MetaOmics Lab

1 Download and install software

1.1 Run from docker image **(Recommended)**

Install docker: https://www.docker.com/

Run command in terminal:

- docker pull metaomics/app
- docker run --rm --name metaOmics -p 3838:3838 metaomics/app

Then you can access the **MetaOmics** pipeline by opening http://127.0.0.1:3838/metaOmics/ on your web browser.

Users can check the docker container ID by bash command **docker ps**.

The default working directory of the docker container is **/srv/shiny-server/metaOmics/**.

To download all output after analysis, run command **docker cp containerID:/srv/shiny-server/metaOmics/local_path_to_output_folder** in terminal.

(Go to https://docs.docker.com/engine/reference/commandline/run/ or use **docker run –help** in terminal for more instructions about docker.)

1.2 Install the app in R

a) Download software from Github.

Download the zip file from https://github.com/metaOmics/metaOmics by clicking on **"Clone or download"** and extract to a working directory.

Or type **git clone https://github.com/metaOmics/metaOmics** in the command line under your working directory

b) Open R and install "shiny" in the R console:

install.packages('shiny'), choose a CRAN mirror and install the package

c) Starting up:

In R, run **shiny::runApp('metaOmics', port=9987, launch.browser=T)**

d) Set the working directory in the "**Setting**" page:

Under "**Directory for Saving Output Files**", select a local directory to automatically save the outputs to.

e) Install dependent R packages of the modules:

Under "**Toolsets**", click to install the desired modules if the "**status**" shows "**not installed**". The installation progress is updated in the notification icon. After modules are installed to R, restart the shiny application again so that the interface is updated with the installed modules. Now a new "**Toolsets**" tab is shown on the browser that contains all installed modules.

However, this shiny app can be safely run in R version 3.3, but for other R version, you may encounter errors in installing dependencies of the modules.

**For more details, please refer to the original MetaOmics paper reference[1].*

2 Prepare Data

2.1 Expression Data

The gene-expression matrix should be prepared as tab-delimited ".txt" or comma-separated ".csv" files. The first column corresponds to the feature ID (e.g., gene symbol, probe ID, or entrez ID) and the rest of columns are the expression data from samples. The first row contains the sample ID. For gene expression profiling from a microarray platform, normalized log-transformed continuous intensities are the default input format. For RNA-seq, both raw count data (e.g., those generated by HTSeq or bedtools) and continuous data (e.g. FPKM, RPKM or TPM converted using Cufflinks) are allowed. Since conversion from count data to FPKM/TPM requires genome annotation and other information that is constantly updated, this task is expected to be done by users before data input.

2.2 Clinical Data

The first column of clinical data set corresponds to the sample ID, and the rest of columns contain the clinical information of the samples (e.g., case/control labels). Sample IDs of the clinical data (on rows) should be ordered in the same way as the gene-expression data (on columns) to avoid any mismatch issues.

2.3 Example Data

There are 3 datasets (Verhaak, et al., 2009; Balgobind, et al., 2010; Kohlmann, et al., 2008) using samples from acute myeloid leukemia (AML) with three known chromosomal translocation subtypes as summarized in **Table 1**: "**inv(16)**" (inversions in chromosome 16), "**t(15;17)**" (translocationsbetween chromosome 15 and 17), "**t(8;21)**" (translocations between chromosome 8 and 21). These AML subtypes have been well studied with different survival, treatment response and prognosis outcomes. **The data is under the folder "/data/example/leukemia".**

Our main purpose was to demonstrate the use of the **MetaClust** module to simultaneously cluster samples of the three studies and see if the clustering can reproduce the three well-known subtypes. Meanwhile, we will also illustrate how to use **MeteQC** module for quality control, **MetaDE** module for DE analysis and simple pathway analysis, as well as **MetaNetwork** for network analysis.

[1] Ma T, Huo Z, Kuo A, et al. MetaOmics: analysis pipeline and browser-based software suite for transcriptomic meta-analysis. *Bioinformatics*. 2019;35(9):1597-1599. doi:10.1093/bioinformatics/bty825

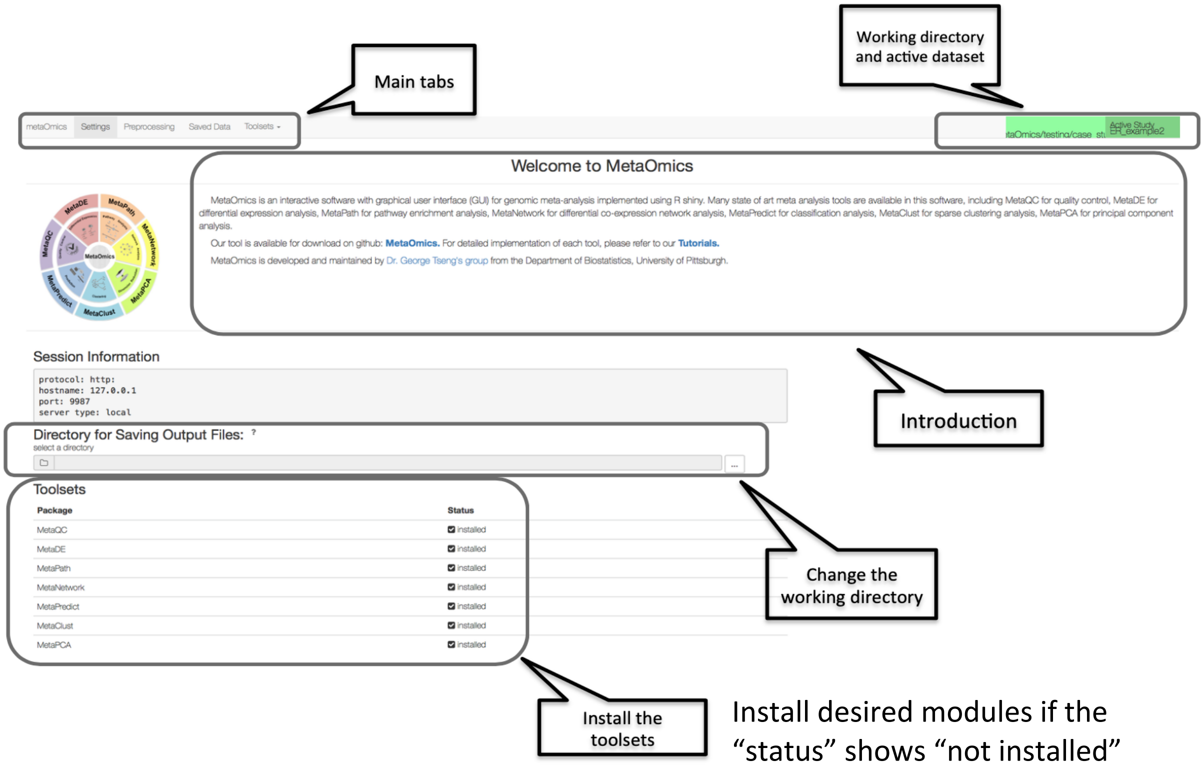
Table 1: Multi-study acute myeloid leukemia (AML) gene-expression profiles. All three studies are from Affymetrix Human Genome U133plus2 with 5,135 genes. Three subtypes of leukemia are defined as the chromosomal translocation, including inversion of chromosome 16 - inv(16), translocation of chromosome 15 and 17 - t(15:17) and translocation of chromosome 8 and 21 - t(8:21).

Study	Source	# Samples	# Samples by subtypes
			inv(16)/t(15:17)/t(8,21)

Study 1	Verhaak et al. (2009)	89	33/21/35
Study 2	Balgobind et al. (2010)	74	27/19/28
Study 3	Kohlmann et al. (2008)	105	28/37/40

3 Setting

After modules are installed, restart the shiny application again so that the interface is updated with the installed modules. Now a new "Toolsets" tab is shown on the browser that contains all installed modules.



Welcome to MetaOmics

MetaOmics is an interactive software with graphical user interface (GUI) for genomic meta-analysis implemented using R shiny. Many state of art meta analysis tools are available in this software, including MetaQC for quality control, MetaDE for differential expression analysis, MetaPath for pathway enrichment analysis, MetaNetwork for differential co-expression network analysis, MetaPredict for classification analysis, MetaClust for sparse clustering analysis, MetaPCA for principal component analysis.

Our tool is available for download on github: **MetaOmics.** For detailed implementation of each tool, please refer to our **Tutorials.**

MetaOmics is developed and maintained by Dr. George Tseng's group from the Department of Biostatistics, University of Pittsburgh.

Session Information

```
protocol: http:
hostname: 127.0.0.1
port: 9987
server type: local
```

Directory for Saving Output Files: ?
select a directory

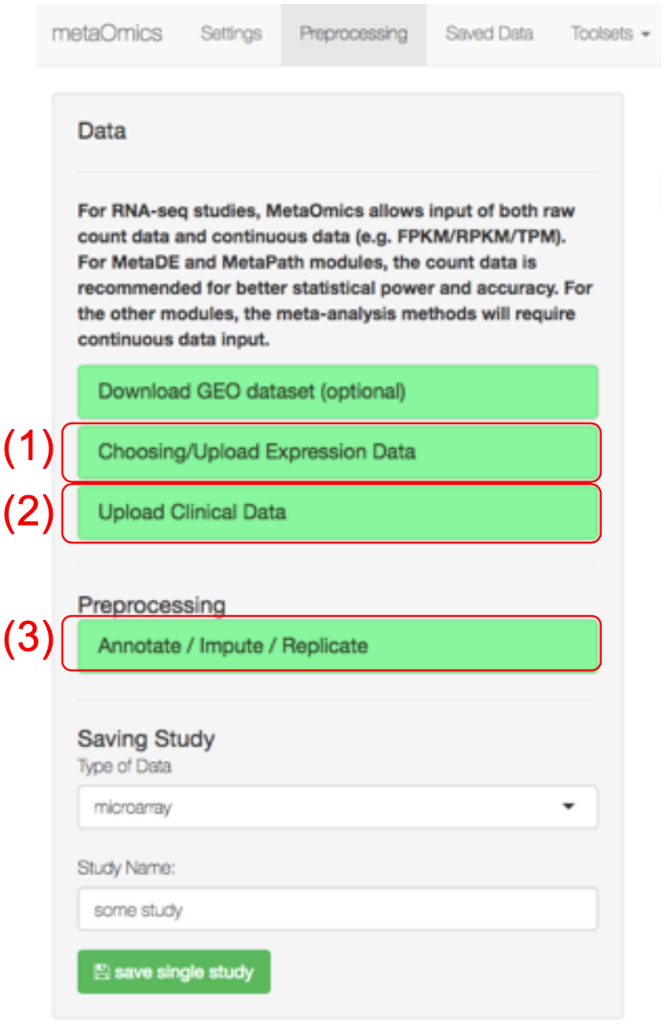
Toolsets

Package	Status
MetaQC	☑ installed
MetaDE	☑ installed
MetaPath	☑ installed
MetaNetwork	☑ installed
MetaPredict	☑ installed
MetaClust	☑ installed
MetaPCA	☑ installed

Install desired modules if the "status" shows "not installed"

4 Meta Preprocessing

4.1 Step 1 Uploading data:



For RNA-seq studies, MetaOmics allows input of both raw count data and continuous data (e.g. FPKM/RPKM/TPM). For MetaDE and MetaPath modules, the count data is recommended for better statistical power and accuracy. For the other modules, the meta-analysis methods will require continuous data input.

Type of Data: microarray

Study Name: some study

- Go to the **Preprocessing** page
- Upload expression data files or choose the existing saved data files by clicking (1)
- Optionally upload clinical data by clicking (2)
- Optionally preprocess the data for feature annotation, missing value imputation, and multiple probes reduction for the same gene in (3)
- By clicking the “**save single study**” button, the gene-expression profile will be uploaded, and the data can be previewed on the right side of the page.

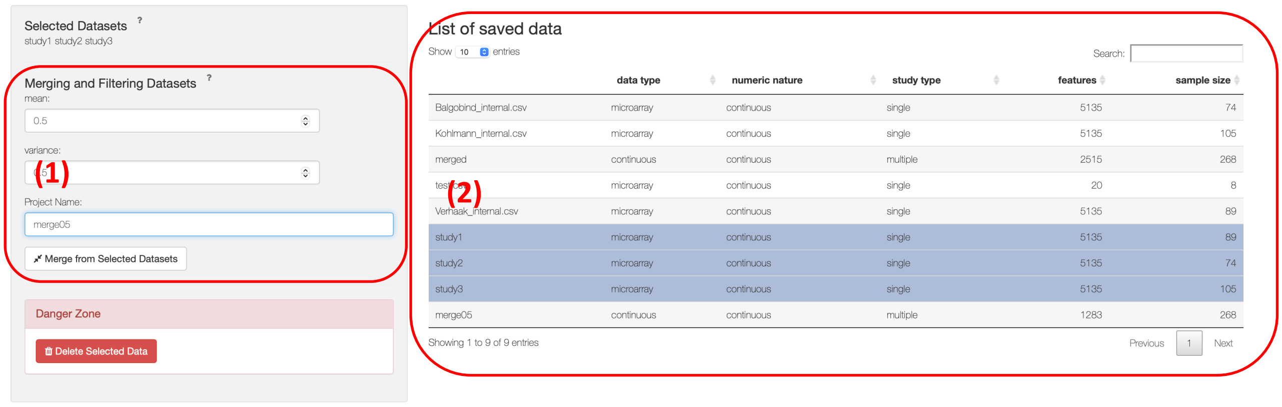
4.2 Step 2 Merge datasets and make active dataset

After uploading all studies with clinical data, turn to the **Saved Data** page. We are creating 3 merged data sets with different filtering criteria

1) Select the 3 datasets we uploaded. By (1), filter out genes with low expression level (here we use mean expression lower than 80th percentile) or low variance (here we use variance lower than 80th percentile), and name the merged data set as “**merge08**”. The merged data only keeps 206 genes.

2) Similarly, select the 3 studies, and filter out genes with mean expression lower than 50th percentile or variance lower than 50th percentile. Name the merged data set as “**merge05**”, and it keeps 1283 genes

3) Select the merged data set and click (3) to set it as the active study that shows up on the top-right corner of the page. The active dataset serves as the input for all the analytical modules.



Selected Datasets ?
study1 study2 study3

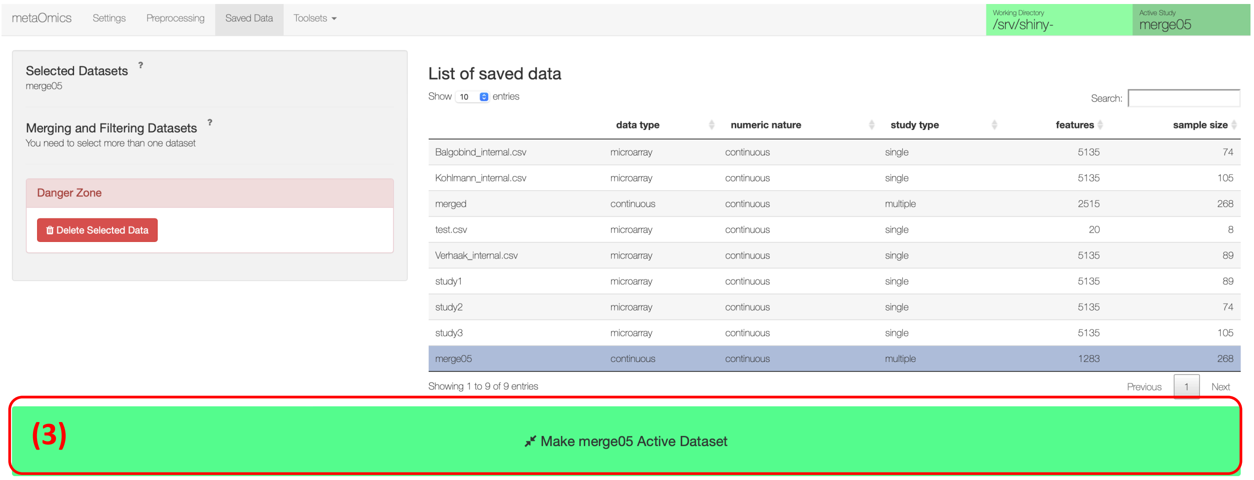
Merging and Filtering Datasets ?
mean:
0.5
variance:
0.5
Project Name:
merge05

List of saved data

Show 10 entries Search:

	data type	numeric nature	study type	features	sample size
Balgobind_internal.csv	microarray	continuous	single	5135	74
Kohlmann_internal.csv	microarray	continuous	single	5135	105
merged	continuous	continuous	multiple	2515	268
test.csv	microarray	continuous	single	20	8
Verhaak_internal.csv	microarray	continuous	single	5135	89
study1	microarray	continuous	single	5135	89
study2	microarray	continuous	single	5135	74
study3	microarray	continuous	single	5135	105
merge05	continuous	continuous	multiple	1283	268

Showing 1 to 9 of 9 entries



Working Directory
/srv/shiny-

Active Study
merge05

Selected Datasets ?
merge05

Merging and Filtering Datasets ?
You need to select more than one dataset

List of saved data

Show 10 entries Search:

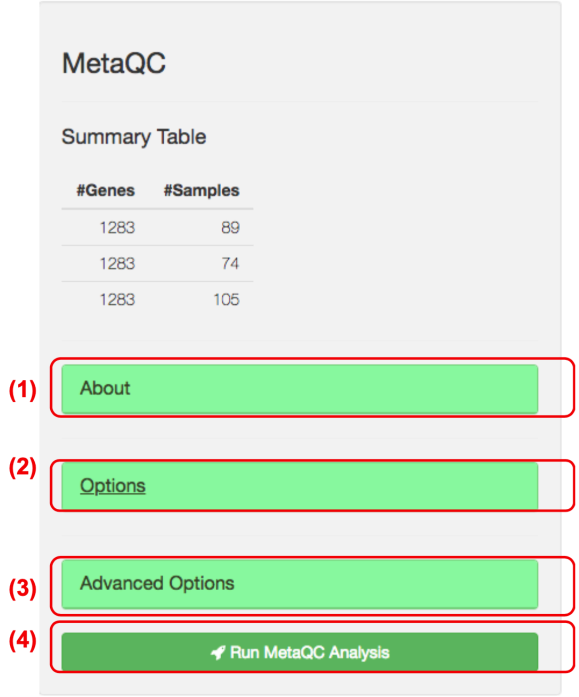
	data type	numeric nature	study type	features	sample size
Balgobind_internal.csv	microarray	continuous	single	5135	74
Kohlmann_internal.csv	microarray	continuous	single	5135	105
merged	continuous	continuous	multiple	2515	268
test.csv	microarray	continuous	single	20	8
Verhaak_internal.csv	microarray	continuous	single	5135	89
study1	microarray	continuous	single	5135	89
study2	microarray	continuous	single	5135	74
study3	microarray	continuous	single	5135	105
merge05	continuous	continuous	multiple	1283	268

Showing 1 to 9 of 9 entries

5 Meta Analysis

5.1 MetaQC

The **MetaQC** module provides an objective and quantitative tool to help determine the inclusion/exclusion of studies for meta-analysis. More specifically, **MetaQC** provides users with six quantitative quality control (QC) measures: internal homogeneity of co-expression structure among studies (**IQC**); external consistency of co-expression pattern with pathway databases (**EQC**); accuracy and consistency of differentially expressed gene detection (**AQCg** and **CQCg**) as well as accuracy and consistency of enriched pathway identification (**AQCp** and **CQCp**). In addition, visualization plots and summarization tables are generated using principal component analysis (PCA) biplots and standardized mean ranks (**SMR**) to assist in visualization and decision.

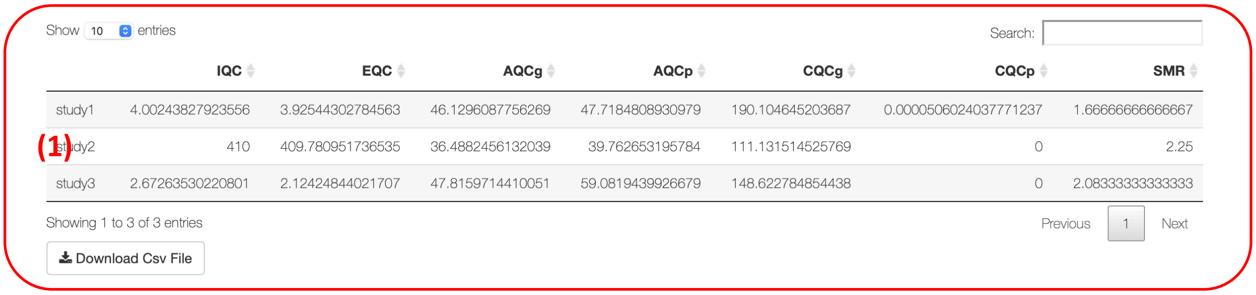


MetaQC

Summary Table

#Genes	#Samples
1283	89
1283	74
1283	105

- Go to **Toolsets** drop-down menu and select **MetaQC** module
- (1) provides the description about this module
- Drop-down menu in (2) can perform gene filtering to reduce computational cost, specify the approach and cutoff to select potentially DE genes, and specify the approach and cutoff to select potentially enriched pathways
- Drop-down menu in (3) can tune other parameters, and it is suggested not to modify the option setting in this section without knowing the method
- Click (4) to perform **MetaQC.**



	IQC	EQC	AQCg	AQCp	CQCg	CQCp	SMR
study1	4.00243827923556	3.92544302784563	46.1296087756269	47.7184808930979	190.104645203687	0.0000506024037771237	1.66666666666667
study2	410	409.780951736535	36.4882456132039	39.762653195784	111.131514525769	0	2.25
study3	2.67263530220801	2.12424844021707	47.8159714410051	59.0819439926679	148.622784854438	0	2.08333333333333

Showing 1 to 3 of 3 entries

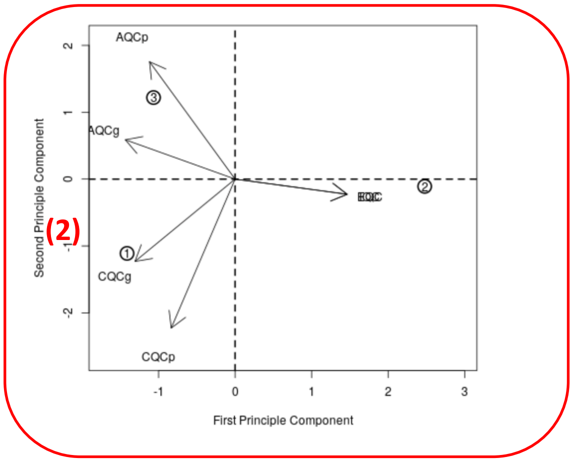


Figure 1

Figure 1: **MetaQC** Results with default parameters for merged dataset "**merge05**". *Figure 1* **(1)** includes seven columns, with the first six columns corresponding to the six quantitative quality control measures of all studies (a larger value indicates a better quality), and the seventh column is the rank of summary statistics of all six quality measures (a lower rank indicates a better quality).

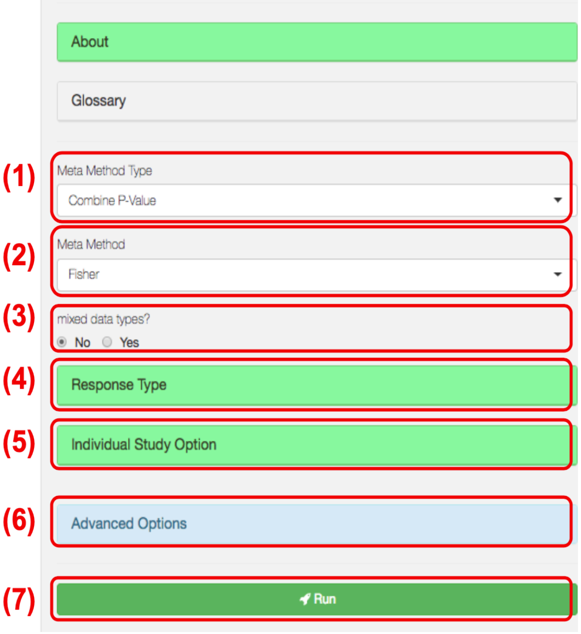
In addition, **MetaQC** also generates a PCA biplot (*Figure 1* **(2)**) based on the six quality control measures, where the circled number is the study index and arrows indicate different measures. If a study (a circled number) is along the direction of the majority of the six QC directions, then the study has higher quality and is consistent with other studies. In general, if a study has larger standard mean rank (**SMR**) values, it is considered to have lower quality and is inconsistent with other studies. We can see here the arrows pointed to different directions and none of the 3 studies are clear outliers. This is also reflected in the **SMR** values, where all the 3 studies have similar ranks. Therefore, we cannot exclude any study.

5.2 MetaDE

The **MetaDE** module implements 12 major meta-analysis methods with 22 variations for differential expression analysis that fall into three main categories: combining p-values, combining effect sizes, and others (e.g., combining ranks, etc.). Depending on the type of outcome, the package can perform two-classes comparison, multi-class comparison, and association with continuous or survival outcomes. The package allows the input of either microarray (continuous intensity) and/or RNA-seq data (count or FPKM/RPKM) for individual study analysis.

After obtaining differentially expressed (DE) genes from the differential expression analysis, users can further perform post-hoc pathway enrichment analysis using the declared DE genes.

5.2.1 Differential expression analysis



Meta Method Type: Combine P-Value

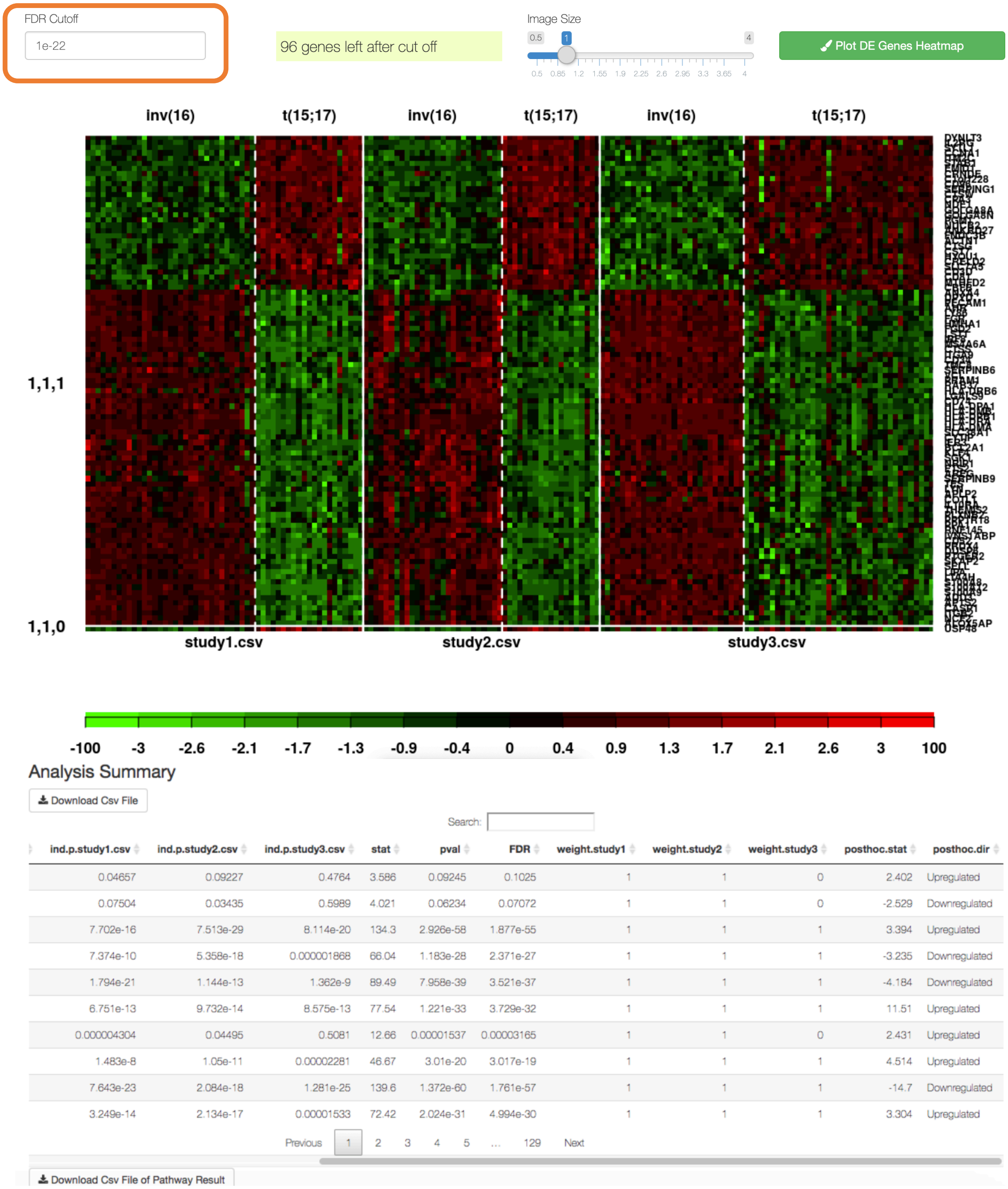
Meta Method: Fisher

mixed data types? ● No ○ Yes

- Go to **Toolsets** drop-down menu and select **MetaDE** module
- Click (1) to choose the type of meta-analysis method from: combining p-values, combining effect sizes and others.
- Given the selected meta method type, choose a meta-analysis method from (2).
- Click No in (3), unless the data combine both RNA-seq and microarray studies.
- (4) is to specify the outcome variable, and control/case group label.
- (5) specifies the data type of each individual study and DE analysis method. For continuous data (e.g., microarray), available methods include LIMMA (default) and SAM. For discrete data (e.g., RNA-seq count), available methods are edgeR, DESeq2, and Voom.
- Default settings in (6) is suggested.

Figure 2: **MetaDE** results for "**merge05**" between group "**inv(16)**" (control) and "**t(15,17)**" (case). The heatmap of DE genes is rendered after specifying the **FDR Cutoff** for selection of DE genes and clicking on **"Plot DE Genes Heatmap"**. The image size can be adjusted by dragging the scrolling bar. In the heatmap, rows refer to the declared DE genes under the specified FDR cutoff, columns refer to samples, and solid white lines are used to separate different studies. The dashed white lines are used to separate groups. Colors of the cells correspond to scaled expression level, as indicated in the color key below. For the results generated by "**AW-Fisher∗**", there is one additional column of cross-study weight distribution on the left end of the heatmap, and the genes in the heatmap are sorted by their weight distribution. We can see that the majority of DE genes are commonly up-regulated or down-regulated (weight=1,1,1), indicating a generally homogeneous signal across the 3 studies.

The summary table of meta-analysis results is at the bottom, including information of test statistic, p-value of individual study, meta-analysis p-value, FDR, etc.



FDR Cutoff: 1e-22

96 genes left after cut off

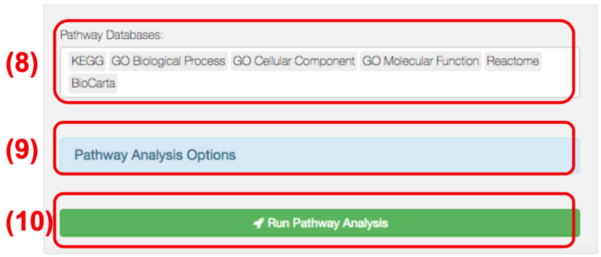
Analysis Summary

ind.p.study1.csv	ind.p.study2.csv	ind.p.study3.csv	stat	pval	FDR	weight.study1	weight.study2	weight.study3	posthoc.stat	posthoc.dir
0.04657	0.09227	0.4764	3.586	0.09245	0.1025	1	1	0	2.402	Upregulated
0.07504	0.03435	0.5989	4.021	0.06234	0.07072	1	1	0	-2.529	Downregulated
7.702e-16	7.513e-29	8.114e-20	134.3	2.926e-58	1.877e-55	1	1	1	3.394	Upregulated
7.374e-10	5.358e-18	0.000001868	66.04	1.183e-28	2.371e-27	1	1	1	-3.235	Downregulated
1.794e-21	1.144e-13	1.362e-9	89.49	7.958e-39	3.521e-37	1	1	1	-4.184	Downregulated
6.751e-13	9.732e-14	8.575e-13	77.54	1.221e-33	3.729e-32	1	1	1	11.51	Upregulated
0.000004304	0.04495	0.5081	12.66	0.00001537	0.00003165	1	1	0	2.431	Upregulated
1.483e-8	1.05e-11	0.00002281	46.67	3.01e-20	3.017e-19	1	1	1	4.514	Upregulated
7.643e-23	2.084e-18	1.281e-25	139.6	1.372e-60	1.761e-57	1	1	1	-14.7	Downregulated
3.249e-14	2.134e-17	0.00001533	72.42	2.024e-31	4.994e-30	1	1	1	3.304	Upregulated

Figure 2

5.2.2 Pathway analysis

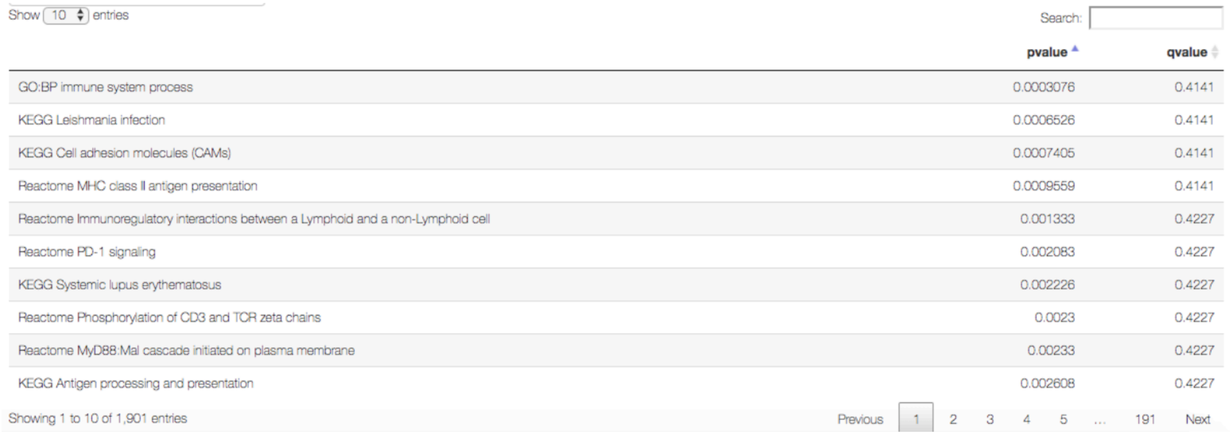
Then, we can perform post-hoc pathway enrichment analysis.



Pathway Databases: KEGG GO Biological Process GO Cellular Component GO Molecular Function Reactome BioCarta

- Choose the pathway database in (8) from 25 available pathway databases.
- In options (9), users need to choose the pathway method from Kolmogorov-Smirnov (KS) test (default option), or the Fisher's exact test, and specify the minimum/maximum gene size of pathways to be included.

Figure 3: the downstream pathway analysis results based on **MetaDE** genes. Each row represents a pathway with its p-value and q-value listed on the right.



Show 10 entries

Search:

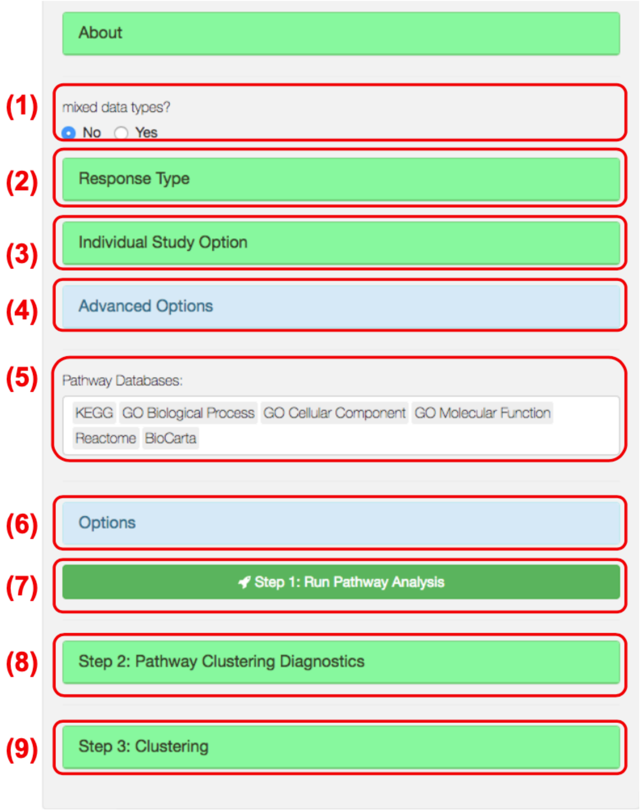
	pvalue	qvalue
GO:BP immune system process	0.0003076	0.4141
KEGG Leishmania infection	0.0006526	0.4141
KEGG Cell adhesion molecules (CAMs)	0.0007405	0.4141
Reactome MHC class II antigen presentation	0.0009559	0.4141
Reactome Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	0.001333	0.4227
Reactome PD-1 signaling	0.002083	0.4227
KEGG Systemic lupus erythematosus	0.002226	0.4227
Reactome Phosphorylation of CD3 and TCR zeta chains	0.0023	0.4227
Reactome MyD88:Mal cascade initiated on plasma membrane	0.00233	0.4227
KEGG Antigen processing and presentation	0.002608	0.4227

Showing 1 to 10 of 1,901 entries

Figure 3

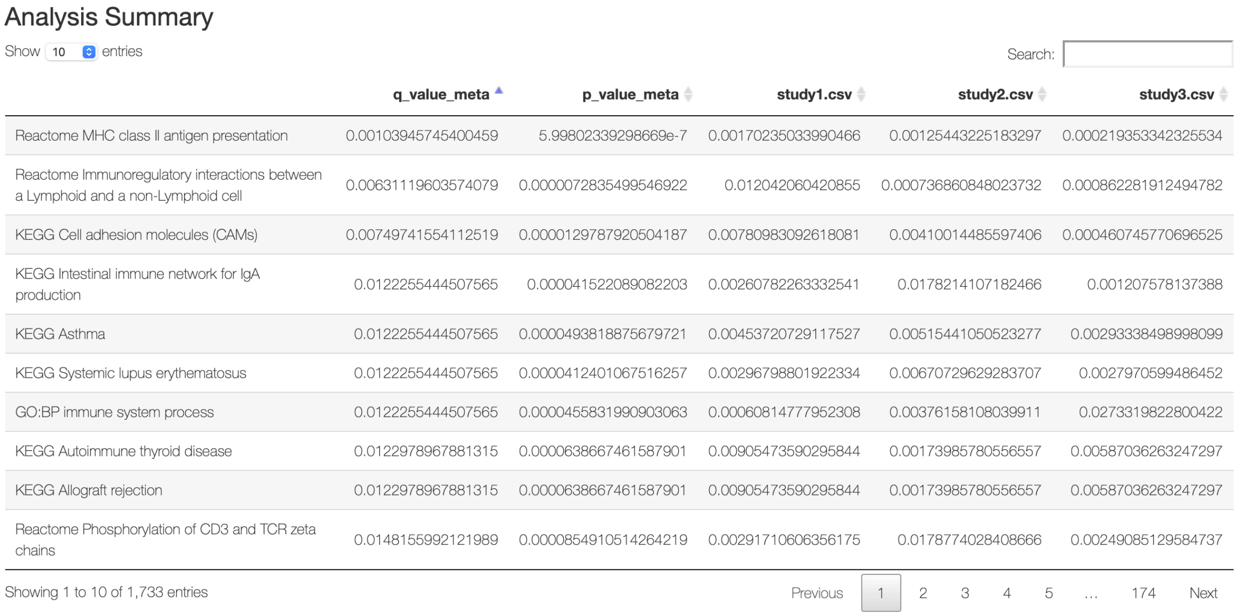
5.3 MetaPath

The **MetaPath** module performs pathway analysis by two advanced meta-analytic pathway analysis tools: Meta-Analysis for Pathway Enrichment (MAPE) and Comparative Pathway Integrator (CPI) (Shen et al., 2010; Fang et al., 2017). Although both **MetaDE** and **MetaPath** could perform pathway enrichment analysis, they are different. **MetaDE** only provides more traditional downstream pathway analysis for the functional annotation of detected DE genes from meta-analysis. On the other hand, **MetaPath** uses more comprehensive and sophisticated methods to jointly perform DE analysis and pathway analysis, and it provides stronger statistical power and more extensive and intuitive biological insights. Also, **MetaPath** performs additional pathway clustering to reduce pathway redundancy and extracts the key words from each cluster via the text mining algorithm to assist with the interpretation.



mixed data types? ● No ○ Yes

- *Step 1:* specify (1) whether the input gene-expression profile is a mix of continuous data and discrete data; (2) response type, case/control labels (similar to **MetaDE**); (3) individual study option (similar to **MetaDE**); (4) advanced options, including whether to adjust for covariates or the direction of hypothesis testing; (5) pathway databases for the enrichment analysis; (6) the method (either MAPE by default or CPI). Then click on (7) to **Run Pathway Analysis**.
- *Step 2:* specify the top enriched pathways by choosing the FDR cutoff in (8). Then, run **Pathway Clustering Diagnostics** to perform consensus clustering analysis to determine the optimum number of clusters K.
- *Step 3:* specify the number of clusters and click on (9) to get pathway clustering results.



Analysis Summary

Show 10 entries

Search:

	q_value_meta ▲	p_value_meta	study1.csv	study2.csv	study3.csv
Reactome MHC class II antigen presentation	0.00103945745400459	5.99802339298669e-7	0.00170235033990466	0.00125443225183297	0.000219353342325534
Reactome Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	0.00631119603574079	0.0000072835499546922	0.012042060420855	0.000736860848023732	0.000862281912494782
KEGG Cell adhesion molecules (CAMs)	0.00749741554112519	0.0000129787920504187	0.00780983092618081	0.00410014485597406	0.000460745770696525
KEGG Intestinal immune network for IgA production	0.0122255444507565	0.000041522089082203	0.00260782263332541	0.0178214107182466	0.001207578137388
KEGG Asthma	0.0122255444507565	0.0000493818875679721	0.00453720729117527	0.00515441050523277	0.00293338498998099
KEGG Systemic lupus erythematosus	0.0122255444507565	0.0000412401067516257	0.00296798801922334	0.00670729629283707	0.0027970599486452
GO:BP immune system process	0.0122255444507565	0.0000455831990903063	0.00060814777952308	0.00376158108039911	0.0273319822800422
KEGG Autoimmune thyroid disease	0.0122978967881315	0.0000638667461587901	0.00905473590295844	0.00173985780556557	0.00587036263247297
KEGG Allograft rejection	0.0122978967881315	0.0000638667461587901	0.00905473590295844	0.00173985780556557	0.00587036263247297
Reactome Phosphorylation of CD3 and TCR zeta chains	0.0148155992121989	0.0000854910514264219	0.00291710606356175	0.0178774028408666	0.00249085129584737

Showing 1 to 10 of 1,733 entries

Figure 4

Figure 4 is the MetaPath analysis summary of “**merge05”** data after *Step 1* by choosing the **method CPI**. This table shows analysis results of all pathways, including individual study association analysis p-value, meta pathway analysis p-value/FDR, etc. Click the p value meta “up arrow” button to sort these pathways and search the pathway name in the search bar. The full table is automatically saved in the working directory specified previously.

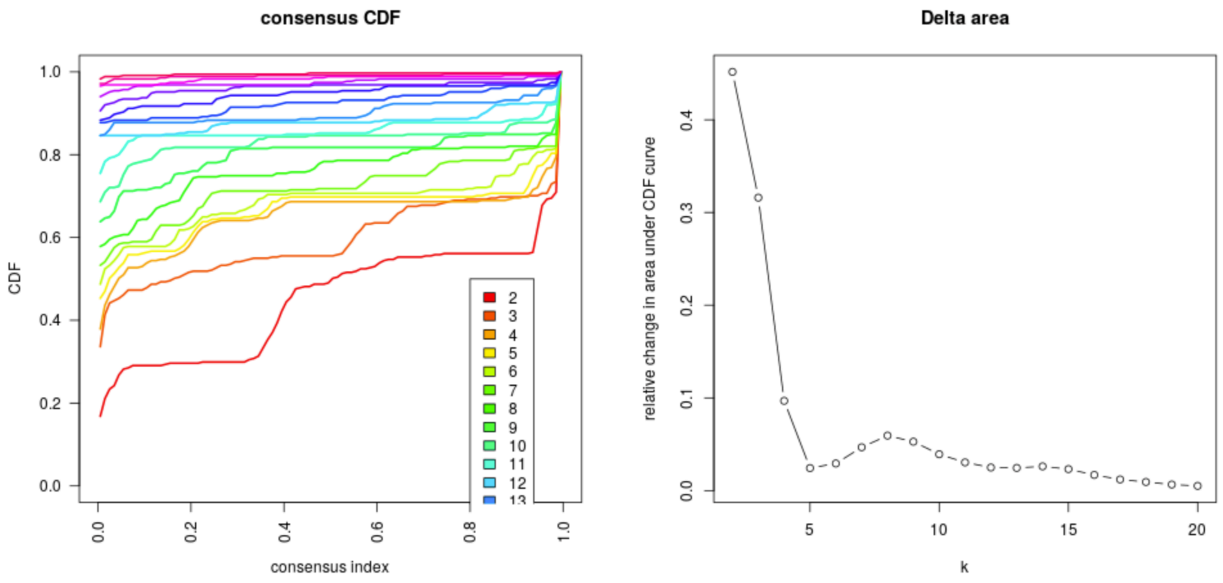


Figure 5

Figure 5 shows the results after *Step 2* by setting the **FDR cutoff as 0.4** which selects 27 pathways. Both of the consensus CDF plot and Delta area plot assist in finding the optimal number of clusters K (refer to Monti et al. (2003) for detailed interpretation of the two plots). To be brief, the cumulative density function (CDF) of the consensus matrix for each K (indicated by colors) is estimated by a histogram of 100 bins. The CDF reaches an approximate maximum, implying consensus and cluster confidence is at a maximum at this K. The Delta area shows the relative change in area under the CDF curve comparing K and K − 1, thus allowing to determine K at which there is no appreciable increase in CDF (which drops as the number of cluster increases). In the example, K = 5 is chosen since it locates at the elbow turning point (i.e., where the magnitude of incremental decrease in delta area diminishes)

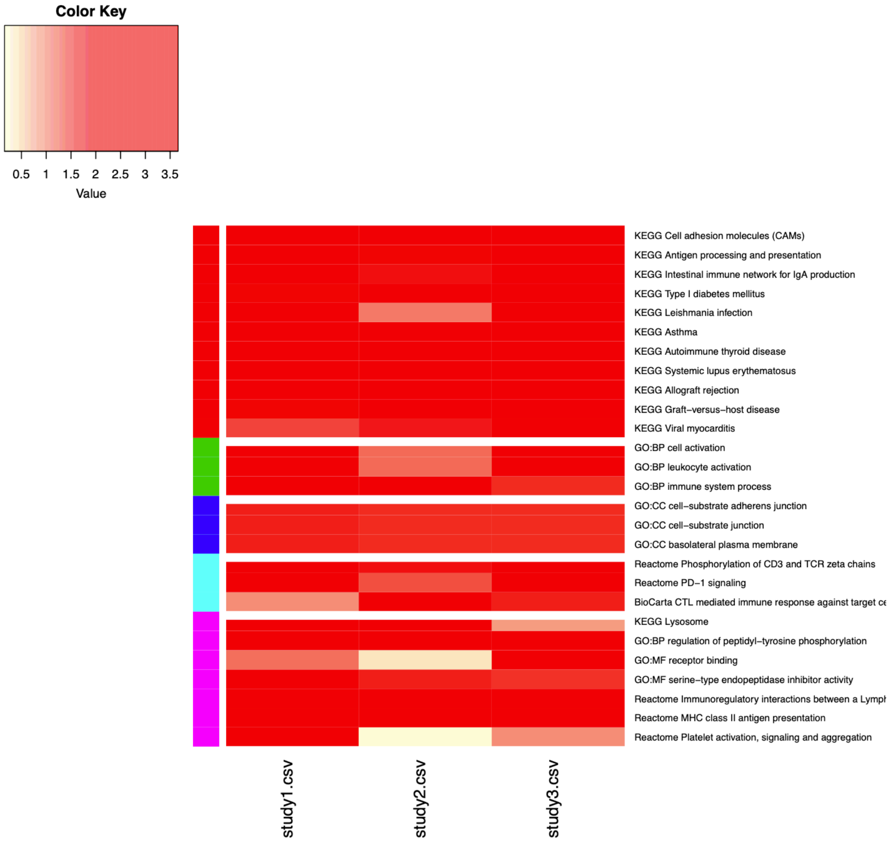


Figure 6

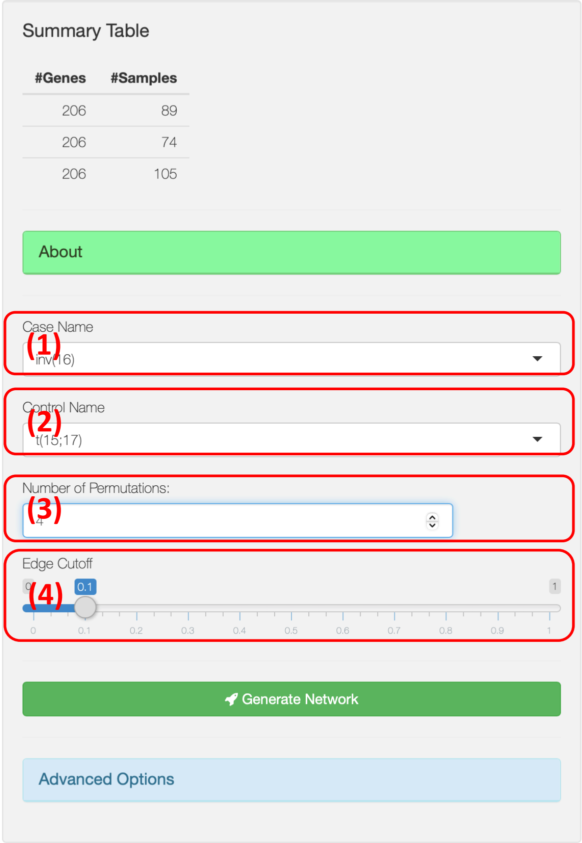
The heatmap in *Figure 6* shows the -log10 transformed p-value of enrichment analysis in each study from *Step 3*. Studies are on columns, and the selected pathways are on rows. The color bar on left indicates group of pathways (K=5). In the heatmap, red indicates "more enriched," and yellow indicates "not significantly enriched." The color key is on the top left corner. The pathways are sorted by the pathway clusters, as indicated by the colors on the left side of the heatmap. In addition, key words of each cluster of pathways are extracted and analyzed by a built-in text mining algorithm. One file named "Clustering Summary.csv" is saved to the working directory, which shows a summary of the text-mining results.

5.4 MetaNetwork

MetaNetwork aims to detect gene modules with different co-expression patterns under different biological conditions. We use the correlation to quantify the co-expression level. High correlation in a gene module, either positive or negative, means the genes inside this module are functionally related, which may indicate they are controlled by the same transcriptional regulatory program, or member of the same pathway. The module includes three steps to get differentially co-expressed networks, including generating networks, searching for basic modules, and assembling super-modules.

5.4.1 Generate Network

After clicking the **Generate Network** button, the screen will show a message indicating the algorithm is running to generate the network.



Summary Table

#Genes	#Samples
206	89
206	74
206	105

(1) Case Name: inv(16)

(2) Control Name: t(15;17)

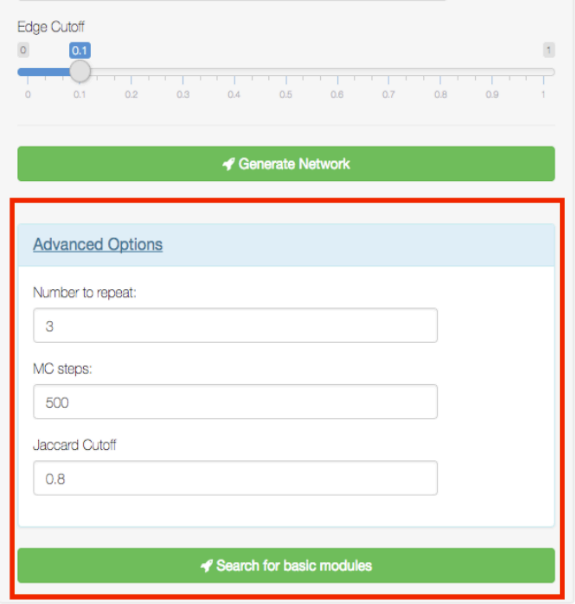
(3) Number of Permutations: 4

(4) Edge Cutoff: 0.1

- Go to **Toolsets** drop-down menu and select **MetaNetwork** module
- Select case and control group name by (1) (2).
- Set the number of permutations in (3) that are used to generate the null distribution for edge energy and will be further used to calculate edge FDR. Given a reasonable number of edges, 3-10 permutations are recommended.
- The edge cutoff in (4) determines the proportion of edges to be kept in the network. Only edges with correlations above this cutoff will be kept as connected. Decreasing the edge cutoff will result in a denser network and add computation time. We recommend starting with a large cutoff (looser network), especially for large numbers of genes and then gradually decreasing it (increasing network density) for a desirable network.

5.4.2 Search for Basic Modules

When searching for basic modules, we try multiple repeats with different seeds to avoid local optimum. After clicking the **Search for basic modules** button, the screen will show a message indicating the algorithm is running to search for basic modules. After this step is done, tables summarizing two kinds of basic modules: one highly connected in case group but loosely connected in control group, and the other one with reverse pattern.


Edge Cutoff: 0.1

Advanced Options

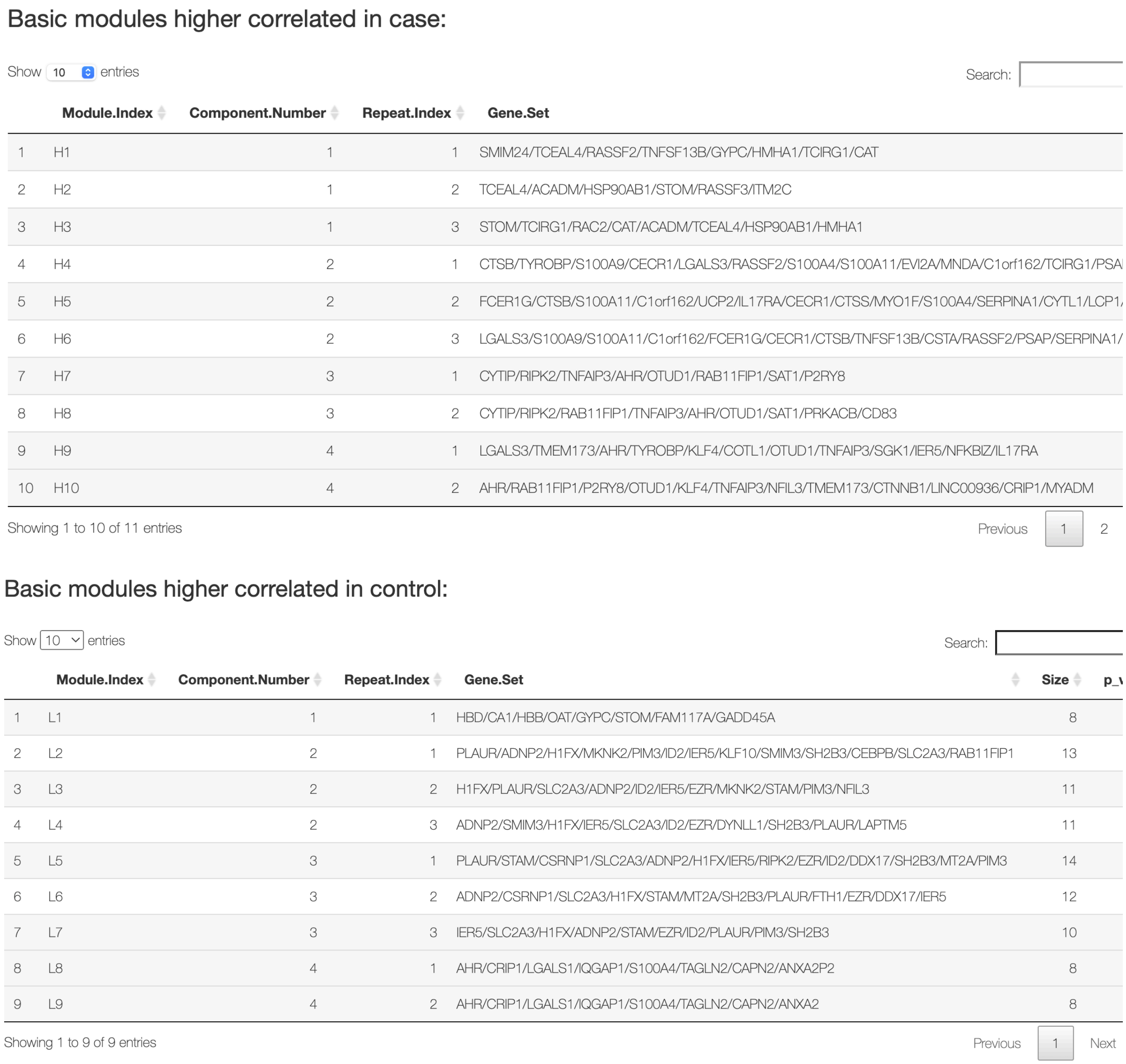
Number to repeat: 3

MC steps: 500

Jaccard Cutoff: 0.8

- Advanced options are recommended not to change (including the number of repeats used for each initial seed modules, the maximum Monte Carlo steps for the simulated annealing algorithm, and the maximum pairwise Jaccard index allowed for basic modules).

Since **MetaNetwork** requires a large computing time, we use the merged leukemia data “**merge08**” (keeping 206 genes), 4 permutations, and other parameters as default. In this example we only compared two phenotypes: “**inv(16)”** and “**t(15;17)**”. In general, the **MetaNetwork** module is time consuming for large datasets (for both the **Generate Network** and **Search for basic modules** steps). We should carefully restrict the number of genes (e.g., less than a thousand) for a test run before applying them to a large gene set. After the **Generate Network** step completes, no output will show up on the screen. Instead, a message box will show up indicating several .Rdata files are saved in the **MetaNetwork** folder under the working directory. After the **Search for basic modules step** is done, the screen will show a table of basic modules higher correlated in case or control, as in *Figure 4*.



Basic modules higher correlated in case:

	Module.Index	Component.Number	Repeat.Index	Gene.Set
1	H1	1	1	SMIM24/TCEAL4/RASSF2/TNFSF13B/GYPC/HMHA1/TCIRG1/CAT
2	H2	1	2	TCEAL4/ACADM/HSP90AB1/STOM/RASSF3/ITM2C
3	H3	1	3	STOM/TCIRG1/RAC2/CAT/ACADM/TCEAL4/HSP90AB1/HMHA1
4	H4	2	1	CTSB/TYROBP/S100A9/CECR1/LGALS3/RASSF2/S100A4/S100A11/EVI2A/MNDA/C1orf162/TCIRG1/PSA
5	H5	2	2	FCER1G/CTSB/S100A11/C1orf162/UCP2/IL17RA/CECR1/CTSS/MYO1F/S100A4/SERPINA1/CYTL1/LCP1/
6	H6	2	3	LGALS3/S100A9/S100A11/C1orf162/FCER1G/CECR1/CTSB/TNFSF13B/CSTA/RASSF2/PSAP/SERPINA1/
7	H7	3	1	CYTIP/RIPK2/TNFAIP3/AHR/OTUD1/RAB11FIP1/SAT1/P2RY8
8	H8	3	2	CYTIP/RIPK2/RAB11FIP1/TNFAIP3/AHR/OTUD1/SAT1/PRKACB/CD83
9	H9	4	1	LGALS3/TMEM173/AHR/TYROBP/KLF4/COTL1/OTUD1/TNFAIP3/SGK1/IER5/NFKBIZ/IL17RA
10	H10	4	2	AHR/RAB11FIP1/P2RY8/OTUD1/KLF4/TNFAIP3/NFIL3/TMEM173/CTNNB1/LINC00936/CRIP1/MYADM

Showing 1 to 10 of 11 entries

Basic modules higher correlated in control:

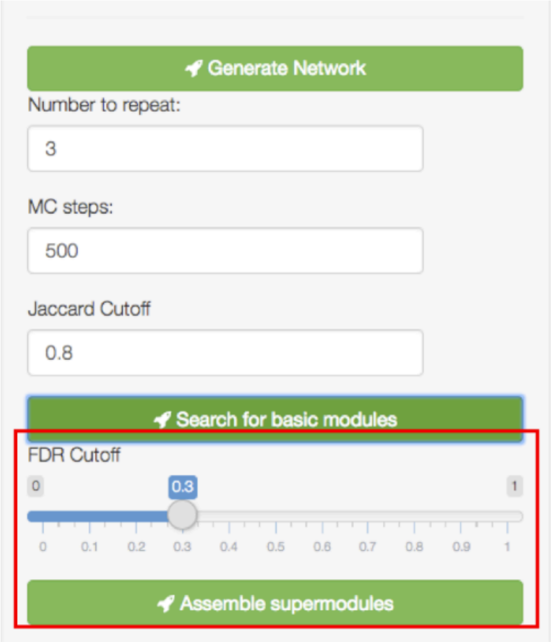
	Module.Index	Component.Number	Repeat.Index	Gene.Set	Size	p_v
1	L1	1	1	HBD/CA1/HBB/OAT/GYPC/STOM/FAM117A/GADD45A	8	
2	L2	2	1	PLAUR/ADNP2/H1FX/MKNK2/PIM3/ID2/IER5/KLF10/SMIM3/SH2B3/CEBPB/SLC2A3/RAB11FIP1	13	
3	L3	2	2	H1FX/PLAUR/SLC2A3/ADNP2/ID2/IER5/EZR/MKNK2/STAM/PIM3/NFIL3	11	
4	L4	2	3	ADNP2/SMIM3/H1FX/IER5/SLC2A3/ID2/EZR/DYNLL1/SH2B3/PLAUR/LAPTM5	11	
5	L5	3	1	PLAUR/STAM/CSRNP1/SLC2A3/ADNP2/H1FX/IER5/RIPK2/EZR/ID2/DDX17/SH2B3/MT2A/PIM3	14	
6	L6	3	2	ADNP2/CSRNP1/SLC2A3/H1FX/STAM/MT2A/SH2B3/PLAUR/FTH1/EZR/DDX17/IER5	12	
7	L7	3	3	IER5/SLC2A3/H1FX/ADNP2/STAM/EZR/ID2/PLAUR/PIM3/SH2B3	10	
8	L8	4	1	AHR/CRIP1/LGALS1/IQGAP1/S100A4/TAGLN2/CAPN2/ANXA2P2	8	
9	L9	4	2	AHR/CRIP1/LGALS1/IQGAP1/S100A4/TAGLN2/CAPN2/ANXA2	8	

Showing 1 to 9 of 9 entries

Figure 7

Figure 7: **MetaNetwork** output from **Search for basic modules** step, summarizing two kinds of basic modules: one highly connected in cases but loosely connected in control (labeled as H1, . . ., H11), and the other one with reverse pattern (labeled as L1, . . ., L9). The actual gene sets are listed for each basic module.

5.4.3 Assemble Supermodules



Number to repeat: 3

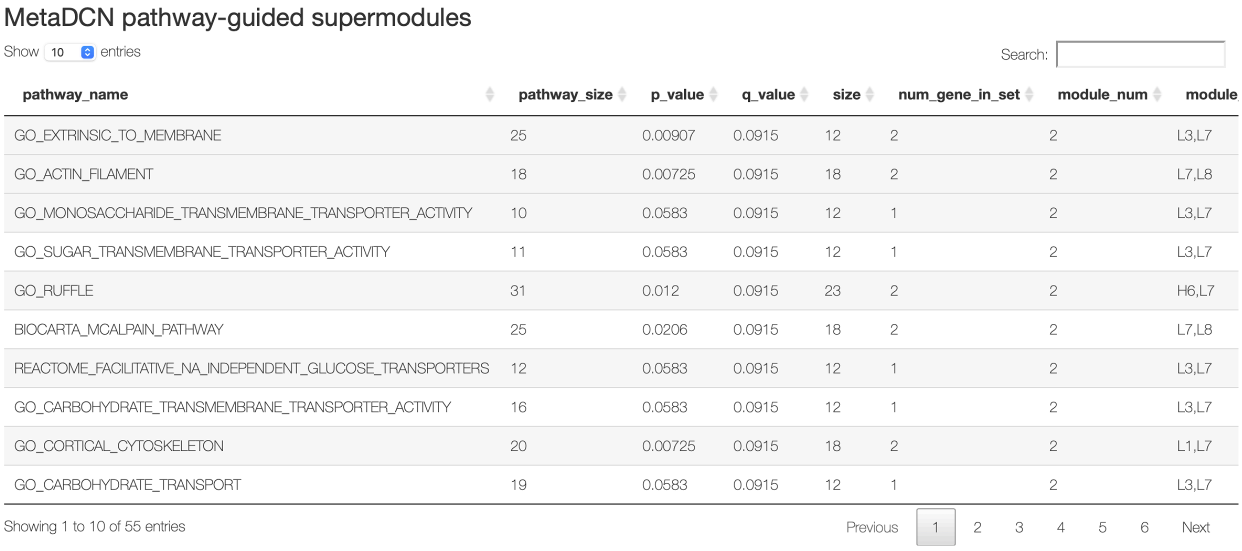
MC steps: 500

Jaccard Cutoff 0.8

FDR Cutoff 0.3

- Decide the **FDR Cutoff** to select basic modules for super-module assembly.
- After clicking **Assemble supermodules** button, the screen will show message indicating the algorithm is running to assemble supermodules. A table for basic modules, supermodules, and their network visualization will be shown on the right panel of the screen.

After the **Assemble supermodules** complete, the screen will show a table of super-modules (*Figure 5*). Users can also select basic modules to plot (*Figure 6*). Meanwhile, in the **MetaNetwork** folder under the working directory, the files of top super-modules in text format designed to input to a Cytoscape plug-in "MetaDCNExplorer" (http://tsenglab.biostat.pitt.edu/software.htm) are automatically generated for improved visualization and dynamic exploration.



MetaDCN pathway-guided supermodules

Show 10 entries

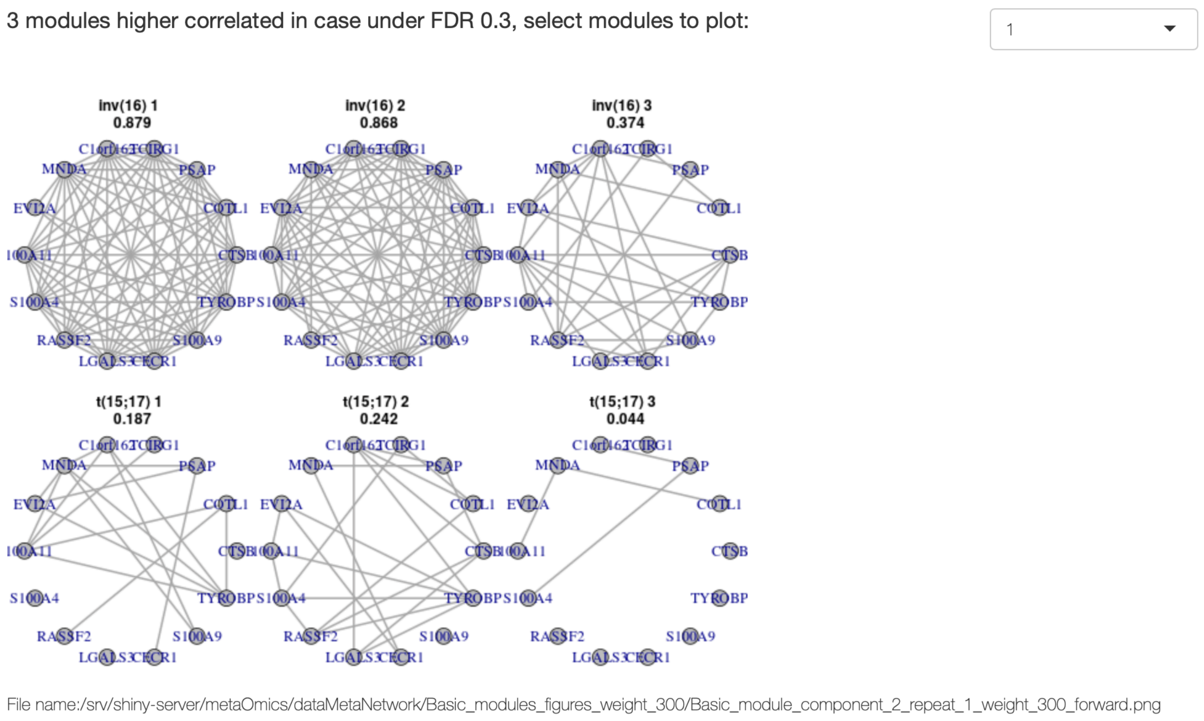
Search:

pathway_name	pathway_size	p_value	q_value	size	num_gene_in_set	module_num	module
GO_EXTRINSIC_TO_MEMBRANE	25	0.00907	0.0915	12	2	2	L3,L7
GO_ACTIN_FILAMENT	18	0.00725	0.0915	18	2	2	L7,L8
GO_MONOSACCHARIDE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	10	0.0583	0.0915	12	1	2	L3,L7
GO_SUGAR_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	11	0.0583	0.0915	12	1	2	L3,L7
GO_RUFFLE	31	0.012	0.0915	23	2	2	H6,L7
BIOCARTA_MCALPAIN_PATHWAY	25	0.0206	0.0915	18	2	2	L7,L8
REACTOME_FACILITATIVE_NA_INDEPENDENT_GLUCOSE_TRANSPORTERS	12	0.0583	0.0915	12	1	2	L3,L7
GO_CARBOHYDRATE_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	16	0.0583	0.0915	12	1	2	L3,L7
GO_CORTICAL_CYTOSKELETON	20	0.00725	0.0915	18	2	2	L1,L7
GO_CARBOHYDRATE_TRANSPORT	19	0.0583	0.0915	12	1	2	L3,L7

Showing 1 to 10 of 55 entries

Figure 8

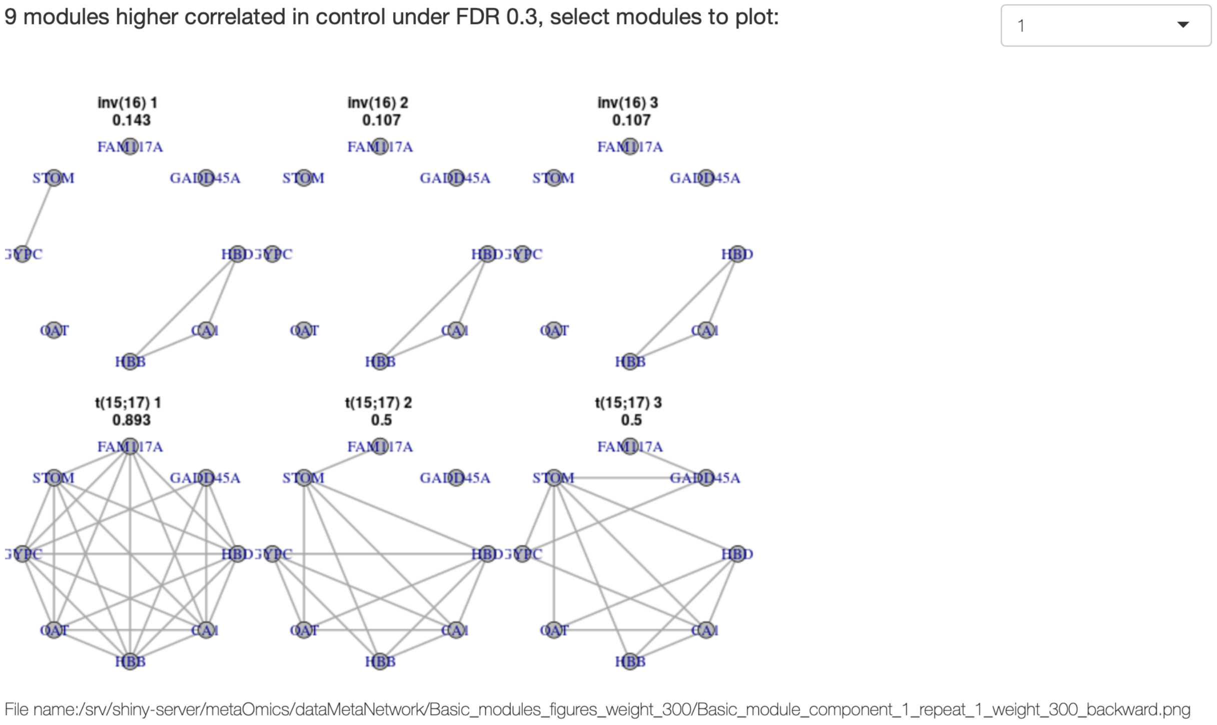
Figure 8: **MetaNetwork** supermodules table. The second column shows the pathway size. The third and fourth column show the p-value and q-value of the detected supermodule. The last column is the size of the supermodule.



3 modules higher correlated in case under FDR 0.3, select modules to plot:

1

File name:/srv/shiny-server/metaOmics/dataMetaNetwork/Basic_modules_figures_weight_300/Basic_module_component_2_repeat_1_weight_300_forward.png



9 modules higher correlated in control under FDR 0.3, select modules to plot:

1

File name:/srv/shiny-server/metaOmics/dataMetaNetwork/Basic_modules_figures_weight_300/Basic_module_component_1_repeat_1_weight_300_backward.png

Figure 9

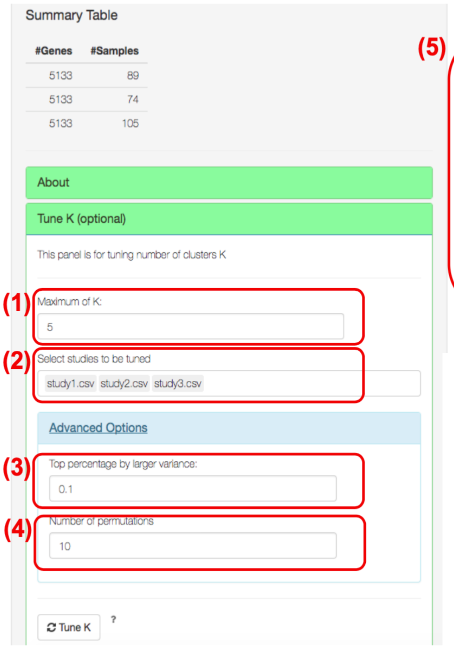
Figure 9: **MetaNetwork** select basic modules to plot. Each dot represents a gene. An edge represents the two genes are highly correlated. The network density is marked on top of each network. The top module show higher correlation in "**inv(16)**" over all the 3 individual studies, and the bottom module show high correlation in "**t(15:17)**" for all the 3 studies.

5.5 MetaClust

MetaClust module aims to perform sample clustering analysis combining multiple transcriptomic studies while ignoring the outcome labels. The resulting clustering from meta-analysis is more robust and accurate than single study analysis. It includes two optional steps before running meta sparse K-means clustering: tune **K** (the number of clusters) and tune **Wbound** (for feature selection) by gap statistic.

We used the merged leukemia data “**merge05**” again to demonstrate the **MetaClust** module. The clusters are well separated in each study based on the unified feature selected across all studies, and the cluster patterns are consistent across the studies.

5.5.1 Tune K



Summary Table

#Genes	#Samples
5133	89
5133	74
5133	105

About

Tune K (optional)

This panel is for tuning number of clusters K

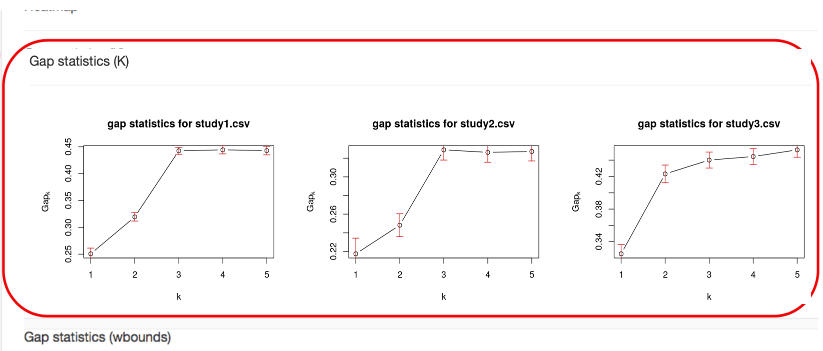
(1) Maximum of K: 5

(2) Select studies to be tuned: study1.csv study2.csv study3.csv

Advanced Options

(3) Top percentage by larger variance: 0.1

(4) Number of permutations: 10



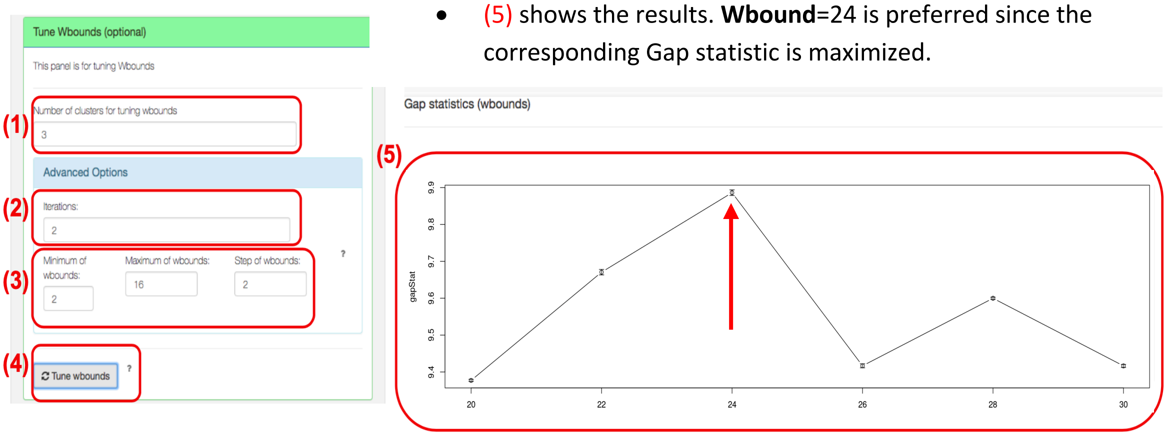
(5) Gap statistics (K)

Gap statistics (wbounds)

- Go to **Toolsets** drop-down menu and select **MetaClust** module
- (1) specifies the maximum number of K
- Use the top (3) percentage large variance genes to perform Gap statistics
- At least 50 bootstrap samples in (4) are suggested for a stable result.
- Click button **Tune K**, and (5) is the tuning results. A good K is selected such that the Gap_k is maximized or stabilized across all studies. From the figure, K = 3 is preferred since the gap statistics from all three studies become flat.

5.5.2 Tune Wbound

- (1) specifies the number of clusters (obtained from last step) for tuning **Wbound**
- Leave the advanced option (2) as default, and change the searching range of wbounds between 20 and 30 by step 2 on (3)
- (5) shows the results. **Wbound**=24 is preferred since the corresponding Gap statistic is maximized.



Tune Wbounds (optional)

This panel is for tuning Wbounds

(1) Number of clusters for tuning wbounds: 3

Advanced Options

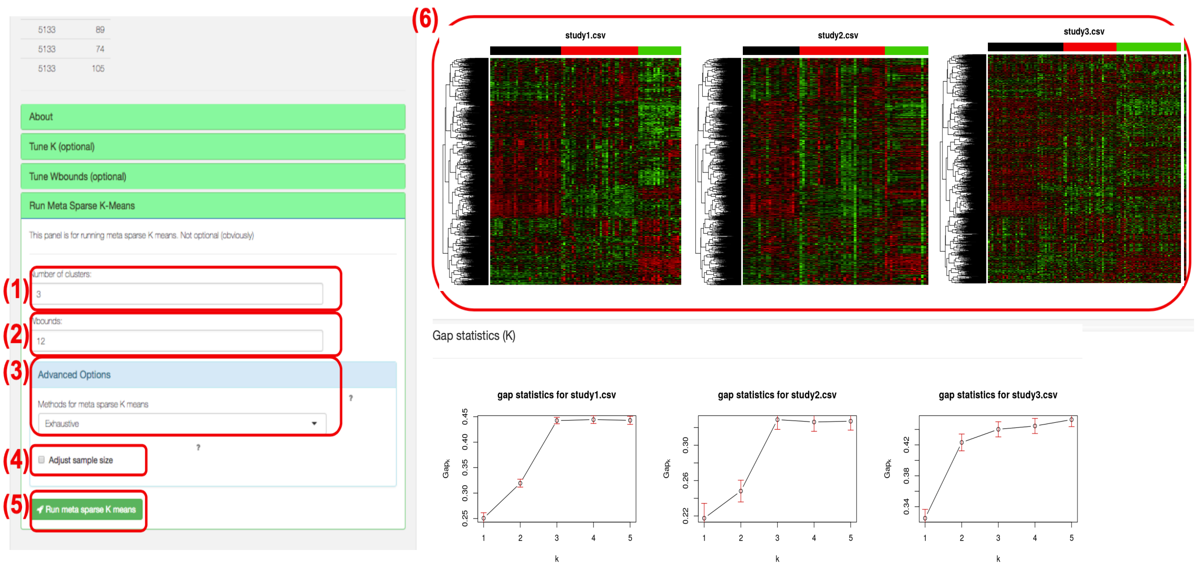
(2) Iterations: 2

(3) Minimum of wbounds: 2 | Maximum of wbounds: 16 | Step of wbounds: 2

Gap statistics (wbounds)

5.5.3 Run meta sparse K-means

Based on the results of **Tune K** and **Tune wbounds**, Run Meta Sparse K-Means by specifying the number of clusters as 3 in (1), wbounds as 24 in (2), and leaving the advanced options (3) (4) by default. The results of heatmap (6) on top shows the gene-expression profiles of the three studies with selected features. Each row represents a gene, and each column represents a sample. Note that the three studies share a common set of genes. The color bar on top of the heatmap represents the subtype labels. For instance, the black bar on top of each study represents the same subtype for all studies. Clearly, we could see distinct subtype patterns, and these patterns are consistent across studies.



5133	89
5133	74
5133	105

About

Tune K (optional)

Tune Wbounds (optional)

Run Meta Sparse K-Means

This panel is for running meta sparse K means. Not optional (obviously)

(1) Number of clusters: 3

(2) Wbounds: 12

(3) Advanced Options

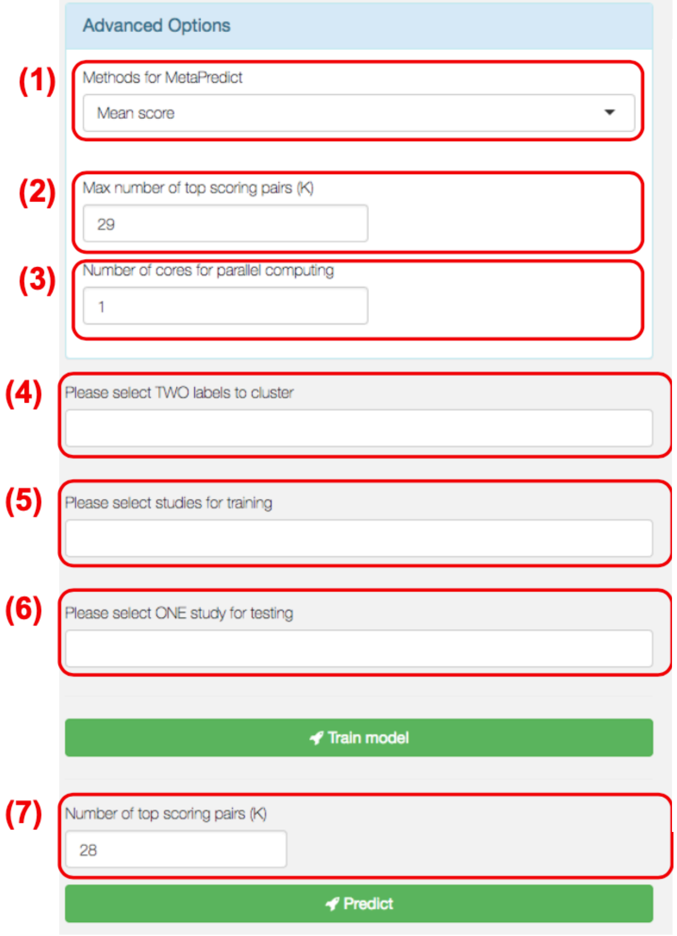
Methods for meta sparse K means: Exhaustive

(4) □ Adjust sample size

Gap statistics (K)

5.6 MetaPredict

The **MetaPredict** module combines multiple transcriptomic studies to build a prediction model and shows improved prediction accuracy as compared to single study analysis. Top scoring pairs (TSP) is a robust algorithm for predicting gene-expression profiles, which adopts non-parametric rank-based prediction rule. The **MetaPredict** is a meta-analysis version of the TSP algorithm that combines multiple transcriptomic studies to build a prediction model and shows improved prediction accuracy as compared to single study analysis. By clicking the **Toolsets** tab and then choosing **MetaPredict**, we are directed to the **MetaPredict** homepage.



Advanced Options

Methods for MetaPredict: Mean score

Max number of top scoring pairs (K): 29

Number of cores for parallel computing: 1

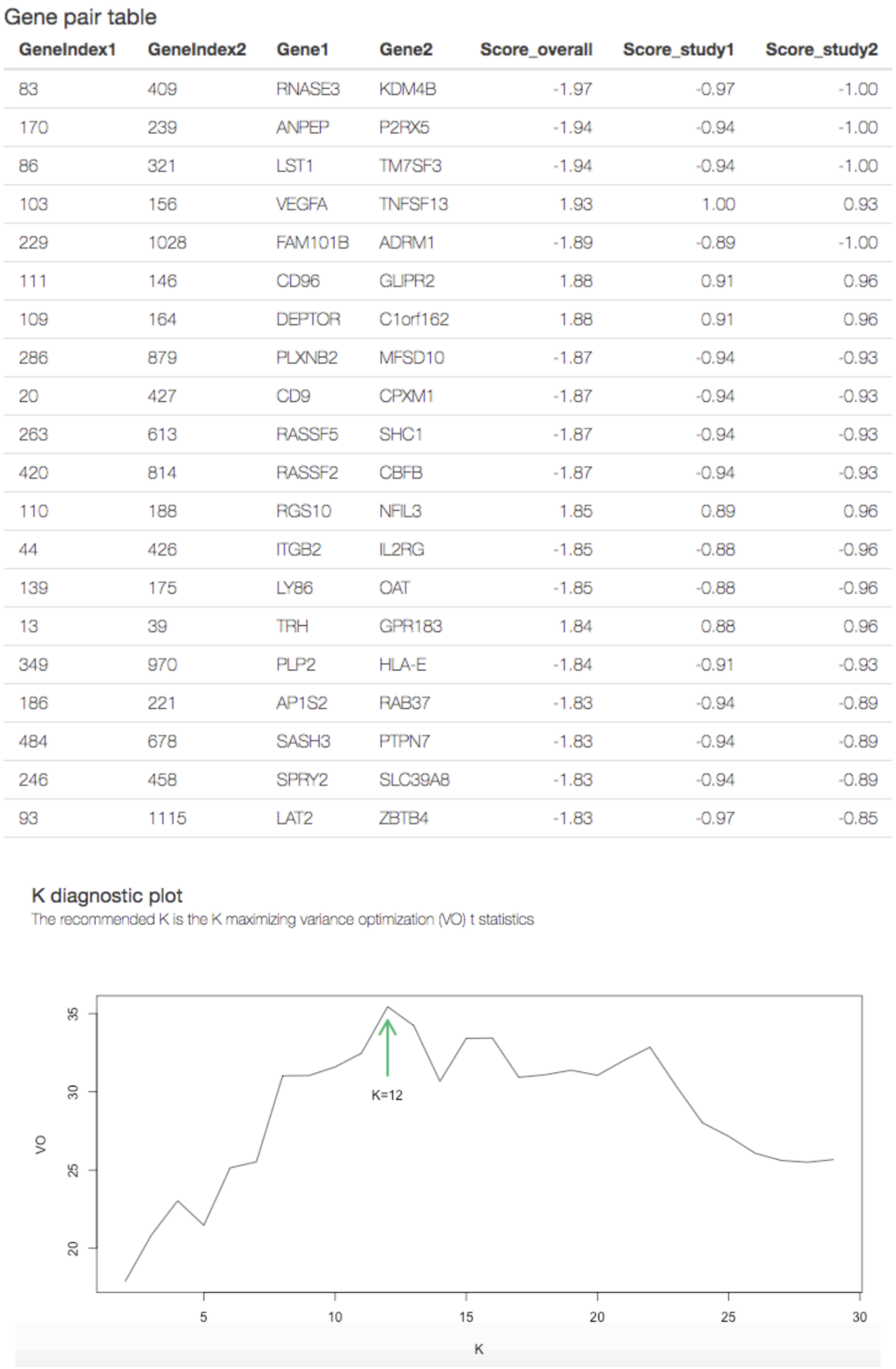
Please select TWO labels to cluster

Please select studies for training

Please select ONE study for testing

Number of top scoring pairs (K): 28

- *Step 1*: first specify (1), (2), (3) under the **Advanced Options** to decide a method to select K top scoring gene pairs from multiple studies, the maximum number of top scoring pairs K (algorithm will search from 1 up to K with default K = 29) and the number of cores for parallel computing. Then click on (4) to choose two labels and select the dataset as training data and testing on (5) and (6) respectively. Click the **Train model** button to build prediction model.
- *Step 2*: decide the K on (7) based on the diagnostic plot generated in Step 1 where the suggested value is shown as a green arrow. Then predict the class label of testing data. Finally, a confusion matrix is output to show the prediction results.



Gene pair table

GeneIndex1	GeneIndex2	Gene1	Gene2	Score_overall	Score_study1	Score_study2
83	409	RNASE3	KDM4B	-1.97	-0.97	-1.00
170	239	ANPEP	P2RX5	-1.94	-0.94	-1.00
86	321	LST1	TM7SF3	-1.94	-0.94	-1.00
103	156	VEGFA	TNFSF13	1.93	1.00	0.93
229	1028	FAM101B	ADRM1	-1.89	-0.89	-1.00
111	146	CD96	GLIPR2	1.88	0.91	0.96
109	164	DEPTOR	C1orf162	1.88	0.91	0.96
286	879	PLXNB2	MFSD10	-1.87	-0.94	-0.93
20	427	CD9	CPXM1	-1.87	-0.94	-0.93
263	613	RASSF5	SHC1	-1.87	-0.94	-0.93
420	814	RASSF2	CBFB	-1.87	-0.94	-0.93
110	188	RGS10	NFIL3	1.85	0.89	0.96
44	426	ITGB2	IL2RG	-1.85	-0.88	-0.96
139	175	LY86	OAT	-1.85	-0.88	-0.96
13	39	TRH	GPR183	1.84	0.88	0.96
349	970	PLP2	HLA-E	-1.84	-0.91	-0.93
186	221	AP1S2	RAB37	-1.83	-0.94	-0.89
484	678	SASH3	PTPN7	-1.83	-0.94	-0.89
246	458	SPRY2	SLC39A8	-1.83	-0.94	-0.89
93	1115	LAT2	ZBTB4	-1.83	-0.97	-0.85

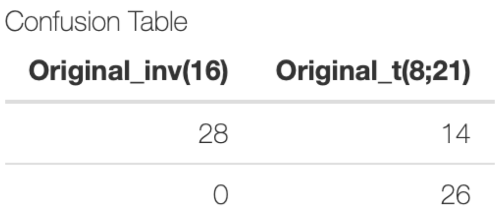
K diagnostic plot

The recommended K is the K maximizing variance optimization (VO) t statistics

Figure 10

Figure 10 shows the results of "**merge05"** data after *Step 1* that uses the **"study1.csv" and "study2.csv" for training and "study3.csv" for testing to predict between two labels "inv(16)" and "t(8;21)"**. The upper part lists the top predictive pairs of genes. A score measures the correlation between the pairs of genes. The bottom plot guides selection of K by maximizing variance optimizaiton (VO) t statistics. The x-axis is the number of top scoring pairs K, and the y-axis is the variance optimization (VO) t-statistics.

Based on the diagnostic plot showing the optimum K=12 in *Figure 10*, a confusion matrix is output after *Step 2*, showing the prediction results as below. The prediction results are also saved in the working directory.

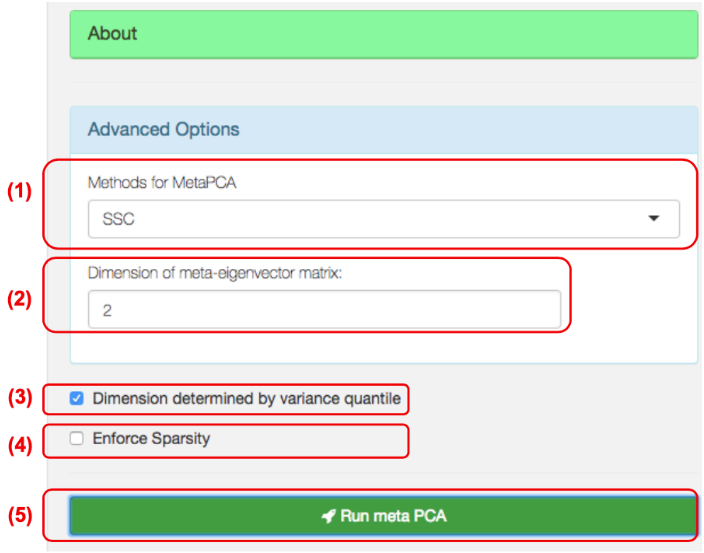


Confusion Table

Original_inv(16)	Original_t(8;21)
28	14
0	26

5.7 MetaPCA

MetaPCA module aims to combine multiple omics datasets and perform simultaneous dimensional reduction in all studies. The results show improved accuracy, robustness, and better interpretation among all studies. There are two methods: SSC represents MetaPCA via sum of squared cosine (SSC) maximization; SV represents MetaPCA via sum of variance decomposition (SV). By clicking the **Toolsets** tab and then choosing **MetaPCA**, we are directed to the **MetaPCA** homepage.



Methods for MetaPCA: SSC

Dimension of meta-eigenvector matrix: 2

☑ Dimension determined by variance quantile

☐ Enforce Sparsity

- Specify the method (SSC is suggested) on (1), and the dimension of the output meta-eigenvector matrix on (2).
- The checkbox of "**Dimension determined by variance quantile**" (3) is suggested to be checked so that the dimension size of each study's eigenvector matrix (SSC) is determined by the pre-defined level of variance quantile 80%. Checking checkbox of "**Enforce Sparsity**" (4) will need to first tune parameter for sparsity before **Run meta PCA**.

Figure 11 shows the **MetaPCA** result based on "merge05" data. The x-axis (horizontal) is the first principal component, and the y-axis (vertical) is the second principal component. Each dot represents a sample in a study with the sample label marked to the top right of the figure. The figures show nice separations between three groups. These figures and eigenvectors are saved to the MetaPCA folder.

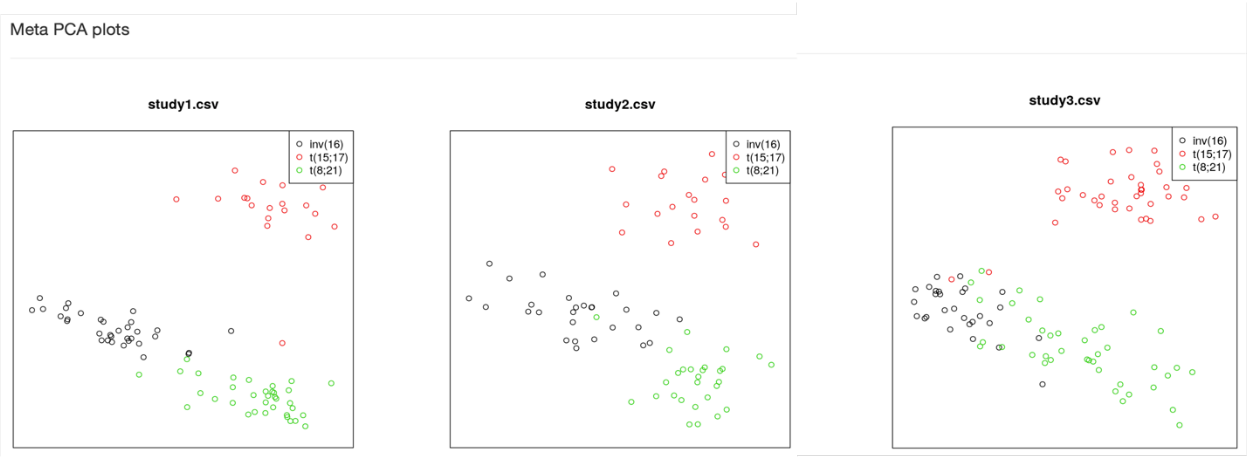

Meta PCA plots
study1.csv
inv(16)
t(15;17)
t(8;21)
study2.csv
inv(16)
t(15;17)
t(8;21)
study3.csv
inv(16)
t(15;17)
t(8;21)

Figure 11