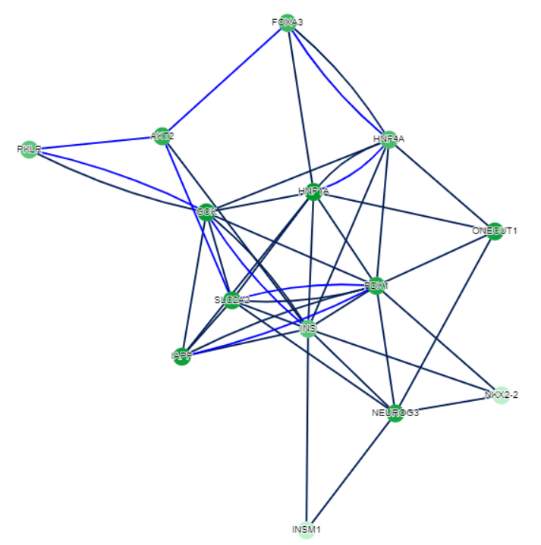
* 1. PPI Evidence.

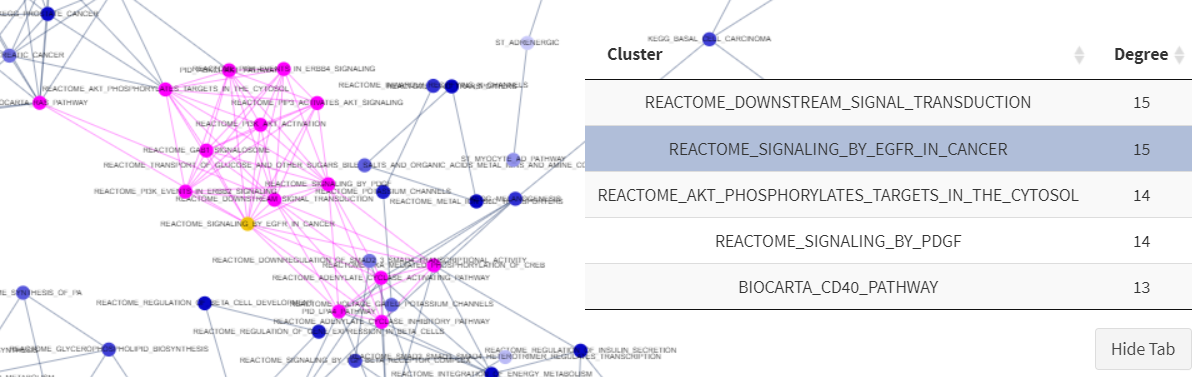
Figure 11. Six different networks from STRING are visualized.

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> ). Note that basic edge represents the combined score and the text-mining evidence is excluded.

1. Switch in the middle of the pannel (① in Fig 8) is used to highlight gene-sets. It will highlight given number (**in Select Gene-set Cluster**) cluster gene-set nodes. For example, gene-set nodes in cluster 1 will be highlighted with Magenta color.
2. ‘Gene-set network’ button will restore the gene-set networks.
3. **Hub**

GScluster supports three types of hubs. (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. Clicking on the network background will cancel the highlight.

Figure 12. Gene-set hub and its neighbor nodes are highlighted

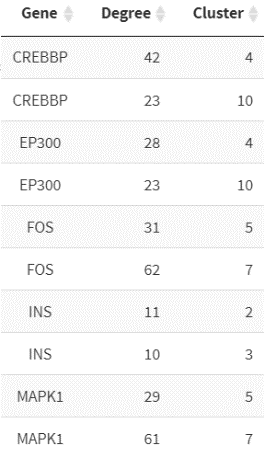
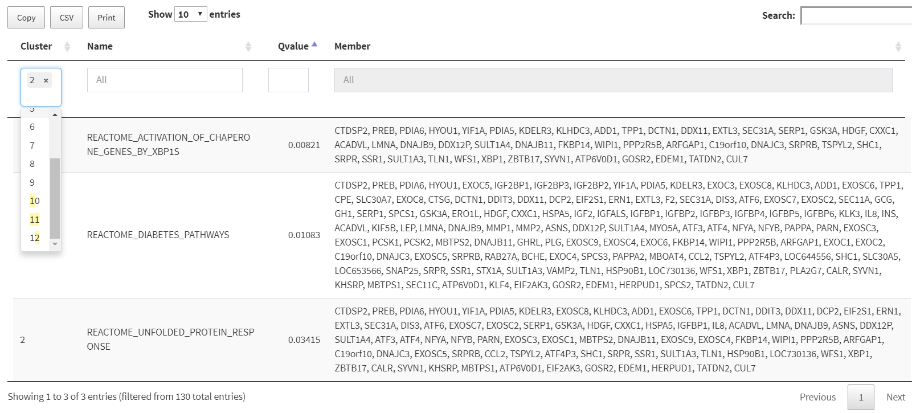
* 1. ‘Gene hub’ is visualized in the same way.
  2. ‘Multi-cluster Hub’ shows gene hubs included in at least two clusters. For example, CREBBP is a hub gene in both cluster 4 and cluster 10 with 23 and 42 neighbors, respectively.

Figure 13. Example of Multi-Cluster Hubs

1. **Clustering Results.**



Using DT (Datatable), GScluster also provides table form of gene-set clusters. Filters for each column are available: cluster number, Gene-set name or Gene-set Qvalue.

4. Additional R functions of GScluster.

1. DownloadData (species) : Used to download the STRING network data for other species. Currently nine species are supported: arabidopsis, celegans, eColi, fly, human, mouse, rat, yeast, and zebrafish. For human, Hippie 2.0 networks (<http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/>) are also provided by the name “hippie.RData” for GScluster.
2. GetGSASeqData

This function is used to transform GSAseq ‘s result into GScluster input format. It takes parameter as filename of GSAseq result.

1. GetGSASNP2Data

This function is used to transform GSA-SNP2’s result into GScluster input format. It takes parameter as filename of GSA-SNP2 result.