**GScluster User Manual**

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

2019-01-03

1. **Install and Launch**

GScluster is provided as an independent R package, which can be installed via Github ( [**github.com/unistbig/GScluster**](https://github.com/unistbig/GScluster) ). Also planed to submit on Bioconductor.

Not only the Core packages, additional data to run GScluster can be also downloaded through Github ([**github.com/unistbig/GScluster-Data/**](https://github.com/unistbig/GScluster-Data/)) or implemented Function (**DownloadData**) .

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

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

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

**> library(GScluster)**

**> GScluster()**

\* Notice that Manual was built based on Human sample data ( DIAGRAM consortium data) with StringDB for PPI

**\* IMPORTANT, library “rcytoscapejs” is on updating their version, which isn’t finished yet. So user need to install the legacy version of rcytoscapejs package manually from** [**github.com/cytoscape/cyjShiny/releases**](https://github.com/cytoscape/cyjShiny/releases)

1. **How to use**

GScluster takes 7 input parameters.

1. GSAresult : object for GSA result file. Which is a matrix, consists with 3 or 4 columns.

Geneset Name : name for gene-sets

Genelist : member gene of each gene-sets

Qvalues : Each Gene-sets score ( optional ? )

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.

Notice header exists and Genes should be separated with white space.

1. GeneScores : objects for DE analysis result. Which is a matrix, consists with 2 columns.

Gene : Gene name, only Gene symbol will be accepted.

Score : DE result, numeric value with range 0 ~ 1. Exponential format like 1E-3 is possible.

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

Figure 2. Example for score.

Notice header doesn’t exist.

1. PPI : objects for Protein-Protein Interaction, which is matrix with numeric value 0~1 with row , column name is Gene symbol. default is StringDB Human data.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **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 PPI

Notice it’s symmetric and Diagonal value is 0.

1. Species : single character for indicate species of data. Default is H

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

1. Alpha : numeric value between 0 ~ 1. Default is 1.

Network weight value for pMM distances. So if Alpha is 0, 0 % of network will be used, so same with MM. if alpha is 1, 100% of network will be used.

1. GsQCutoff : numeric value between 0 ~ 1. Default is 0.25

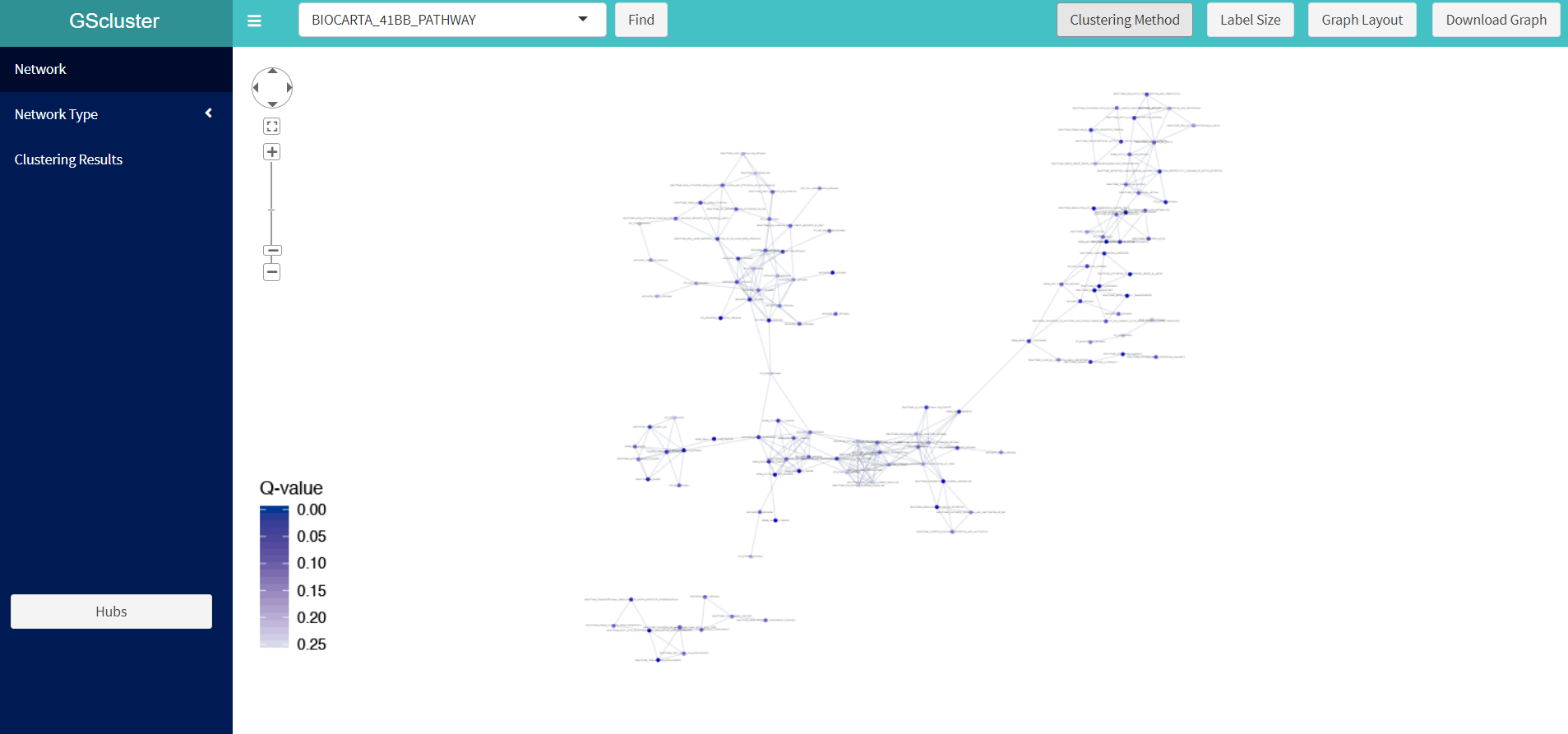
This parameter stands Geneset Qvalue cutoff, Gene-sets with larger qvalue with cutoff will not included in GScluster.

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

This parameter stands Genescore Qvalue cutoff, Genes with larger qvalue with cutoff will not included in GScluster.

**3. Exploring the Network**

After run GScluster, the Gene-set network graph will be displayed. Default gene-set clustering options are Method : pMM, Seed size : 3, Distance : same percentage value of 50% value of MM. The detailed clustering result is represented as table 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**

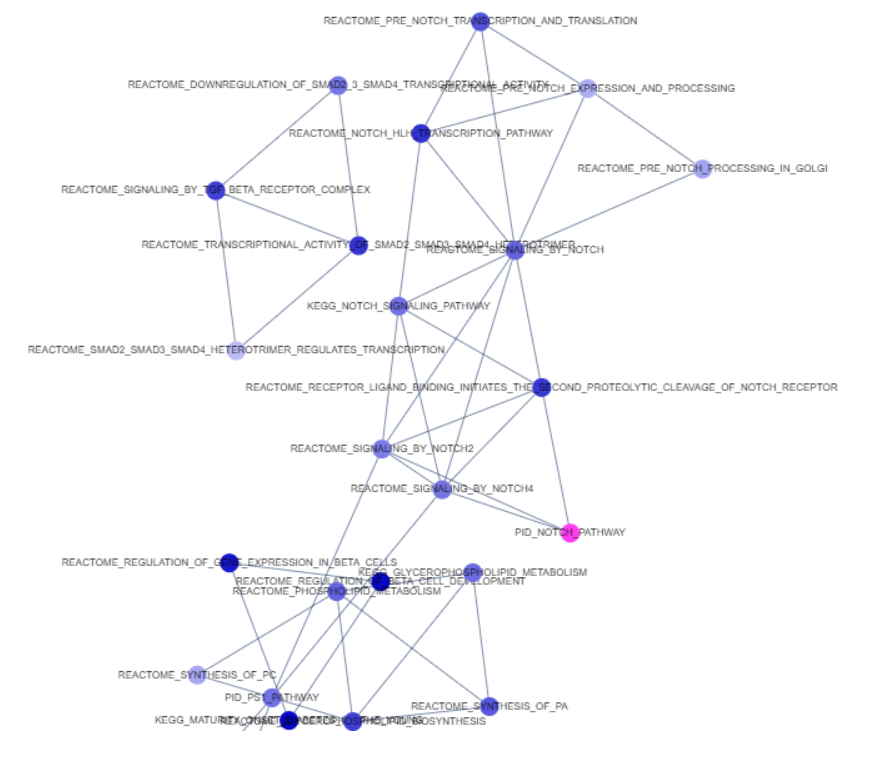
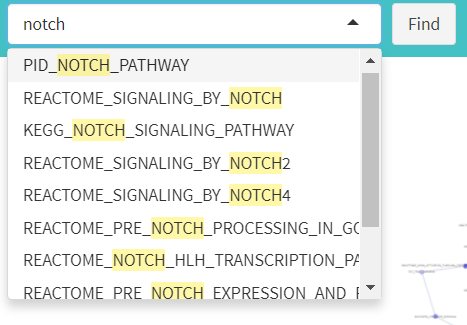
**↓⑧**

**↑⑨**

Figure 4. Result panel. Each number indicates useful function introduced below.

**GScluster Functions**

1. **Graph control panel:** Users can zoom in/out, fit or move the graph by simple mouse control or using graph control panel (① in Fig 4)in the top left of the result panel. 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 for few seconds. (Fig. 5).



**1. Type a keyword**

**2. Select target & Find**

**3. Node will Highlighted**

Figure 5. Search for the position of 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’** to change the gene-set network graph. We present detailed explanation for each parameter.

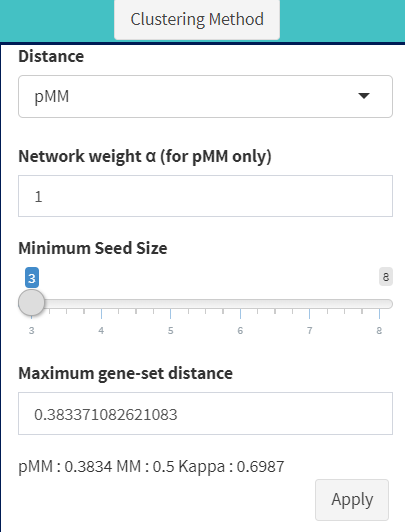
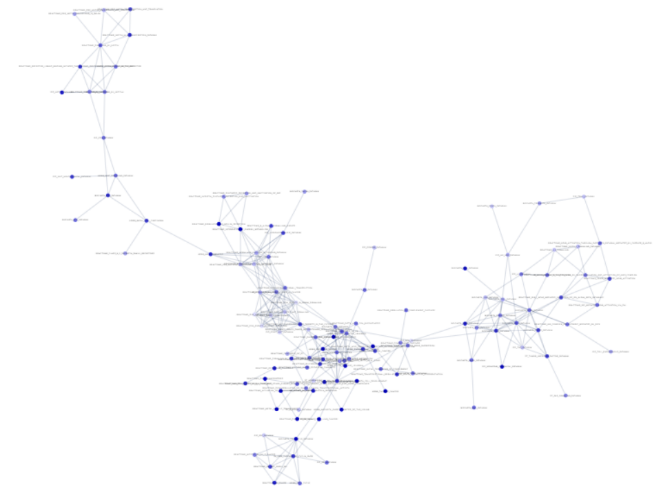
****

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 (Fg. 7).

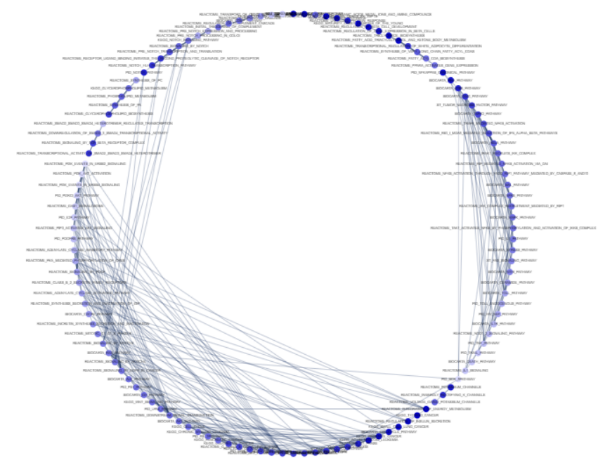
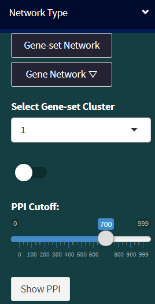


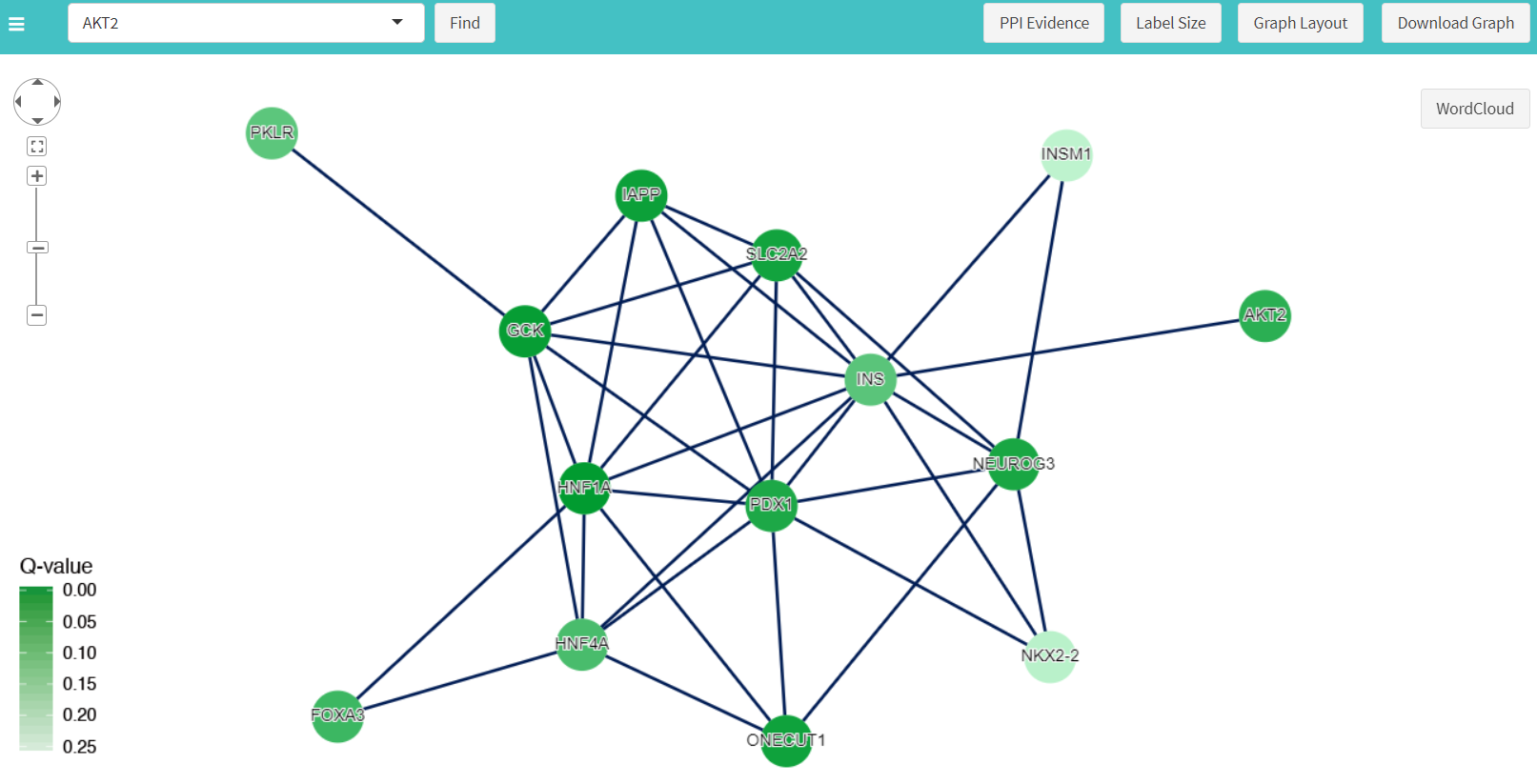
Figure 7. Circle (left) and Cola (right) layout

1. **Download Graph:** Users can download a vector image file (.SVG) for current plot by clicking ‘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 cluster based on STRING human PPI data. For example, if you want to see the gene network in cluster 1, do 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 most strongest )
   4. Click **‘Show PPI’** button. Then it will show the network for genes in cluster 1.
   5. Change to Gene network will show 2 additional function (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 which gathered from DisGeNET ([www.disgenet.org](http://www.disgenet.org)), notice that this can be apply with Human Data.it will show wordcloud any common disease (frequent > 1) between significant genes

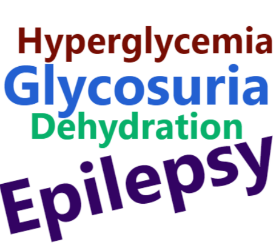


Figure 10. Wordcloud generated from a specific cluster

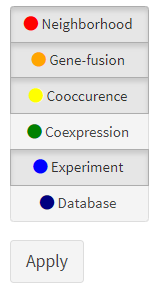
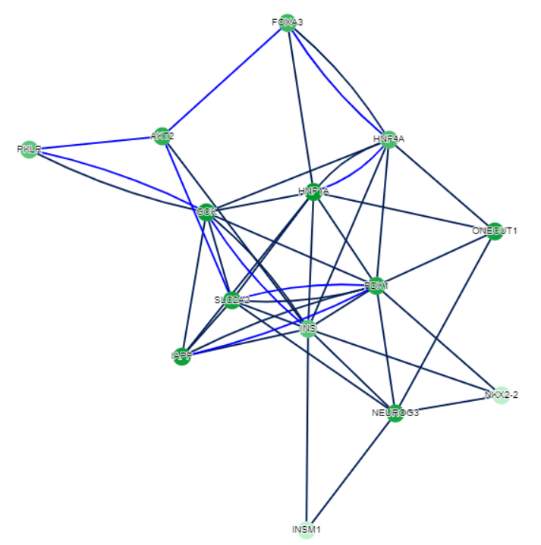
* 1. PPI Evidence.

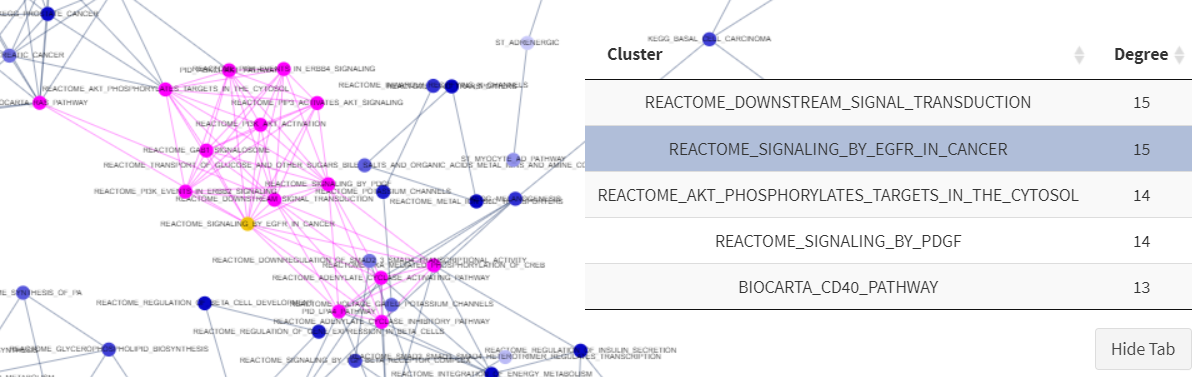
Figure 11. PPI Evidence and it’s result

PPI Evidence will show additional information by edge**.** GScluster provide 6 edge types such as Neighborhood, Gene fusion, Co-occurrence, Co-expression, Experiments, Databases. Detailed explanation for each PPI evidence type is described in STRING web page ( <https://string-db.org/cgi/help.pl> ), notice basic edge is Combined score and Text-mining evidence has discarded. In Example, 4 evidence were selected but only Existing Experiment evidence showed.

1. Switch in midpoint (① in Fig 8) is used to Highlight gene-sets. It will highlight given number (**in Select Gene-set Cluster**) cluster gene-set nodes. In example, gene-set nodes in cluster 1 will highlighted with Magenta color.
2. Gene-set Network Button will restore gene-set network which recently calculated.
3. **Hub**

GScluster supports 3 types of Hub. (Hub means highly connected Node or has high degree ).

* 1. Geneset Hub, it will show top 5 hub Gene-set in Gene-set network. By click them in table, network will highlighted Hub : yellow, Hub neighbor : magenta. Clicking network background ( empty space ) will cancle highligt.

Figure 12. Geneset hub and it’s neighbor nodes are highlighted

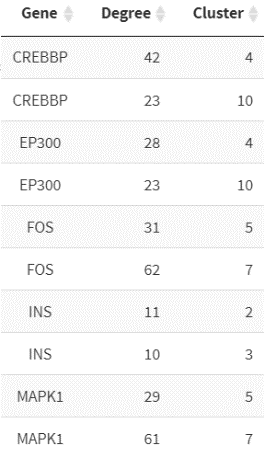
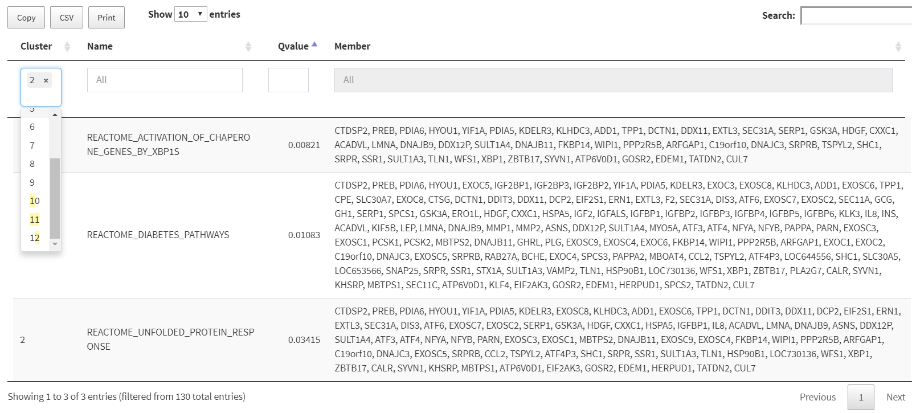
* 1. Gene hub, work same with Gene-set hub except target is Gene not Gene-set
  2. Multi-cluster Hub, it will show Gene hubs within at least two clusters. For example result, gene CREBBP is hub gene in both cluster 4 and cluster 10 with 23 and 42 linked genes.

Figure 13. Example of Multi-Cluster Hub

1. **Clustering Results.**



Using DT (Datatable), GScluster also provides table form of Gene-set cluster. Which can apply various filters : cluster number, Gene-set name or Gene set Qvalue.

4. Additional R functions of GScluster.

1. DownloadData (species) : Used to Download additional PPI Data for another species.  
   possible value for species is arabidopsis, celegans, eColi, fly, human, mouse, rat, yeast, zebrafish. if species is human, additional PPI processed from Hippie 2.0 (<http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/>) will downloaded with name “hippie.RData”. which can be load and used directly to GScluster.
2. GetGSASeqData

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