A Tutorial for MetaOmics

Contents

1 Introduction

MetaOmics is a browser-based software suite for transcriptomic meta-analysis with R Shiny-based graphical user interface (GUI). Many state-of-the-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, and MetaPCA for principal component analysis. For detailed explanation of these omics data-analysis approaches, please refer to (to be updated once the manuscript is accepted).

In this tutorial, we will go through the installation and usage of MetaOmics step by step using real data examples. The MetaOmics software suite is publicly available at `https://github.com/metaOmics/metaOmics`. This tutorial can be found at `https://github.com/metaOmics/tutorial/blob/master/metaOmics_turtorial.pdf`. Each MetaOmics module will be introduced in later sections, and the associated R packages are available on GitHub at `https://github.com/metaOmics`.

1.1 Abbreviation terms

- General terms:
  - CV: Cross validation
  - DE: Differentially expressed
  - FDR: False discovery rate
  - FC: Fold change
  - FPKM: Fragments Per Kilobase Million mapped reads
  - QC: Quality control
  - RPKM: Read Per Kilobase Million mapped reads
  - TPM: Transcripts Per Kilobase Million mapped reads
  - CDF: Cumulative density function.
- Methods or tools:
  - AW-Fisher: Adaptively weighted Fisher's method
  - CPI: Comparative pathway integrator
  - FEM: Fixed effects model
  - KS test: Kolmogorov-Smirnov test
  - MAPE: Meta analysis pathway enrichment method
  - PCA: Principal component analysis
  - REM: Random effects model
  - SMR: Standardized mean ranks
  - TSP: Top scoring pair algorithm
  - KNN: K-nearest neighbor algorithm
  - MCMC: Markov Chain Monte Carlo

- PR: Pearson rank correlation
- SR: Spearman rank correlation
- rankProd: product of ranks from multiple studies
- SAM: R package for Significance Analysis of Microarrays
- LIMMA: R package for the analysis of gene expression data arising from microarray or RNA-Seq technologies
- DESeq2: R package for moderated estimation of fold change and dispersion for RNA-seq data
- edgeR: R package for Empirical Analysis of Digital Gene Expression Data
- voom: R package for precision weights unlock linear model analysis tools for RNA-seq read counts

2 Preliminaries

2.1 Citing MetaOmics

MetaOmics software suite implements many meta-analysis methods from different authors. Please cite appropriate papers if you use MetaOmics, by which the authors will receive professional credits for their work.

- MetaOmics software suite itself can be cited as:
  - Ma, T., Huo, Z., Kuo, A., Zhu, L., Fang, Z., Zeng, X., Lin, C.-W., Liu, S., Wang, L., Rahman, T., Chang, L.-C., Kim, S., Li, J., Park, Y., Song, C., Oesterreich, S., Sibille, E. and Tseng, G. C. MetaOmics: Comprehensive Analysis Pipeline and Browser-based Software Suite for Transcriptomic Meta-Analysis.
- Review, comparative papers and published R packages:
  - Tseng, G. C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research*, 40(9):3785–3799.
  - Chang, L.-C., Lin, H.-M., Sibille, E., and Tseng, G. C. (2013). Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC bioinformatics*, 14(1):368.
  - Wang, X., Kang, D. D., Shen, K., Song, C., Lu, S., Chang, L.-C., Liao, S. G., Huo, Z., Tang, S., Ding, Y., Kaminski, N., Sibille, E., Lin, Y., Li, J., and Tseng, G. C. (2012). An r package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. *Bioinformatics*, 28(19):2534–2536.
- MetaQC:
  - (MetaQC) Kang, D. D., Sibille, E., Kaminski, N., and Tseng, G. C. (2012). Metaqc: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. *Nucleic acids research*, 40(2):e15–e15.
- MetaDE:
  - (Fisher) Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.
  - (AW-Fisher) Li, J. and Tseng, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics*, 5(2A):994–1019.
  - (AW-Fisher) Huo, Z., Tang, S., Park, Y., and Tseng, G. (2017). P-value evaluation, variability index and biomarker categorization for adaptively weighted fisher's meta-analysis method in omics applications. *arXiv preprint arXiv:1708.05084*.
  - (REM/FEM) Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19(suppl 1):i84–i90.

  - (rOP) Song, C. and Tseng, G. C. (2014). Hypothesis setting and order statistic for robust genomic meta-analysis. *The annals of applied statistics*, 8(2):777.
  - (minMCC) Lu, S., Li, J., Song, C., Shen, K., and Tseng, G. C. (2009). Biomarker detection in the integration of multiple multi-class genomic studies. *Bioinformatics*, 26(3):333–340.
  - (Stouffer) Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M. (1949). The american soldier: adjustment during army life.(studies in social psychology in world war ii, vol. 1.)
  - (RankProd) Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., and Chory, J. (2006). Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–2827

- MetaPath:
  - (MAPE) Shen, K. and Tseng, G. C. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, 26(10):1316–1323.
  - (CPI) Fang, Z., Zeng, X., Lin, C.-W., Ma, T., and Tseng, G. C. (2017). Comparative pathway integrator: a framework of meta-analytic integration of multiple transcriptomic studies for consensual and differential pathway analysis. *In preparation.*

- MetaNetwork:
  - (MetaDCN) Zhu, L., Ding, Y., Chen, C.-Y., Wang, L., Huo, Z., Kim, S., Sotiriou, C., Oesterreich, S., and Tseng, G. C. (2016). Metadcn: meta-analysis framework for differential co-expression network detection with an application in breast cancer. *Bioinformatics*, 33(8):1121–1129.

- MetaPredict:
  - (MetaKTSP) Kim, S., Lin, C.-W., and Tseng, G. C. (2016). MetaKTSP: A Meta-Analytic Top Scoring Pair Method for Robust Cross-Study Validation of Omics Prediction Analysis. *Bioinformatics*, 32(13):1966–73.

- MetaClust:
  - (MetaSparseKmeans) Huo, Z., Ding, Y., Liu, S., Oesterreich, S., and Tseng, G. (2016). Meta-analytic framework for sparse k-means to identify disease subtypes in multiple transcriptomic studies. *Journal of the American Statistical Association*, 111(513):27–42.

- MetaPCA:
  - (MetaPCA) Kim, S., Kang, D., Huo, Z., Park, Y., and Tseng, G. C. (2017). Meta-analytic principal component analysis in integrative omics application. *Bioinformatics*, 1:8.

2.2 How to start MetaOmics

The full instruction of how to install and start MetaOmics software suite is also available at `https://github.com/metaOmics/metaOmics`. There are two ways to start the MetaOmics software: via R software or via docker.

2.2.1 Start from R

Requirement:

- R >= 3.3.1
- Shiny >= 0.13.2

Note:

- We recommend users to use R 3.3 to implement our tool. If you are using R 3.4 (released on 4/24/2017), you may encounter errors in installing dependencies of the modules. You can manually install the dependencies by running the following commands in R:

  install.packages(c('GSA', 'combinat', 'samr', 'survival', 'cluster', 'gplots', 'ggplot2', 'irr', 'shape', 'snow', 'snowfall', 'igraph', 'doMC', 'PMA'))

  source('https://bioconductor.org/biocLite.R')

  biocLite(c('multtest', 'Biobase', 'edgeR', 'DESeq2', 'impute', 'limma', 'AnnotationDbi', 'ConsensusClusterPlus', 'genefilter', 'GSEABase', 'Rgraphviz'))

- For Windows, users need to run the following command in R to install the package 'doMC':

  install.packages('doMC', repos='http://R-Forge.R-project.org')

How to install the software:

- At the MetaOmics home page `https://github.com/metaOmics/metaOmics`, clone the project by clicking on "Clone or download," and extract to a working directory, or type in the following in the command line:

  git clone `https://github.com/metaOmic/metaOmics`

How to start the software:

1. Open R.
2. Set the working directory such that the MetaOmics folder is included.

   install.packages('shiny')

   shiny::runApp('metaOmics', port=9987, launch.browser=T)

2.2.2 Run from docker image:

1. Install docker (`https://www.docker.com`).
2. In terminal:

   docker pull metaomics/app

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

2.3 MetaOmics setting page

After starting MetaOmics, the first page is the MetaOmics setting page as shown in Figure 1. There are four tabs on the top of this page (see Figure 1 (1)), which will direct users to specific functional modules of the software including "Setting," "Preprocessing," "Saved Data," and "Toolsets." Below these tabs is a **Welcome to MetaOmics** section, which briefly introduces the software and other information about the authors and maintainers. Further below, there are two sections: **Session Information** and **Directory for Saving Output Files** (see Figure 1 (2)). By clicking the "..." button, users can set the working directory, in which all the meta-analysis results will be saved. The current working directory is displayed on the top-right corner (see Figure 1 (3)). There is one more section with the header **Toolsets** (see Figure 1 (4)), where users can click the blue button to install the desired modules if the "Status" shows "not installed." If the packages are already installed, an icon of "checked installed" will show up in "Status." The installation progress may take up to a few minutes for each module. A notification icon will pop up at the bottom-right corner upon finishing the installation process. After the modules are installed, users need to restart the MetaOmics software suite so that the Shiny application interface is updated with the installed modules. The current active dataset is shown in Figure 1 (5), which is introduced in Section 4.1 **Step 2**.

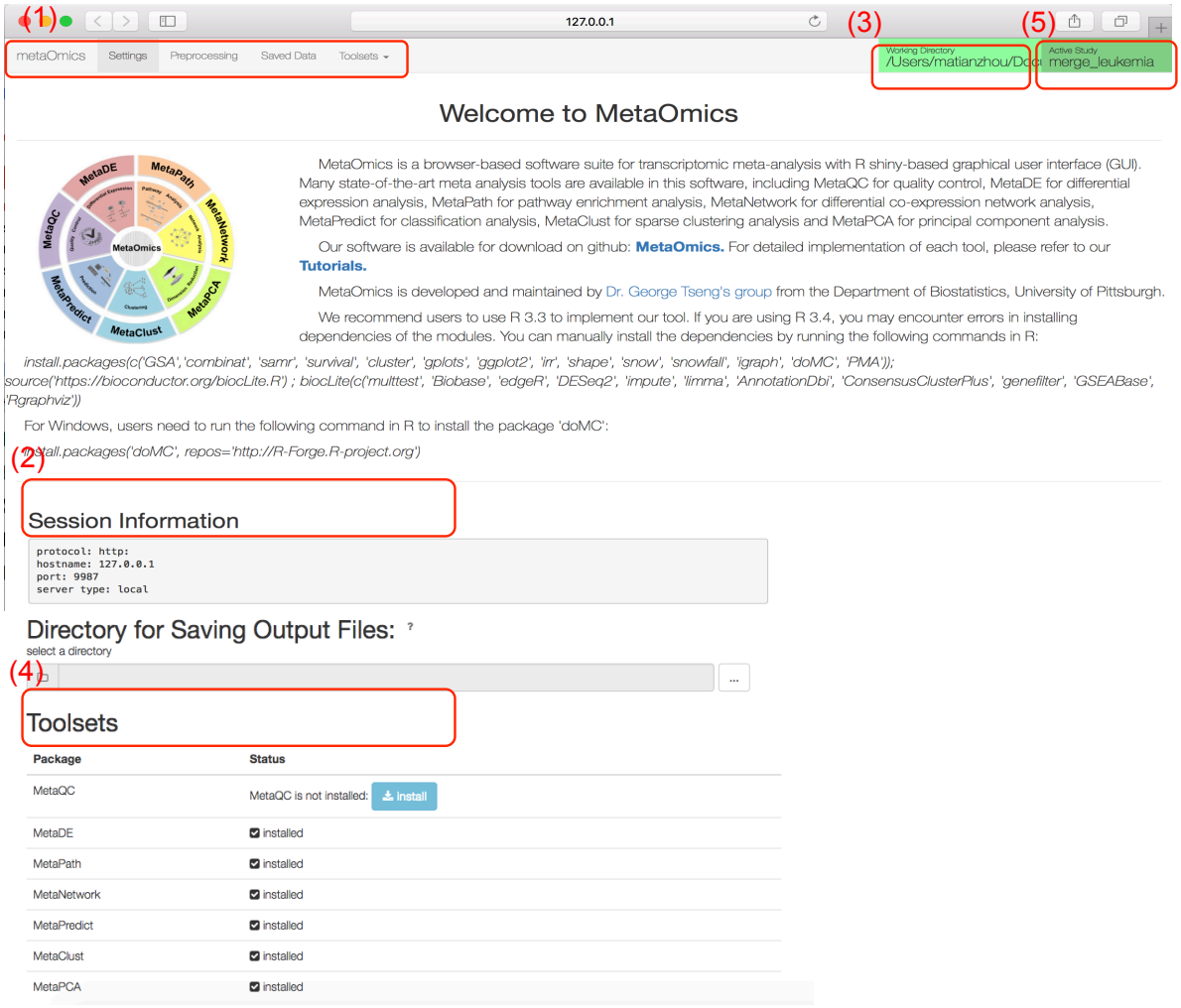


Figure 1: MetaOmics software suite GUI setting page

2.4 Question and bug report

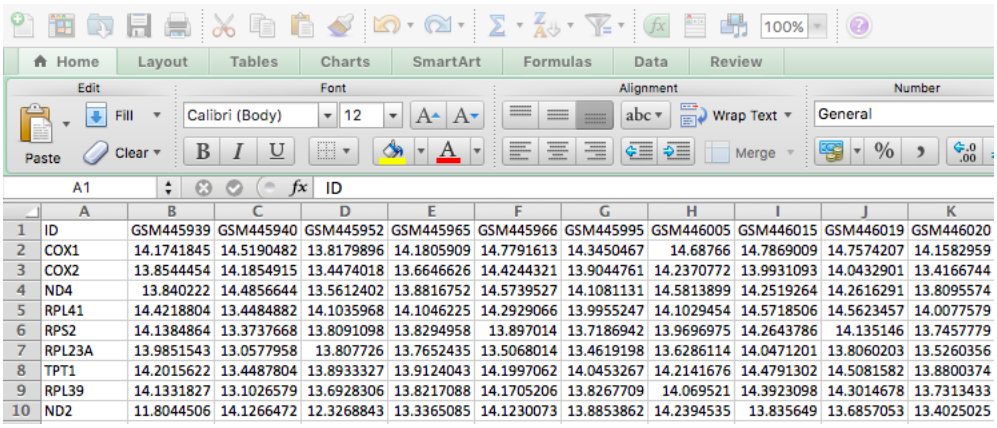
If you encounter errors or bugs, please report to the maintainer, Tianzhou Ma <tim28@pitt.edu>.

3 Prepare data

In this section, we will introduce how to prepare a gene-expression data matrix as well as the clinical-data files for the MetaOmics software suite.

3.1 Raw data

The gene-expression matrix should be prepared as tab-delimited ".txt" or comma-separated ".csv" files (see an example in Figure 2). 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. Valid data types include microarray data, RNA-seq FPKM/RPKM data, and RNA-seq count data. Abbreviation terms for FPKM and RPKM are explained in Section 1.1.

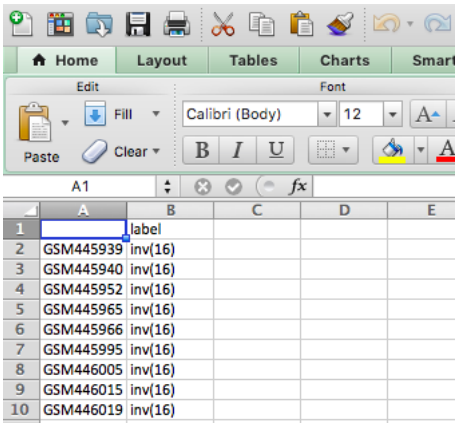


	A	B	C	D	E	F	G	H	I	J	K
1	ID	GSM445939	GSM445940	GSM445952	GSM445965	GSM445966	GSM445995	GSM446005	GSM446015	GSM446019	GSM446020
2	COX1	14.1741845	14.5190482	13.8179896	14.1805909	14.7791613	14.3450467	14.68766	14.7869009	14.7574207	14.1582959
3	COX2	13.8544454	14.1854915	13.4474018	13.6646626	14.4244321	13.9044761	14.2370772	13.9931093	14.0432901	13.4166744
4	ND4	13.840222	14.4856644	13.5612402	13.8816752	14.5739527	14.1081131	14.5813899	14.2519264	14.2616291	13.8095574
5	RPL41	14.4218804	13.4484882	14.1035968	14.1046225	14.2929066	13.9955247	14.1029454	14.5718506	14.5623457	14.0077579
6	RPS2	14.1384864	13.3737668	13.8091098	13.8294958	13.897014	13.7186942	13.9696975	14.2643786	14.135146	13.7457779
7	RPL23A	13.9851543	13.0577958	13.807726	13.7652435	13.5068014	13.4619198	13.6286114	14.0471201	13.8060203	13.5260356
8	TPT1	14.2015622	13.4487804	13.8933327	13.9124043	14.1997062	14.0453267	14.2141676	14.4791302	14.5081582	13.8800374
9	RPL39	14.1331827	13.1026579	13.6928306	13.8217088	14.1705206	13.8267709	14.069521	14.3923098	14.3014678	13.7313433
10	ND2	11.8044506	14.1266472	12.3268843	13.3365085	14.1230073	13.8853862	14.2394535	13.835649	13.6857053	13.4025025

Figure 2: An example input data format

3.2 Clinical data

Clinical data should be prepared as in the example in Figure 3. The first column 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.



	A	B	C	D	E
1		label			
2	GSM445939	inv(16)			
3	GSM445940	inv(16)			
4	GSM445952	inv(16)			
5	GSM445965	inv(16)			
6	GSM445966	inv(16)			
7	GSM445995	inv(16)			
8	GSM446005	inv(16)			
9	GSM446015	inv(16)			
10	GSM446019	inv(16)			

Figure 3: A example clinical data format

3.3 Example data in the MetaOmics software

We collected three multi-cohort example datasets for the MetaOmics software, including an acute myeloid leukemia (AML) example, a breast cancer (BRCA) example, and a prostate cancer example. Table 1 summarizes the first example dataset of three studies on acute myeloid leukemia (AML). Table 2 summarizes the second example dataset of four studies on breast cancer, whereas the first study contains both count data and FPKM data (continuous). Table 3 summarizes the third example dataset of eight studies on prostate cancer. The leukemia data is used to demonstrate MetaPreprocess and all seven analytical modules (MetaQC, MetaDE, MetaPath, MetaNetwork, MetaPredict, MetaClust, and MetaPCA). Considering that all three studies in leukemia data are of high quality, we further run MetaQC on the prostate cancer data with eight studies to show its capability to identify low-quality studies.

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

Table 2: Multi-study breast cancer gene-expression profiles. All gene-expression profiles of all four studies contain 10,330 genes. Study 1 contains both count data and FPKM (continuous) data, so user should **select only one of them**. The other three studies contain only continuous data. The phenotype of interest is the estrogen-receptor (comparing ER+ vs ER-).

Study	Source	Scale	# Samples	# Samples by ER ER+/ER-
Study 1	Weinstein et al. (2013)	count continuous	406	319/87
Study 2	Desmedt et al. (2007)	continuous	198	134/64
Study 3	Wang et al. (2005)	continuous	286	209/77
Study 4	Ivshina et al. (2006)	continuous	245	211/34

Table 3: Multi-study prostate cancer dataset information. Eight prostate cancer gene-expression profiles were measured by different microarray platforms.

Study	Source	# Platform	# Samples	# Samples by label Normal/Primary	# Genes
Study 1	Welsh et al. (2001)	HG-U95A	34	9/25	8798
Study 2	Yu et al. (2004)	HG-U95Av2	146	81/65	8799
Study 3	Lapointe et al. (2004)	cDNA	103	41/62	13579
Study 4	Varambally et al. (2005)	HG-U133 Plus 2	13	6/7	19738
Study 5	Singh et al. (2002)	HG-U95Av2	102	50/52	8799
Study 6	Wallace et al. (2008)	HG-U133A2	89	20/69	12689
Study 7	Nanni et al. (2006)	HG-U133A	30	7/23	12689
Study 8	Tomlins et al. (2006)	cDNA	57	27/30	9703

4 MetaPreprocess

In this section, we introduce how to upload the datasets into the MetaOmics software suite, which is the first step in order to run any of the seven analytical modules. The R package for the MetaPreprocess module can be found at `https://github.com/metaOmics/preproc`.

4.1 Preprocessing page

Step 1 Uploading data:

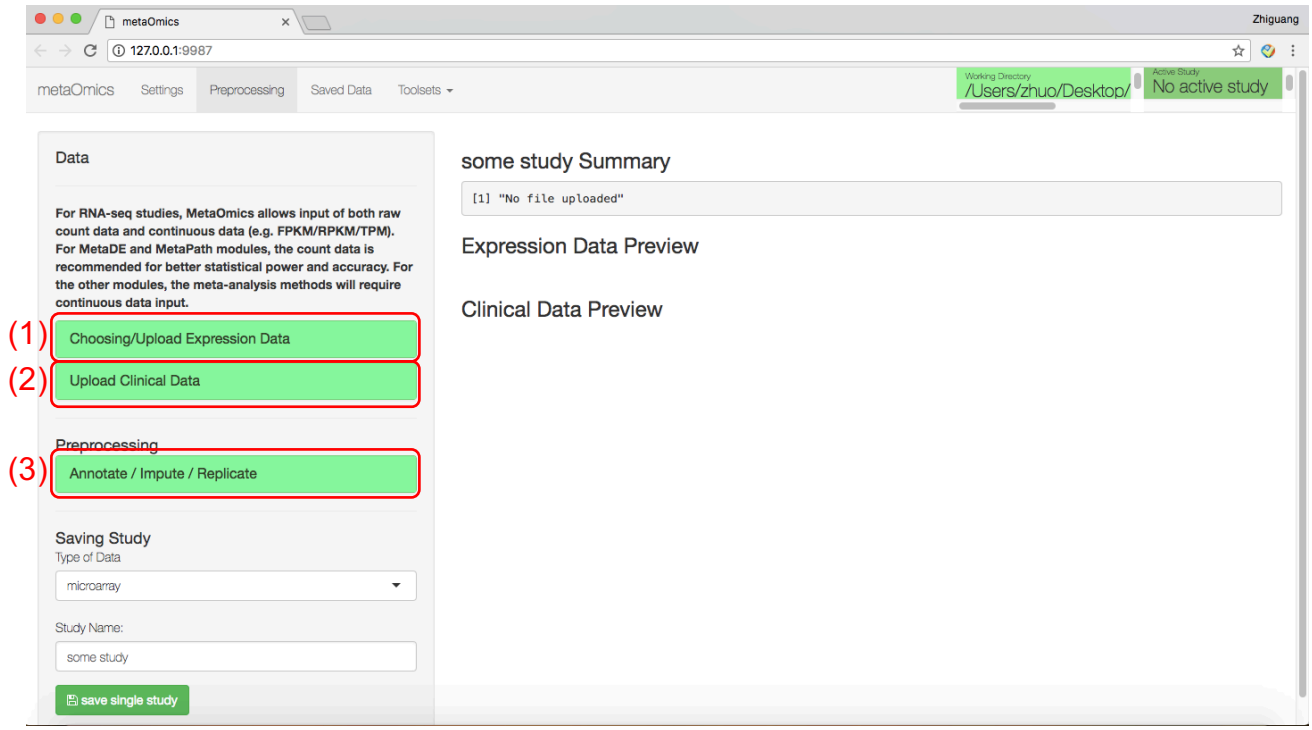


Figure 4: GUI for the preprocessing page

On the preprocessing page, by clicking on the "Choosing/Upload Expression Data" tab, users can upload expression data files (i.e., a data file from each study) or choose the existing saved data files as in Figure 4 (1). The data should be prepared according to Section 3. Users may optionally upload clinical data (see Figure 4 (2)), depending on their biological purposes. Note that all MetaOmics modules require external clinical labels except for the MetaClust module. The three example datasets are available under MetaOmics folder "/metaOmics/data/example/", and we will mainly focus on the AML dataset ("/metaOmics/data/example/leukemia") throughout this tutorial. 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.

Step 2 Preprocessing:

The MetaOmics software suite also provides handlers (see Figure 4 (3)) for feature annotation, missing value imputation, and multiple probes reduction for the same gene. For preprocessing, click on the "Annotate/Impute/Replicate" button to:

1. Annotate the probe ID/reference sequence ID/Entrez ID of an individual dataset (choose the gene symbol if the input data rows are already annotated).
2. Impute the missing value using the K-Nearest Neighbors (KNN) algorithm. The algorithm has been described in Troyanskaya et al. (2001). Note that the missing value will be automatically detected. If there is no missing value, this function will be disabled.

3. Handle the multiple probes matching the same gene issue. For the meta-analysis purpose, if multiple probes match the same gene symbol, only one represented gene will be selected. If each input feature represents a unique gene symbol, this function will be disabled.

A complete introduction of these options is available in Section 6. The right side of Figure 5 will show the summary statistics of uploaded data and preview of the data matrix. There is a search box where users can search for their genes of interest.

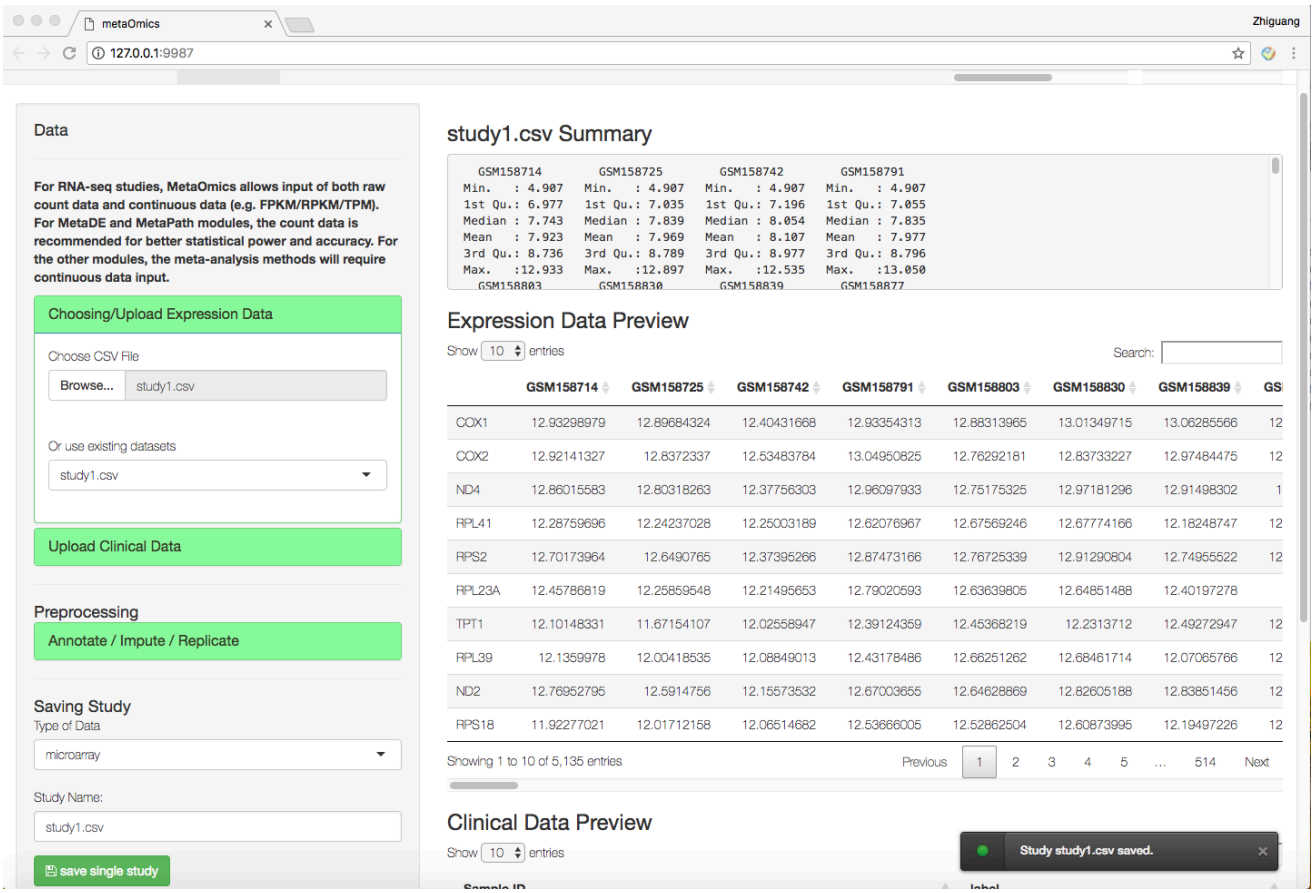


Figure 5: Uploading individual studies

Step 3 Save single study: In the next step, specify the data type (microarray data; RNA-seq FPKM/RPKM data; RNA-seq count data) and study name, and click "save single study." To upload RNA-seq data, the count data and FPKM/RPKM data should be uploaded separately and saved using different names. Abbreviation terms for FPKM and RPKM are explained in Section 1.1.

Step 4 Upload datasets for all studies: Repeat the steps above for all studies in the meta-analysis. All uploaded studies are now available on the "Saved Data" page.

Download GEO datasets (optional):

In the newest revision of our software (2018/09), we have added the "Download GEO dataset (optional)" feature to our preprocessing page to give users convenient options to directly download datasets from GEO via GEOquery. Users can type in the GSE ID they are looking for and click on the green button "Download GEO data" to download the data (Figure 6). The data will be saved under the folder "GEO" in the working directory with file names in the form of "GSExxx_expr.csv" (corresponding to the expression data) and "GSExxx_pheno.csv" (corresponding to the phenotype data).

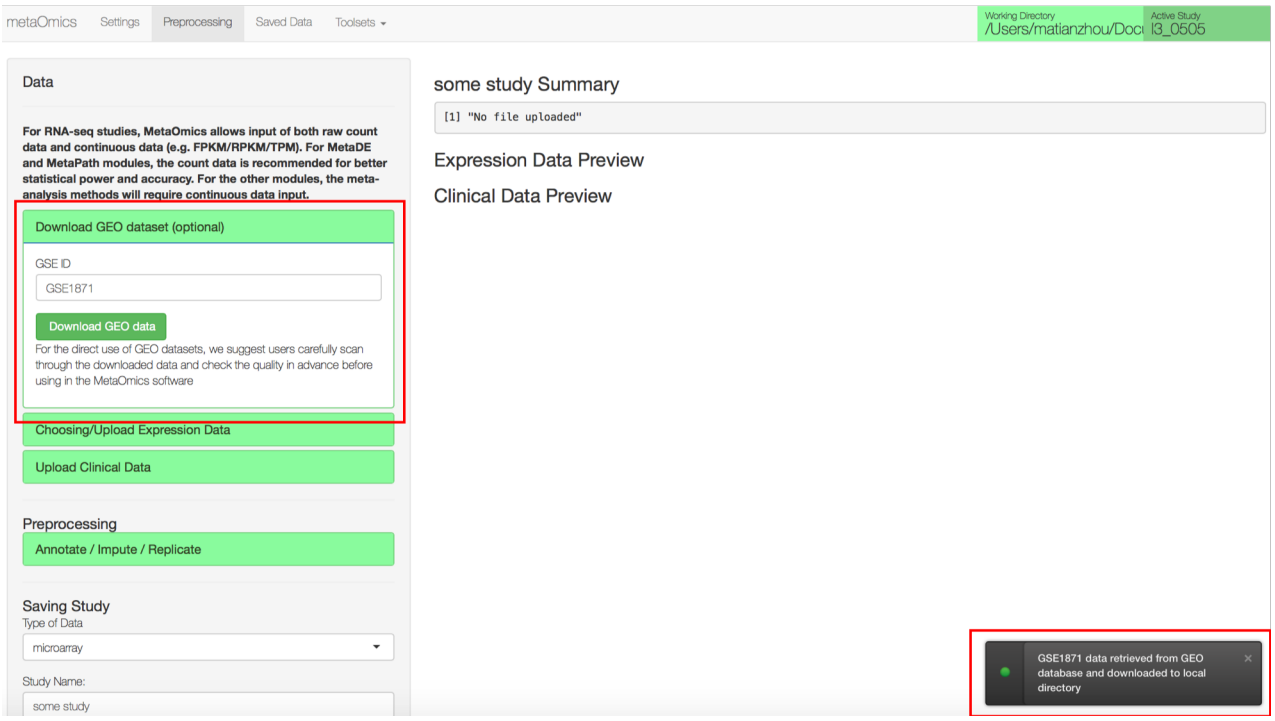


Figure 6: Download GEO dataset

Once retrieved, users can either choose to upload the csv file of the GEO dataset (either GSExxx_expr.csv or the file after processing by users), or directly use the downloaded GEO dataset (choose "Yes") with log transforming option. On the right side of Figure 7, we will see the preview of both expression data and clinical data (by default, the clinical data will be the phenotype data downloaded). Users can then do the standard preprocessing as above and save the study for the next steps.

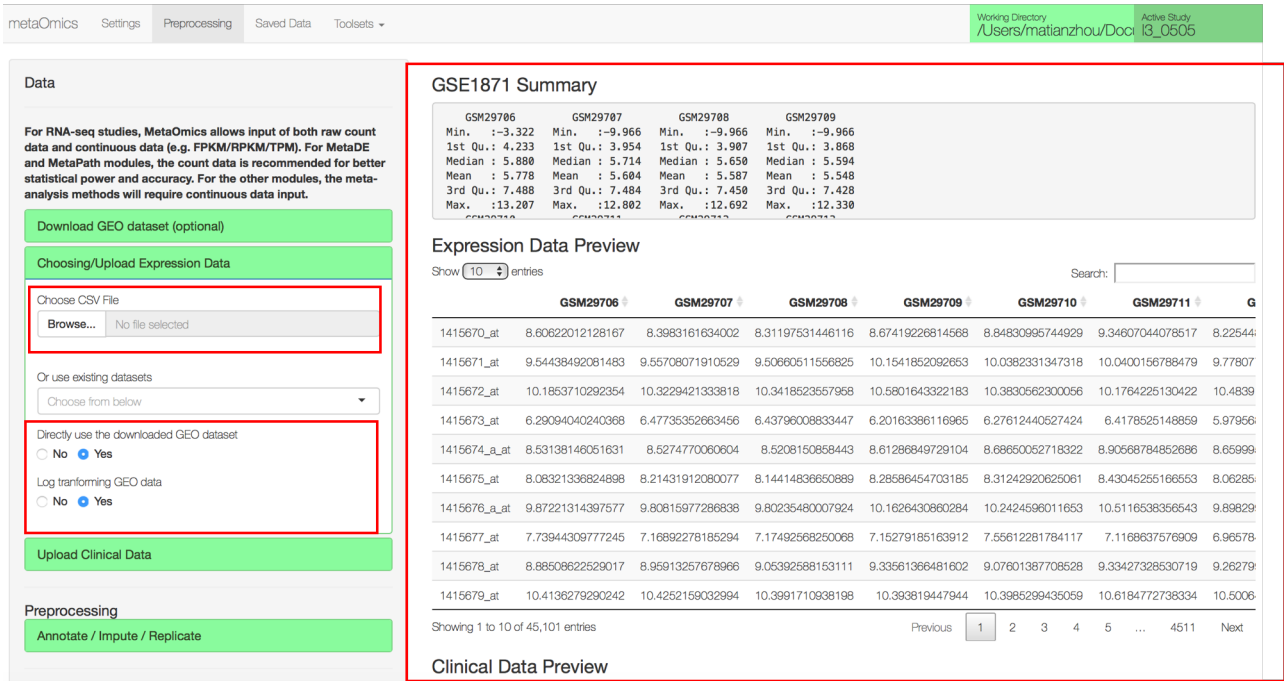


Figure 7: Download GEO dataset

4.2 Saved Data page

After uploading multiple studies with/without clinical data, users can turn to the Saved Data page.

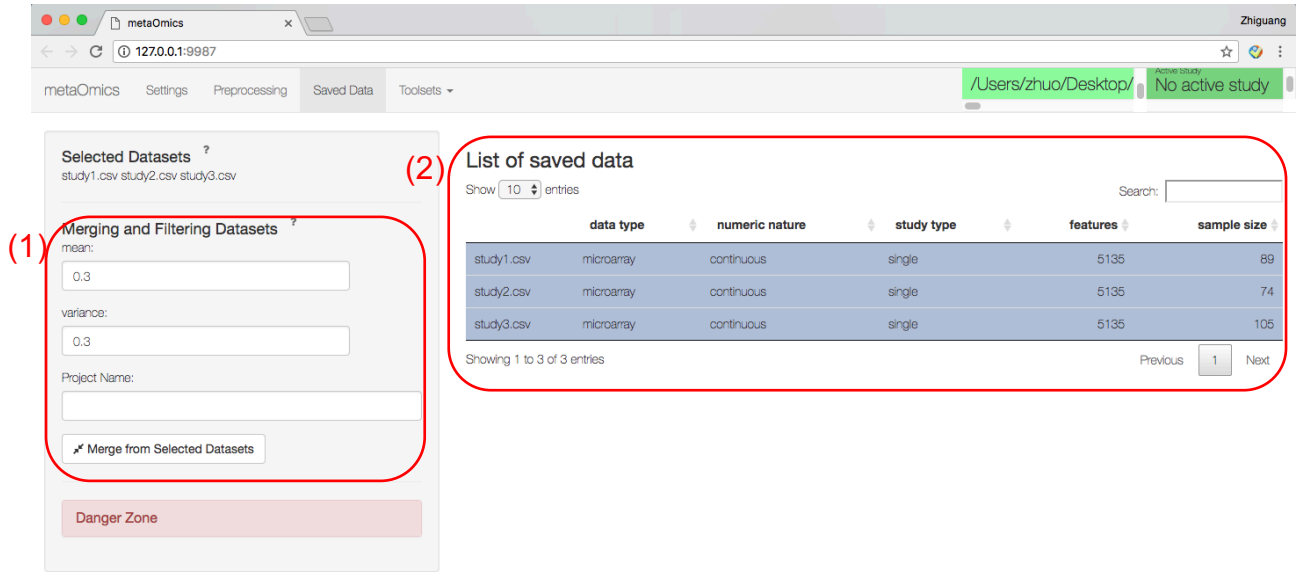


Figure 8: Merge from selected datasets

Step 1 Merging and filtering: All saved datasets from the previous step will be found in Figure 8 (2). Users should select multiple datasets for further meta-analysis. Users can filter out genes with low expression level (by default, mean expression lower than 30^{th} percentile) or low variance (by default, variance lower

than 30^{th} percentile). After specifying filtering criteria, enter "Project Name" and click on the "Merge from Selected Datasets" (see Figure 8 (1)). A merged dataset (study type = "multiple") will appear on the "List of Saved Data" panel (Figure 8 (2)). Creating multiple projects with varying preprocessing criteria is useful. For example, the user can start from a project with harsh filtering criteria (maintain 500-1000 genes) and give a test run through all modules to save time. If successful, a larger project can be created by including more genes. If users want to delete any dataset, they can click on the red "danger zone" button and delete the selected datasets.

Step 2 **Make active dataset:** The last thing to do before using any of the meta-analytic modules is to select the merged data and click on the "Make Your Dataset Active Dataset" button – a big green button in Figure 9. Then the merged data will become 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 in MetaOmics.

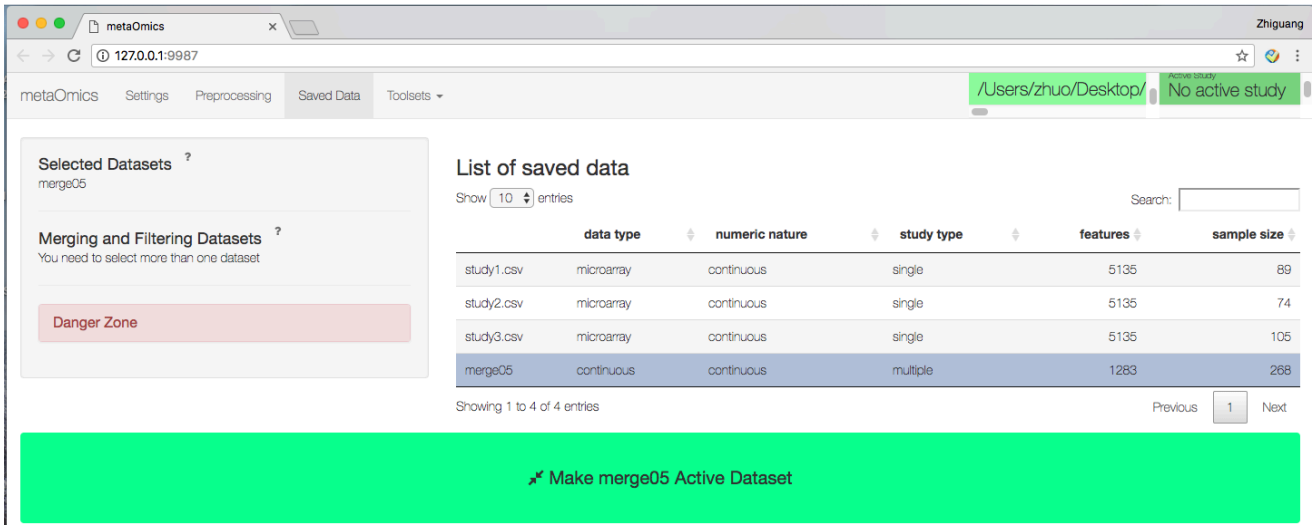


Figure 9: Make merged dataset active

5 Toolsets

After the MetaPreprocess module, the merged multiple studies are ready for the seven analytical modules in MetaOmics. By clicking on the "Toolsets" tab, users can navigate to individual MetaOmics modules. In the next few subsections, we will introduce in detail how to run each of these modules. For each MetaOmics module, a summary table of studies and sample sizes is shown on the top-left corner. There is an "about" drop-down menu that contains a brief introduction of the module. The "options" drop-down menu contains common options users can select or tune in the analysis. The "advanced options" menu contains more technical options, which we generally do not recommend users change unless they are familiar with the methods. After applying the modules, all result files will be automatically saved in the working directory that is specified in Section 2.3. For computationally demanding methods, the procedure may take minutes or hours, depending on the size of datasets. Users can keep track of the progress by checking the R console.

5.1 MetaQC

The MetaQC package 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: IQC, EQC, AQCg, CQCg, AQCp, and CQCp. Details of how each measure is defined and computed can be found in Kang et al. (2012). 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. The MetaQC package itself can be downloaded at (`https://github.com/metaOmics/MetaQC`).

5.1.1 Procedure

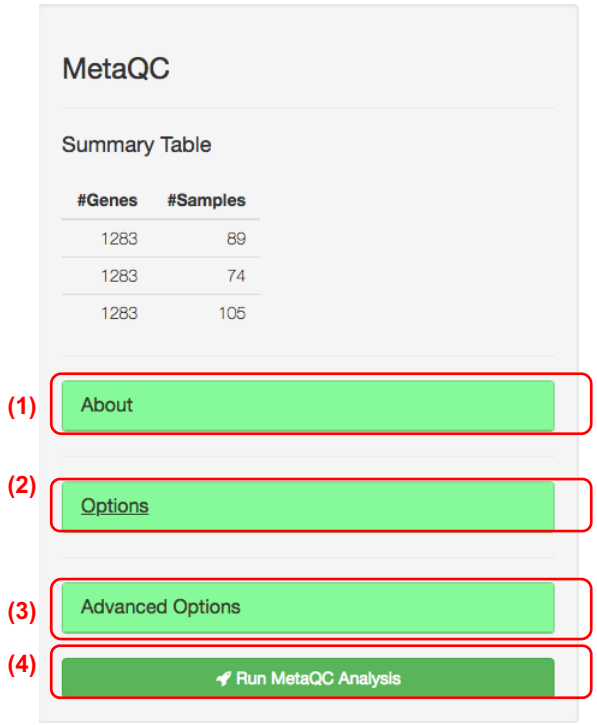


Figure 10: MetaQC options

There are three main options available for the MetaQC package, as shown in Figure 10. The complete list of MetaQC options is available in Section 6.1.

Step 1 Options: Under the drop-down menu "Options" (Figure 10 (2)), users can:

- Perform gene filtering. Gene filtering is suggested to reduce computational cost. Once "Yes" is chosen for gene filtering, users need to further specify the filtering cutoffs by mean and variance. In the demo example, the merged data have already had gene filtering in the merging step, so no further filtering is performed.
- Specify the approach (either by raw p-value or adjusted p-value) and cutoff to select potentially DE genes needed in the computation of IQC, EQC, AQCg, and CQCg.
- Specify the approach (either by raw p-value or adjusted p-value) and cutoff to select potentially enriched pathways needed in the computation of AQCp and CQCp.

Note that the larger the p-value or adjusted p-value is, the more genes will be included in the quality control computing. The default parameters are suggested for the purpose of MetaQC.

Step 2 Advanced options:

Under the drop-down menu "Advanced Options" ((3) in Figure 10), users are allowed to tune other parameters of MetaQC. In particular, it includes the selection of pathways by pathway size and the number of permutations to run to obtain the six measures. A complete list of all options available for the package can be found in Section 6.1. However, these are advanced methods and users are suggested not to modify the option setting in this section without knowing the method.

Step 3 Run MetaQC Analysis:

Once all the above options are specified, users can click on "Run MetaQC Analysis" (Figure 10 (4)) to perform MetaQC.

5.1.2 Results for MetaQC on AML example

(1)

Show 10 entries Search:

	IQC	EQC	AQCg	AQCp	CQCg	CQCp	SMR
study1.csv	4.002	4.214	46.13	47.72	190.1	0.0000506	1.667
study2.csv	410	409.6	36.49	39.76	111.1	0	2.25
study3.csv	2.673	2.287	47.82	59.08	148.6	0	2.083

Showing 1 to 3 of 3 entries Previous 1 Next

Download Csv File

(2)

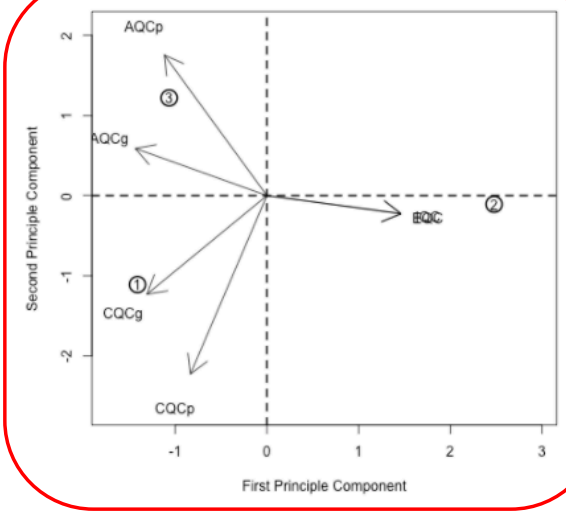


Figure 11: MetaQC Results for AML example. Figure 11 (1) includes seven columns, with the first six columns corresponding to the six quantitative quality control measures of all studies (a larger value indicates a better quality), and the seventh column is the rank of summary statistics of all six quality measures (a lower rank indicates a better quality). Users can download the full table as a .csv file by clicking on "Download .csv File". In addition to the tabular results, MetaQC also generated a PCA biplot 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. In this specific example, since the six QC directions are quite heterogeneous, we cannot exclude any study because of quality issues.

The performance of MetaQC in the AML example is shown in Figure 11. Detailed descriptions of these studies can be found in Table 1. As shown in Figure 11, the MetaQC module generates a summary table of MetaQC results (Figure 11 (1)) as well as a PCA biplot (Figure 11 (2)). In Figure 11 (1), if a study has larger standard mean rank (SMR) values, it is considered to have lower quality and is inconsistent with other studies. In Figure 11 (2), if a study (a circled number) is along the direction of majorities of the six QC directions, then the study has higher quality and is consistent with other studies. Since the AML example has very good data quality, we won't be able to demonstrate how to use MetaQC to identify studies of low quality. Thus, in order to demonstrate how to give suggestions on excluding studies with low quality or incompatibility with other studies, we further performed MetaQC on the prostate cancer example.

5.1.3 Results for MetaQC on prostate cancer example

Figure 12: MetaQC Results for prostate cancer example. Figure 12 (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 the six quality measures (a lower rank indicates a better quality). Users can download the full table as a .csv file by clicking on "Download .csv File". In addition to the tabular results, MetaQC also generated a PCA biplot 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 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. In this specific example, the six QC directions all direct to the left, and study 8 and study 7 are discordant with this direction. As a consequence, study 8 and study 7 are inconsistent with other studies and should be further scrutinized for potential technical or biological causes of their lower quality.

The test data used to demo the MetaQC package here is merged from eight prostate cancer studies with 30% of genes filtered out by mean and 30% filtered out by variance (2077 genes remained for the MetaQC analysis). Detailed descriptions of these studies can be found in Table 3. As shown in Figure 12, the MetaQC module generates a summary table (Figure 12 (1)) of MetaQC results as well as a PCA biplot (Figure 12 (2)). In this specific example, the 7^{th} and the 8^{th} studies have larger SMR values than the other studies. In addition, in the biplot, the 7^{th} and the 8^{th} studies are inconsistent with the six QC directions. Therefore, both the 7^{th} and the 8^{th} studies have relatively poorer quality, and they are inconsistent with other studies. Users should pay attention to these potential outlier studies.

5.2 MetaDE

The MetaDE package 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. The R package for the MetaDE module can be found at `https://github.com/metaOmics/MetaDE`. 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. In the two subsections below, we will go over how to perform (1) differential expression analysis and (2) pathway enrichment analysis based on (1) differential expression analysis. Note that MetaDE involves many sophisticated meta-analysis methods; please refer to the MetaOmics paper (supplementary information: Box S3) for a detailed description of each method.

5.2.1 Procedure of differential expression analysis

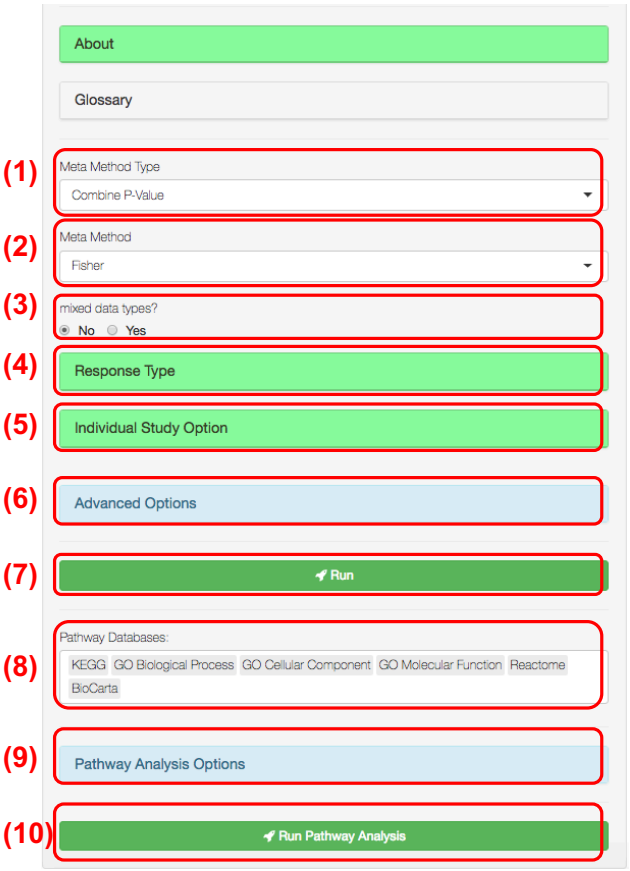


Figure 13: MetaDE options

In Figure 13, the red boxes (1) - (7) are the options and steps to run the differential expression analysis. A complete list of all options is available in Section 6.2.

Step 1 Choose the type of meta-analysis method: There are three types of meta-analysis to choose from: combining p-values, combining effect sizes and others. Note that there are many abbreviation terms. Users could refer to Section 1.1 for their full names.

Step 2 Choose a meta-analysis method:

- For the "combining p-values" category, users can choose from "Fisher", "AW-Fisher*", "maxP", "minP", "rOP*" and "Stouffer", where some of them have the one-sided, corrected versions.
- For the "combining effect size" category, users can choose from "FEM" and "REM*", where "REM*" has choice of six analytical algorithms for implementation.
- For "others", there are three rank-based methods (PR, SR, and rankProd) and minMCC for multi-class meta-analysis. To choose from the overwhelming number of meta-analysis methods, we follow Chang et al. (2013) and mark * for the top performing methods AW-Fisher, REM (HO option), and rOP as recommendations for users.

Step 3 Mixed data type: If this option is selected, MetaDE will allow partial studies with count data from RNA-seq and remaining studies with continuous intensities from microarray.

Step 4 Choose the response type: Under the drop-down menu, users can specify the types of outcome (response) variables to be two-class, continuous, multi-class, or survival. By choosing "two class comparison," users can specify the group label name for the Label Attribute (from the column names of your clinical data). Then for group label (a factor of at least two levels), specify the name for the Control Label and Experimental Label, respectively. For the other types, only the group label name is needed.

Step 5 Choose study design for individual study:

- Individual data types can be either discrete (count) or continuous.
- Under the drop-down menu "Setting Individual Study Method," user can specify individual study methods according to individual data type. For continuous data (e.g., microarray), available options include LIMMA (default method) and SAM. For discrete data (e.g., RNA-seq count), available options include edgeR, DESeq2, and Voom.
- The users can also specify whether each study is paired design or not.

Step 6 Advanced Options If selected, other uncommonly used options will become available. Again, this is not suggested if you are not familiar with the method.

- Parametric: If "no" is selected, permutation will be performed instead of the parametric closed form solution.
- Covariate: Indicate if any covariate will be adjusted.
- Alternative hypothesis: Two-sided or one-sided.

Step 7 Run: Once all the above options are specified, users can click on the "Run" button to perform MetaDE analysis.

5.2.2 Results of differential expression analysis

We used the AML example to demonstrate the MetaDE module. After merging the three datasets and filtering out 50% of genes by mean and 50% of genes by variance, 1283 genes remained. In this example we only compared two phenotypes: inv(16) and t(15;17). Detailed descriptions of these studies can be found in Table 1. Two main outputs from the first "meta differential analysis" step in the procedure are shown in Figure 14.

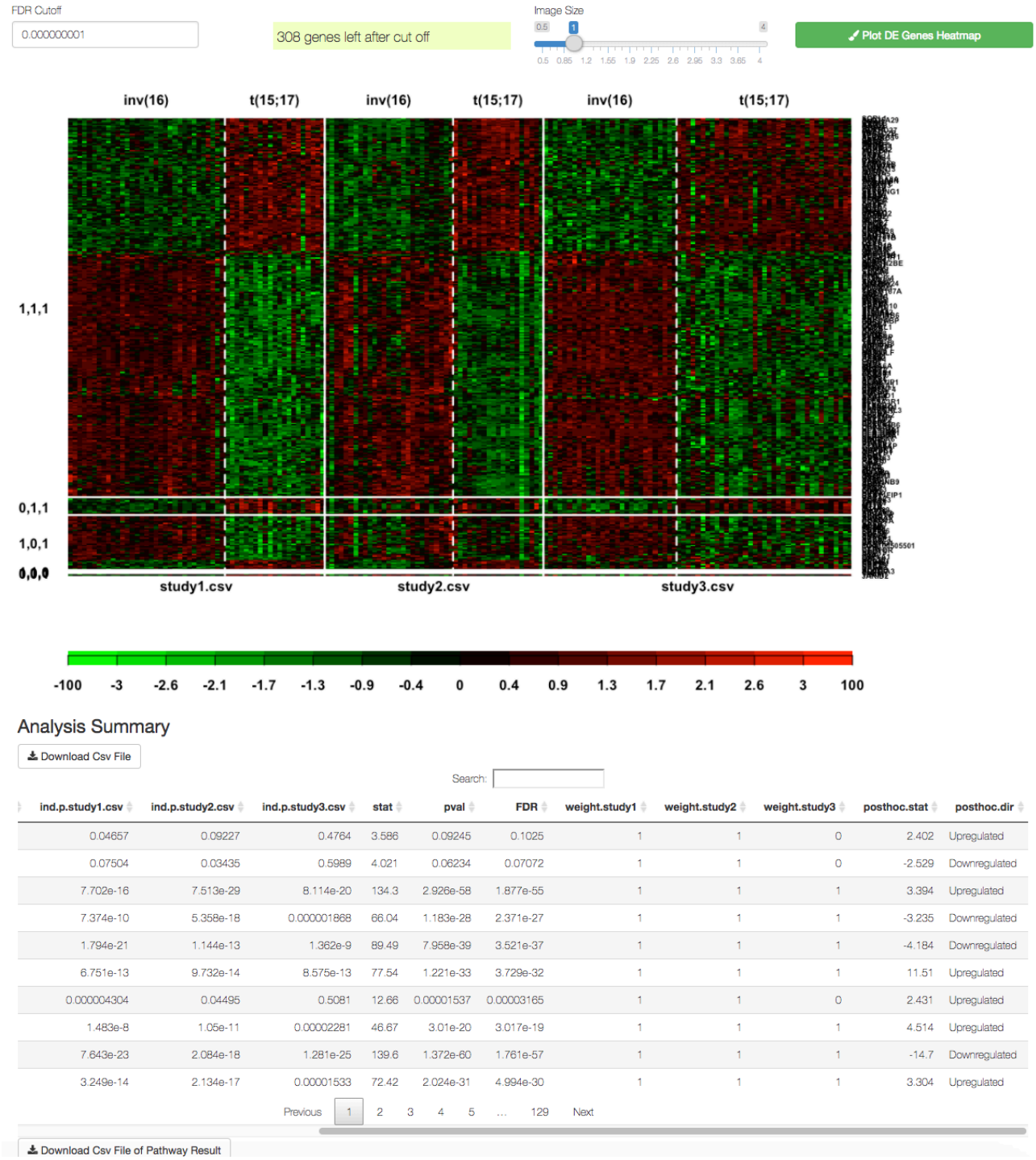


Figure 14: MetaDE Results. The heatmap of DE genes is rendered after specifying the FDR cutoff for selection of DE genes and clicking on "Plot DE Genes Heatmap." The image size can be adjusted by dragging the scrolling bar. In the heatmap, rows refer to the declared DE genes under the specified FDR cutoff, 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. The summary of meta-analysis results is on the bottom, including information of individual test statistics, individual study p-value, meta-analysis p-value, FDR, etc.

5.2.3 Procedure of downstream pathway analysis

Users can then perform pathway enrichment analysis on the declared DE genes from the previous step, the options and steps of which are shown in red boxes (8) - (10) in Figure 13. The procedure is outlined as below.

Step 1 **Choose the pathway database:** Users can select from 25 available pathway databases to perform the pathway enrichment analysis.

Step 2 **Choose the pathway enrichment method and the pathway size range:** In this step users can choose pathway enrichment options with the Kolmogorov-Smirnov (KS) test (default option), or the Fisher's exact test by specifying number of input genes for pathway analysis. Both of these tests could provide the significance measurement on DE gene enrichment for the prespecified pathways. For Fisher's exact test, the input genes can be obtained by either specifying a MetaDE p-value cutoff or specifying the number of top DE genes. The KS-test doesn't require a specific p-value cutoff as it will use all MetaDE p-value information to test the enrichment. Users can also specify the minimum/maximum gene size of pathways to be included for pathway enrichment analysis.

Step 3 **Run:** Once all the above options are specified, users can click on the "Run Pathway Analysis" button to perform pathway enrichment analysis.

5.2.4 Result of downstream pathway analysis

The result of downstream pathway analysis is shown in Figure 15. The summary includes the pathway names, the corresponding enrichment p-value and FDR. In addition to the results shown in the browser, users can download the result by clicking on the "Download .csv File" button on the top left of the summary table.

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 Previous 1 2 3 4 5 ... 191 Next

Figure 15: Downstream pathway analysis based on MetaDE genes. Each row represents a pathway with its p-value and q-value listed on the right.

5.3 MetaPath

In the previous MetaDE package, pathway analysis is performed using the declared DE genes from MetaDE analysis. 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). Instead of directly using the declared DE genes from MetaDE analysis as input, the MAPE algorithm performs pathway analysis in each study and performs meta-analysis on the pathway level. In addition, CPI also includes advanced pathway clustering diagnostics and pathway clustering with text mining to circumvent abundant pathway redundancy in the databases and improve interpretation. The R package for MetaPath module can be found at `https://github.com/metaOmics/MetaPath`.

Note that both MetaDE and MetaPath could perform pathway enrichment analysis. Below are their differences. 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. For example, MAPE_G is more similar to MetaDE in a sense since both of them combine gene level p-values first and then perform pathway analysis, while MAPE_P is different since it performs single-study DE and pathway analysis first and then combines pathway level p-values. MAPE_I takes the advantage of both, and CPI is a more advanced algorithm to further perform meta-analysis on studies with known heterogeneities. In addition, 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.

5.3.1 Procedure

The MetaPath package requires the input of raw expression data as in MetaDE. There are three major steps to implement the package: pathway analysis, pathway clustering diagnostics, and pathway clustering with text mining. As shown in Figure 16, there are nine major options that need to be specified to implement the package. A complete list of all options for the package can be found in Section 6.3

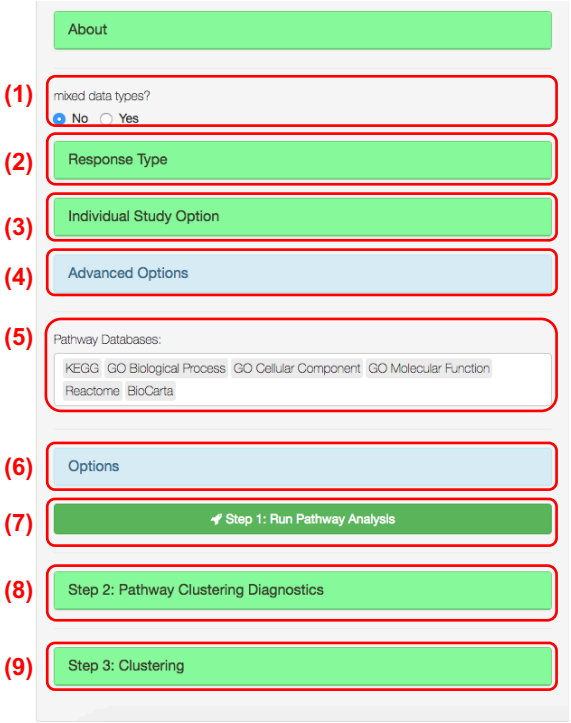


Figure 16: MetaPath options

Setup pathway analysis parameters: As shown in Figure 16, users need to 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. In (5), users can select from 25 available pathway databases for the enrichment analysis. In (6), users can select the MetaPath method (either MAPE or CPI). By default, the MAPE approach is used. Other options include the pathway enrichment method (Fisher's Exact Test or KS Test), the minimum and maximum pathway size. If Fisher's exact test is chosen for the enrichment method, users need to further specify the criteria for selection of DE genes, (e.g., the number of top ranked genes). On the other hand, if KS test is chosen, one needs to further specify whether to use permutation to obtain the enrichment p-value.

Step 1 **Run pathway analysis:** Once the above options are specified, users can click on (7) to "Run Pathway Analysis".

Step 2 **Pathway clustering diagnostics:** Since these pathways may contain redundant information, we want to cluster these pathways into certain groups. The first thing is to determine number of clusters K. Following the previous step (**Step 1**), users can specify the top enriched pathways for further clustering. The top enriched pathways can be specified by choosing the FDR cutoff by expanding the drop-down menu in (8). Then, by clicking on "Pathway Clustering Diagnostics" (Figure 16 (8)), the MetaPath module will perform consensus clustering analysis to determine the optimum number of clusters K.

Step 3 **Pathway clustering with text mining:** From the previous step (**Step 2**), users can determine the optimal number of clusters in the pool of pathways selected. Now, one can specify the number of clusters and click on (9) to get pathway clustering results.

5.3.2 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	0.001039	5.998e-7	0.001702	0.001254	0.0002194
Reactome Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	0.006311	0.000007284	0.01204	0.0007369	0.0008623
KEGG Cell adhesion molecules (CAMs)	0.007497	0.00001298	0.00781	0.0041	0.0004607
KEGG Intestinal immune network for IgA production	0.01223	0.00004152	0.002608	0.01782	0.001208
KEGG Asthma	0.01223	0.00004938	0.004537	0.005154	0.002933
KEGG Systemic lupus erythematosus	0.01223	0.00004124	0.002968	0.006707	0.002797
GO:BP immune system process	0.01223	0.00004558	0.0006081	0.003762	0.02733
KEGG Autoimmune thyroid disease	0.0123	0.00006387	0.009055	0.00174	0.00587
KEGG Allograft rejection	0.0123	0.00006387	0.009055	0.00174	0.00587
Reactome Phosphorylation of CD3 and TCR zeta chains	0.01482	0.00008549	0.002917	0.01788	0.002491

Showing 1 to 10 of 1,733 entries Previous 1 2 3 4 5 … 174 Next

Figure 17: MetaPath analysis summary. This table shows analysis results of all pathways, including individual study association analysis p-value, meta pathway analysis p-value/FDR, etc. Users can sort these pathways by clicking the p_value_meta "up arrow" button. In addition, users can search the pathway name in the search bar.

We used the AML data to demonstrate the usage of the MetaPath module with the same filtering criteria and the phenotype of interest as in the MetaDE module. Detailed descriptions of these studies can be found in Table 1. After **Step 1** is finished, a summary table was generated, as shown in Figure 17 (based on the CPI method). The full table is automatically saved in the working directory specified previously.

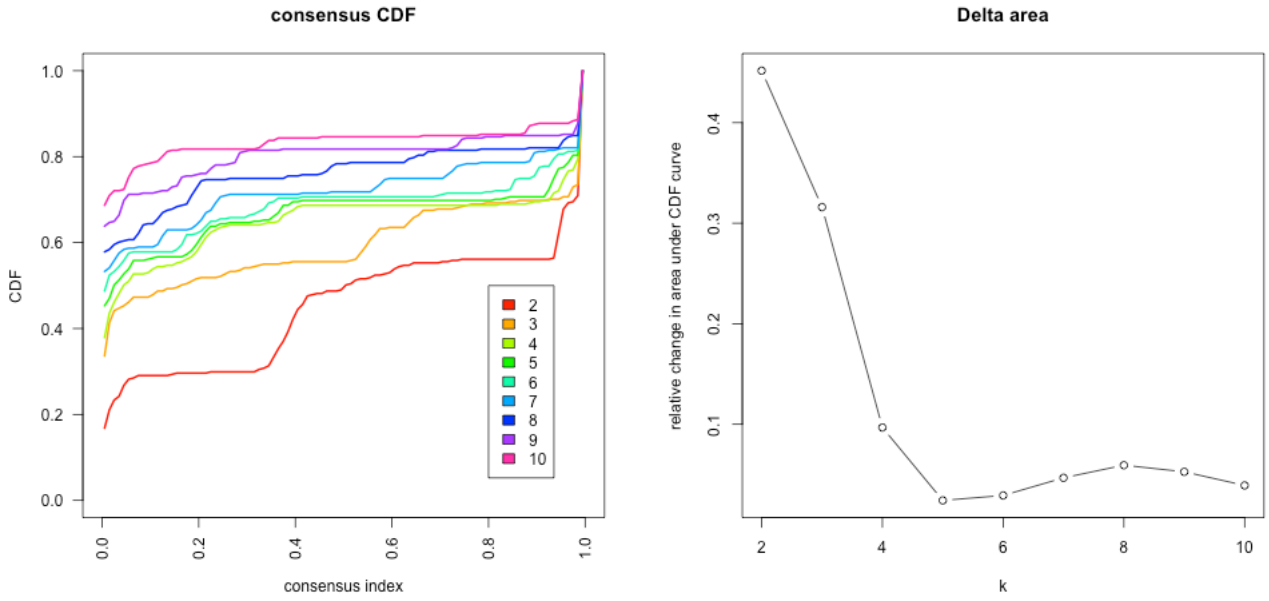


Figure 18: MetaPath results (2). Both plots assist users in finding the optimal number of clusters K, and users may refer to Monti et al. (2003) for detailed interpretation of the two plots. To be brief, the cumulative density function (CDF) of the consensus matrix for each K (indicated by colors) is 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 users 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).

In order to perform pathway clustering analysis, the number of clusters K needs to be determined. By clicking "Pathway Cluster Diagnostics" (**Step 2**), a user generates two plots on the right panel (Figure 18), including a consensus CDF plot and a Delta area plot (both from the "ConsensusClusterPlus" package). Note that in this specific example, we set the FDR cutoff value in **Step 2** as 0.4, and 27 pathways are left under this cutoff. Both plots assist users in finding the optimal number of clusters K, and a brief explanation is described in Figure 18. Users may refer to Monti et al. (2003) for a detailed interpretation of the two plots.

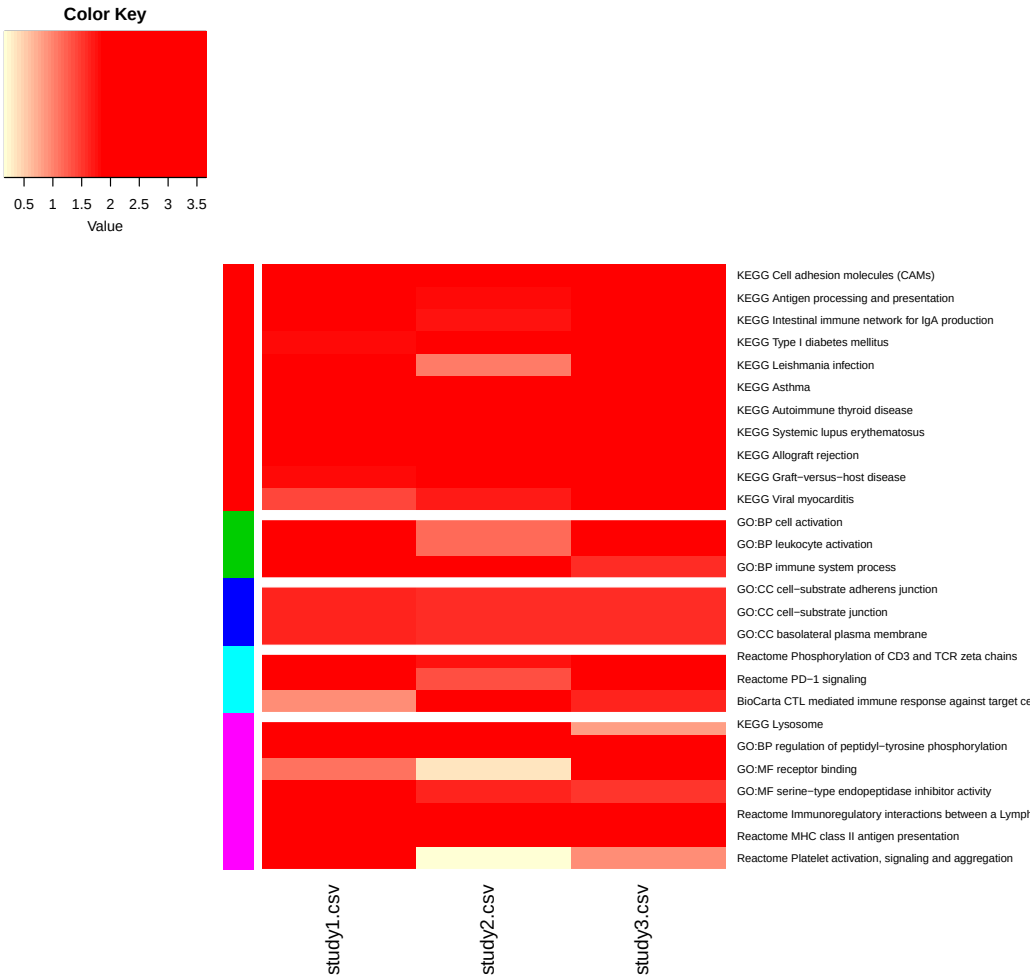


Figure 19: MetaPath results (3). The heatmap in Figure 19 shows the -log10 transformed p-value of enrichment analysis in each study from **Step 3**. Studies are on columns, and the selected pathways are on rows. The color bar on left indicates group of pathways. In the heatmap, red indicates "more enriched," 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.

The heatmap in Figure 19 shows the -log10 transformed p-value of enrichment analysis in each study from **Step 3**. 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

By clicking toolsets and then MetaNetwork, users are directed to the MetaNetwork home page, as Figure 20. 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 R package for the MetaNetwork module can be found at `https://github.com/metaOmics/MetaNetwork`.

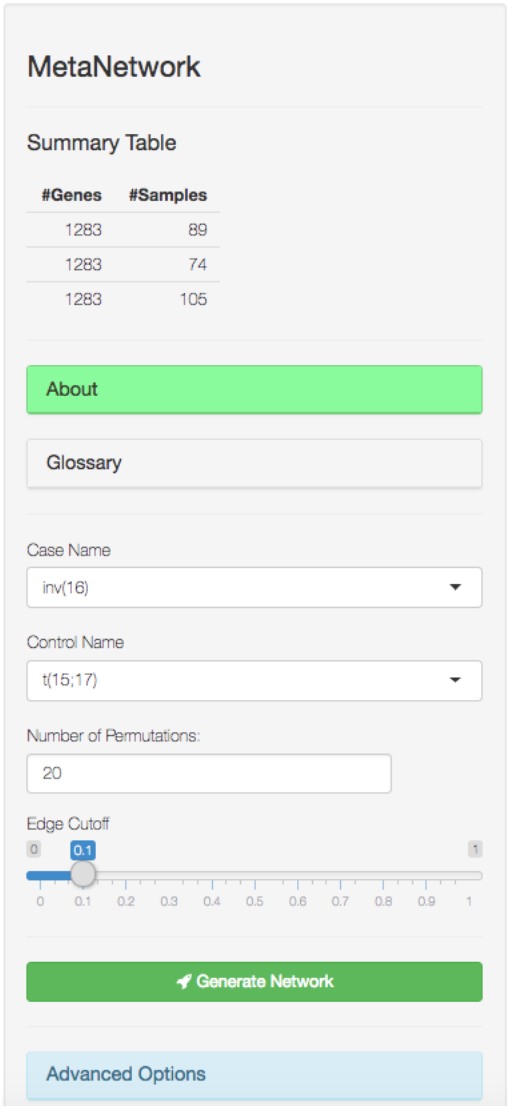


Figure 20: MetaNetwork homepage

MetaNetwork includes three steps to get differentially co-expressed networks, including (1) generate network, (2) search for basic modules, and (3) assemble supermodules. The left side of Figure 20 is the control panel of **Step 1**. The control panel for **Step 3** and **Step 3** will show up after the previous step is completed. The explanation of the complete list of all options are in Section 6.4

5.4.1 Procedure

Step 1 **Generate Network** The first step of MetaNetwork is to generate the co-expression network. In this step, the network for permuted data will also be generated. Users need to select case and control names, the number of permutations, and edge cutoff which determines the proportion of edges to be kept in the network. Permutations are used to generate the null distribution for edge energy, which will be further used to calculate edge FDR. Increasing the number of permutations can provide a more accurate FDR estimate but also increases the computation time if a single CPU is used. Given a reasonable number of edges, 3-10 permutations are recommended. A quantile cutoff for edge correlation is applied to decide if an edge exists or not. 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. After clicking the

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

Step 2 Search for basic modules

The next step is to search for basic modules. Advanced options (users are recommended not to change) include the number of repeats used for each initial seed modules (Number to repeat), the maximim Monte Carlo steps for the simulated annealing algorithm (MC Steps), and the maximum pairwise Jaccard index allowed for basic modules (Jaccard Cutoff), as shown in Figure 21. When searching for basic modules, we try multiple repeats with different seeds to avoid local optimum. Increasing the repeat times will generate more basic modules but will again raise the computation burden. We recommend starting with the default number of three repeats, and only increasing it if the number of basic modules detected is too small. If two repeats from the Monte Carlo simulation are very similar (with a Jaccard index greater than the cutoff), only one repeat with stable configuration (low energy) will be kept in the analysis. Users are advised to use the default options (Jaccard index cutoff = 0.8). Explanations of these technical terms are omitted in this tutorial, but readers can refer to Zhu et al. (2016) for details. After clicking the **Search for basic modules** button in Figure 21, the screen will show a message indicating the algorithm is running to search for basic modules. This step is computationally demanding, depending the on gene size. After this step is done, tables summarizing two kinds of basic modules: one highly connected in cases but loosely connected in control, and the other one with reverse pattern.

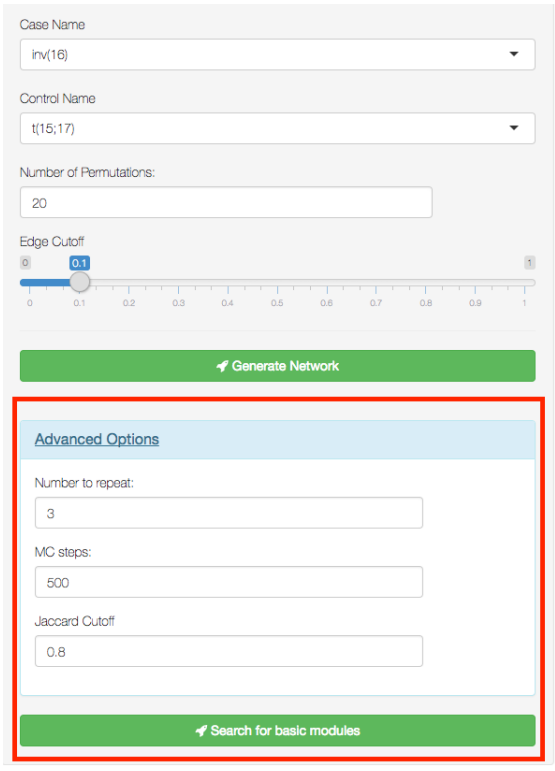
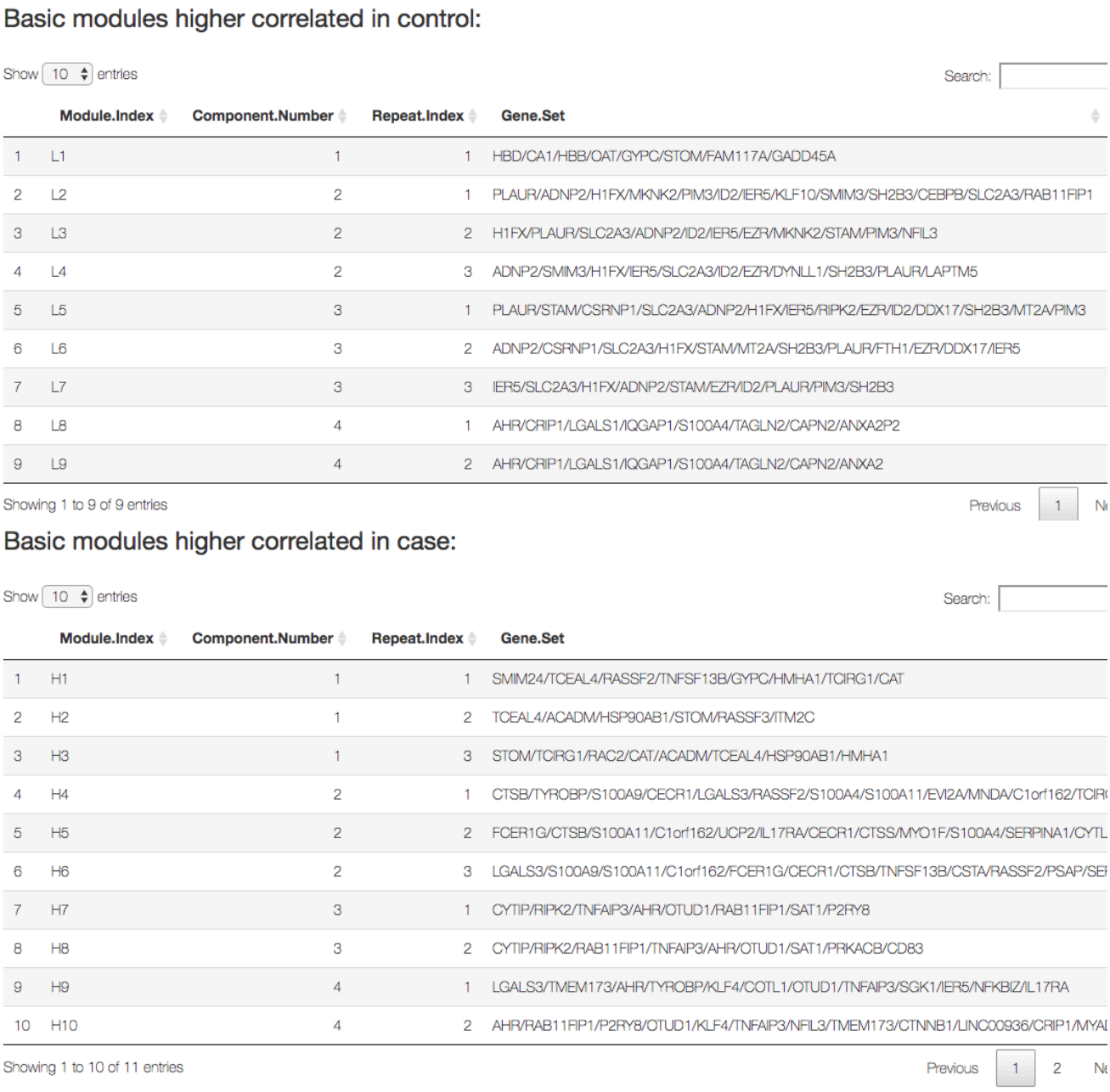


Figure 21: MetaNetwork control panel for search for basic modules

A search for basic modules can be time consuming, especially if a large number of genes are used. After this step is done, the screen will show a table of basic modules higher correlated in case and a table of basic modules higher correlated in control, as in Figure 22.



Basic modules higher correlated in control:

Show 10 entries Search:

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

Showing 1 to 9 of 9 entries Previous 1 Ne

Basic modules higher correlated in case:

Show 10 entries Search:

	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/TCIR
5	H5	2	2	FCER1G/CTSB/S100A11/C1orf162/UCP2/IL17RA/CECR1/CTSS/MYO1F/S100A4/SERPINA1/CYTL
6	H6	2	3	LGALS3/S100A9/S100A11/C1orf162/FCER1G/CECR1/CTSB/TNFSF13B/CSTA/RASSF2/PSAP/SE
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/MYA

Showing 1 to 10 of 11 entries Previous 1 2 Ne

Figure 22: MetaNetwork output from search for basic modules step. These modules higher correlated in case are labeled as H1, . . ., H6, and the modules higher correlated in control are labeled as L1, . . ., L6. The component number and repeat index are for intermediate indexes. The actual gene sets are listed for each basic module.

Step 3 Assemble supermodules

After searching for basic modules step, the control panel becomes what is shown in Figure 23. The last step is to assemble supermodules. Users can decide the FDR cutoff to select basic modules for supermodule assembly. In Figure 23, MCStep denotes the maximum number of iterations in the simulated annealing searching algorithm. A default 500 is recommended. After clicking the **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. MetaNetwork automatically creates files of top supermodules designed to input to a Cytoscape plug-in "MetaDCNExplorer" (`http://tsenglab.biostat.pitt.edu/software.htm`) for improved visualization and dynamic exploration.

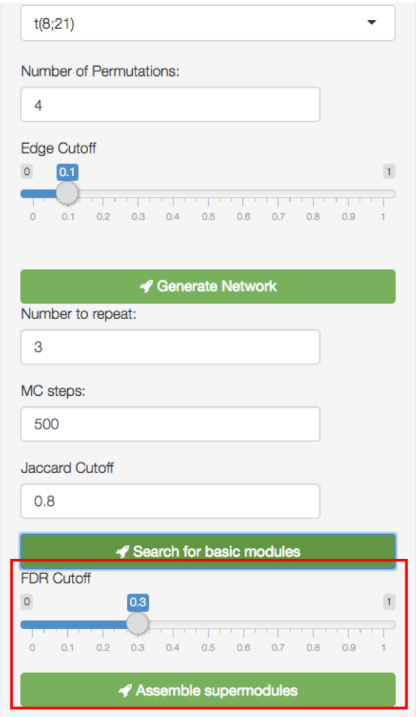


Figure 23: MetaNetwork control panel for the "Assemble supermodules" step

5.4.2 Results

We used the leukemia data to demonstrate the MetaNetwork module. After merging the three datasets by filtering out 80% of genes by mean and 80% by variance, 206 genes remained. In this example we only compared two phenotypes: inv(16) and t(15;17). Detailed descriptions of these studies can be found in Table 1. In general, the MetaNetwork tool is time consuming for large datasets (for both the "network generation" and "search for basic modules" steps). We generally suggest users carefully restrict the number of genes (e.g., less than a thousand) for a test run before applying them to a large gene set. By default, all outputs and several interim RData files will be automatically saved to the folder named "MetaNetwork" under the working directory specified in Section 2.3.

After the **Generate Network** step is done, 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. 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 22. After the **Assemble supermodules** assembly is done, the screen will show a table of supermodules (Figure 24). Users can also select basic modules to plot (Figure 25). Meanwhile several files will be saved in the MetaNetwork folder. Detailed explanations of all these intermediate files are described in Section 6.4.

MetaDCN pathway-guided supermodules

Show 10 entries Search:

pathway_name	pathway_size	p_value	q_value	size	num_gene_in_set	module_num
KEGG_LYSOSOME	121	0.00185	0.0915	24	4	2
GO_ACTIN_FILAMENT	18	0.00725	0.0915	18	2	2
GO_CORTICAL_CYTOSKELETON	20	0.00725	0.0915	18	2	2
GO_CELL_CORTEX_PART	24	0.00725	0.0915	18	2	2
GO_EXTRINSIC_TO_MEMBRANE	25	0.00907	0.0915	12	2	2
REACTOME_TRAFFICKING_AND_PROCESSING_OF_ENDOSOMAL_TLR	14	0.0109	0.0915	22	2	2
GO_RUFFLE	31	0.012	0.0915	23	2	2
REACTOME_IRON_UPTAKE_AND_TRANSPORT	36	0.0154	0.0915	26	2	2
BIOCARTA_MCALPAIN_PATHWAY	25	0.0206	0.0915	18	2	2
GO_CELL_CORTEX	39	0.0206	0.0915	18	2	2

Showing 1 to 10 of 55 entries Previous 1 2 3 4 5 6 Next

Figure 24: 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.

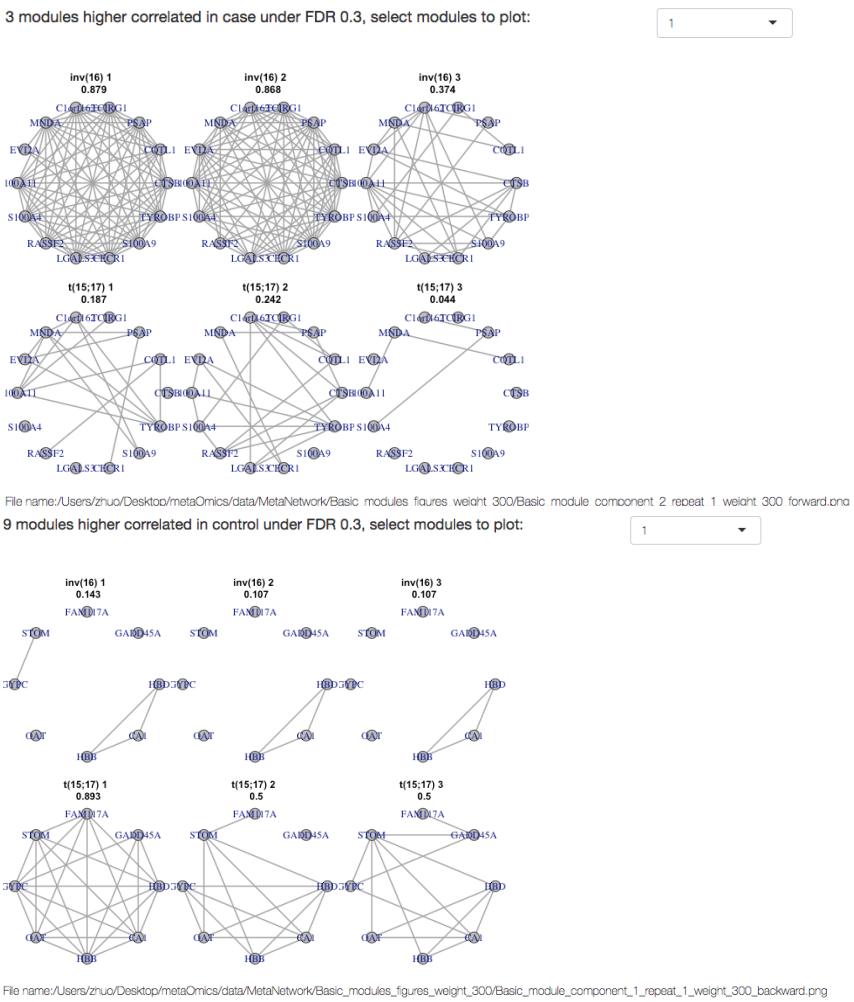


Figure 25: MetaNetwork select basic modules to plot. Each dot represents a gene. An edge represents the two genes are highly correlated. The network density is marked on top of each network. The top three modules show higher correlation in inv(16), and the bottom three modules show higher correlation in t(15:17).

5.5 MetaPredict

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. The R package for the MetaPredict module can be found at `https://github.com/metaOmics/MetaPredict`.

The homepage for MetaPredict is shown in Figure 26. Under advanced options, there is one drop-down menu ("Methods for MetaPredict") (1), "Max number of top scoring pairs (K)" (2), "Number of cores for parallel computing" (3). These will be introduced shortly, but the users are not advised to change them unless they know about the algorithm. The necessary parameters include "Number of top scoring pairs (K)" (7)), three character entries ("Please select TWO labels to cluster" (4), "Please select studies for training" (5), and "Please select studies for testing") (6), and two executing tabs ("Train model" and "Predict").

5.5.1 Procedure

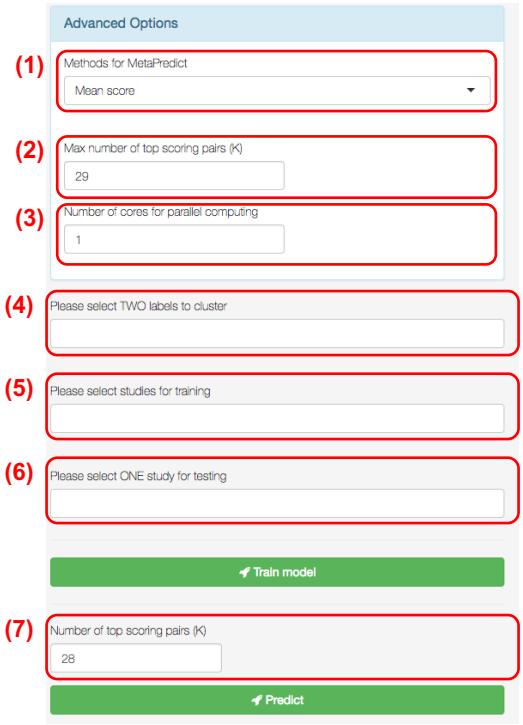


Figure 26: Homepage of MetaPredict

Step 1 Building prediction model based on meta-analysis

First, we need to decide a method to select K top scoring gene pairs from multiple studies (Figure 26). Second, we need to provide 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. Next, we need to select only two labels to build the classification model. In other words, if there exists more than two kinds of labels, we need to choose two from them. If you click on (4), all available labels will pop up. Then, select the dataset as training data and testing respectively, and click the "Train model" button to run the MetaPredict program. It may take a while to run the model.

Step 2 MetaPredict prediction

After the model training is finished, on the top right, it will show up a "Gene pair table" (Figure 27), which presents the top K gene pairs statistics. A diagnostic plot (Figure 28) is generated to assist users in deciding which K to use in the final prediction model. The suggested value is shown in the plot as a green line, which is decided by the VO method we introduced in the original paper. Users may also decide K on their own to predict the class label of testing data. After deciding K, users then hit the button "Predict" (Figure 26). Finally, a confusion matrix is output to show the prediction results (Figure 29). The prediction results are also saved in the working directory. A complete list of options is available in Section 6.5.

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

Figure 27: Results for MetaPredict, top predictive pairs of genes. A score measures the correlation between the pairs of genes.

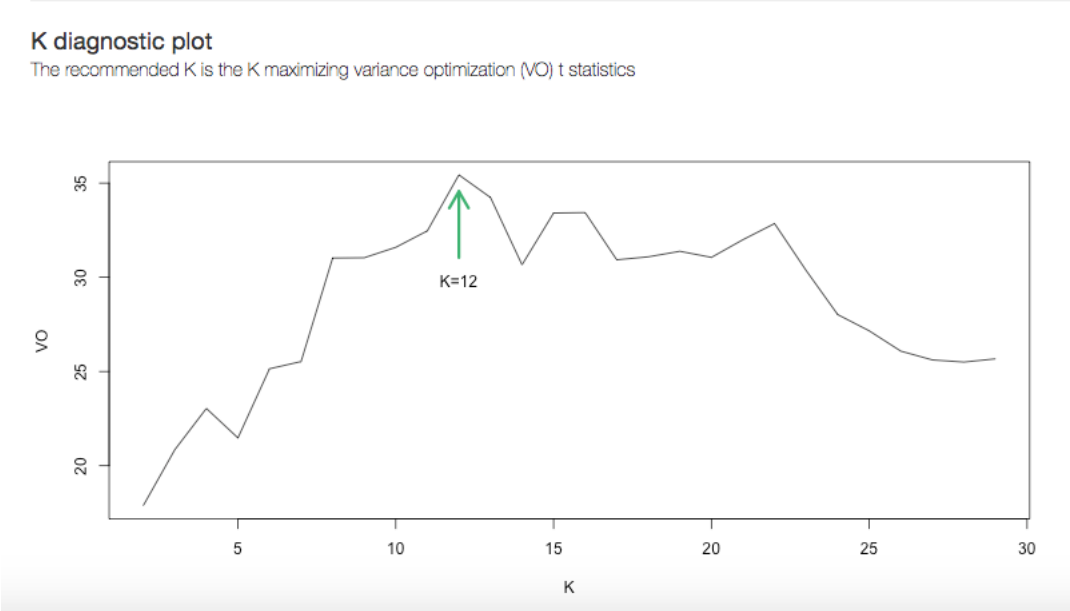


Figure 28: Results for MetaPredict. Select K by maximizing variance optimizaiton (VO) t statistics. The x-axis is the number of top scoring pairs K, and the y-axis is the variance optimization (VO) t-statistics.

Confusion Table

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

Sensitivity = 28/42 = 0.666666666666667
Specificity = 26/26 = 1

Figure 29: Confusion table for MetaPredict testing cohort

5.5.2 Results

We used the leukemia data to demonstrate the MetaNetwork module. After merging the three datasets by filtering 50% of genes by mean and 50% by variance, 1283 genes remained. In this example we only compare two phenotypes: inv(16) and t(15;17). Detailed descriptions of these studies can be found in Table 1. The top predictive pairs of genes are available in Figure 27. The diagnostic plot showing the optimum K is shown in Figure 28. A confusion matrix is output, showing the prediction results in Figure 29). The prediction results are also saved in the working directory.

5.6 MetaClust

By clicking on the Toolset tab and then choosing MetaClust, users are directed to the MetaClust homepage as in Figure 30. MetaClust (Huo et al., 2016) aims to perform sample clustering analysis combining multiple transcriptomic studies. By integrating information from multiple studies of similar biological purposes, MetaClust can identify a unified intrinsic gene set among all studies, perform weighted clustering analysis using the common intrinsic gene set, and match the clustering patterns across studies to define disease subtypes/cluster types. The resulting clustering from meta-analysis is more robust and accurate than single

study analysis. The R package for the MetaClust module can be found at `https://github.com/metaOmics/MetaSparseKmeans`.

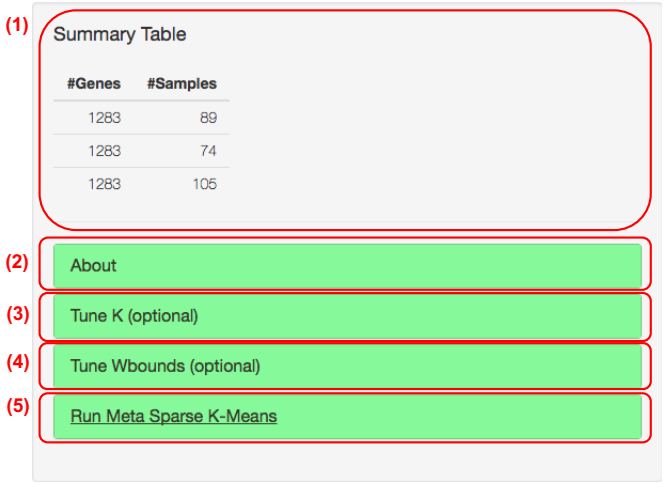


Figure 30: MetaClust homepage

5.6.1 Procedure

Figure 30 shows the homepage of MetaClust. On the top-left panel, users can see the "data summary table (1)". Below there are four tabs. "About tab (2)" describes basic introduction of MetaClust; "Tune K tab (3)" performs tuning parameter selection for the number of clusters K; "Tune Wbounds tab (4)" performs tuning parameter selection for the number of selected features, where a larger Wbound will yield more selected genes; "Run Meta Sparse K-Means tab (5)" performs the MetaClust algorithm. Starting with multiple studies, we could run MetaSparseKmeans (5) with pre-specified number of clusters (K) and gene selection tuning parameter (Wbounds). If you are not sure about what constitutes good K and Wbounds, users are advised to use the "Tune K (3)" and "Tune Wbounds (4)" panel. A complete list of options is available in Section 6.6.

Step 1 Tune K:

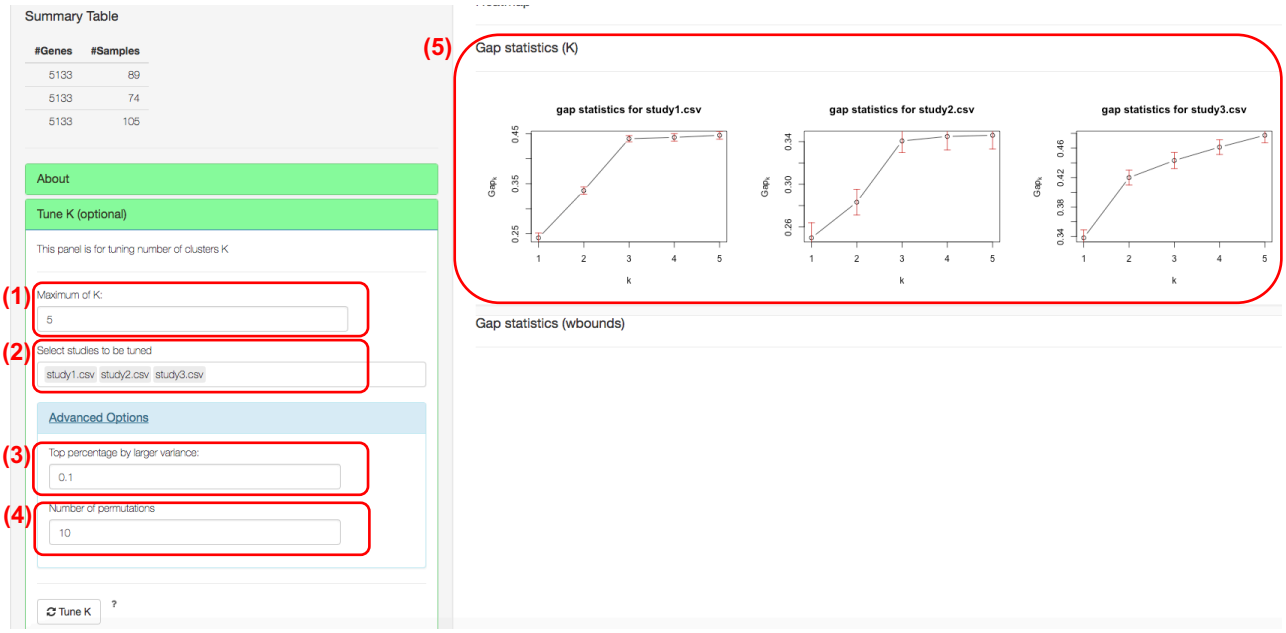


Figure 31: Tuning parameter selection for number of clusters. A good K is selected such that the Gap_k is maximized or stabilized across all studies.

If the users are not sure what is the number of clusters, they can start to use the "Tune K panel" as in Figure 31. Gap statistics will be used to get optimal K for each individual study. Detailed descriptions of the gap statistics can be found Tibshirani et al. (2001). Users need to specify the maximum number of K (1), with which the algorithm will search number of studies from 1 to K. Studies to be tuned can be selected (4). In advanced options, users can further specify the number of top variance genes to be included and number of permutations. But if users don't know the algorithm, please leave them as default. Top percentage p% by larger variance means that we will use top p% larger variance genes to perform gap statistics (3). The number of permutations is the number of bootstrap samples for gap statistics (4). At least 50 bootstrap samples are suggested for a stable result of number of clusters. By clicking the button "Tune K," users will obtain the gap statistics, as in Figure 31. 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).

Step 2 Tune Wbounds:

Wbounds directly control number of features selected by the MetaClust module. A larger number of Wbounds will result in more selected genes. If the users are not sure what is a good Wbound, they can start to use the "Tune Wbounds panel", as in Figure 32.

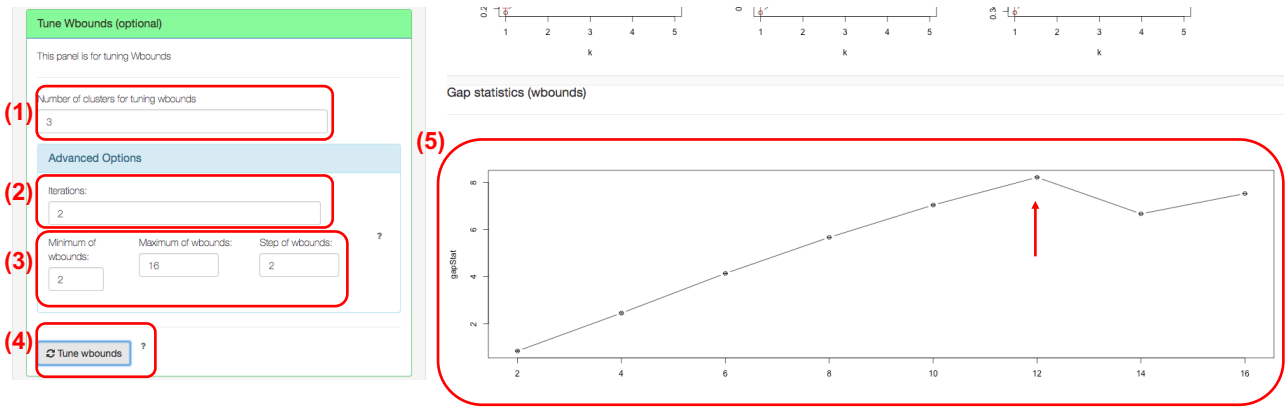


Figure 32: Wbound selection. The Wbound controls the number of selected features. A larger Wbound will yield a larger number of selected genes. The optimum Wbound is selected when the gap statistics is maximized.

Again, gap statistics will be used for tuning Wbounds. Users will specify the number of clusters for tuning Wbounds (1), which could be obtained from the previous step. In advanced options, users can further specify the number of iterations and the range of candidate Wbounds. But if users don't know the algorithm, please leave them as default. Iterations (2) specify the number of bootstrap samples for gap statistics. Users also need to specify the searching space of Wbounds by minimum of Wbounds, maximum of Wbounds, and step of Wbounds (3). After all these steps are set, the user can click on the "Tune Wbounds" button (4). The results will be shown in Figure 32 (5). Wbound=12 is preferred since the corresponding gap statistics is maximized (where the red arrow indicates).

Step 3 Run MetaClust:

Under the "Run Meta Sparse K-Means" panel, the user can specify the number of clusters (1) and Wbounds (2), and run MetaClust (5), as in Figure 33.

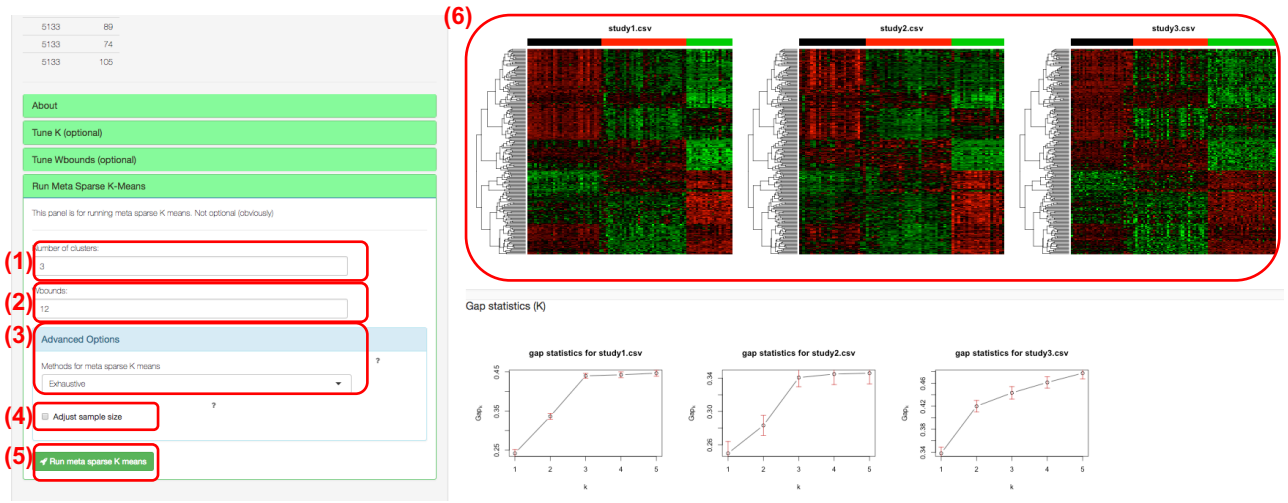


Figure 33: Result for MetaClust. The heatmap on top shows the gene-expression profiles of the three studies with selected features. Each row represents a gene, and each column represents a sample. Note that the three studies share a common set of genes. The color bar on top of the heatmap represents 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.

In advanced options (which users are not adviced to change if they are not familiar with the algorithm), there are three clustering matching methods (3): exhaustive, linear, and MCMC. Exhaustive is suggested if the data is not large. Currently, only the exhaustive search method is implemented. "Adjust sample size" checkbox (at position (5)) allows users to adjust sample size effect. After the number of clusters and Wbounds are specified, users can click on the "Run meta sparse K means" and obtain results, as shown in Figure 33.

5.6.2 Results

We used the leukemia data to demonstrate the MetaClust module. After merging the three datasets, we didn't filter out any genes (filter 0% genes by mean and 0% by variance); thus 5133 genes remained. Detailed descriptions of these studies can be found in Table 1. In this example, we do not need the extra label information. The result is shown in Figure 33 (5). We obtained unified feature selection across all studies. The clusters are well separated in each study, and the cluster patterns are consistent across all studies. The clustering heatmaps and labels are saved in the MetaClust folder.

5.7 MetaPCA

Dimension reduction is a popular data-mining approach for transcriptomic analysis. MetaPCA aims to combine multiple omics datasets of identical or similar biological hypothesis and perform simultaneous dimensional reduction in all studies. The results show improved accuracy, robustness, and better interpretation among all studies. By clicking the "Toolsets" tab and then choosing MetaPCA, users are directed to the MetaPCA homepage, as shown in Figure 34. The R package for the MetaPCA module can be found at `https://github.com/metaOmics/metaPCA`.

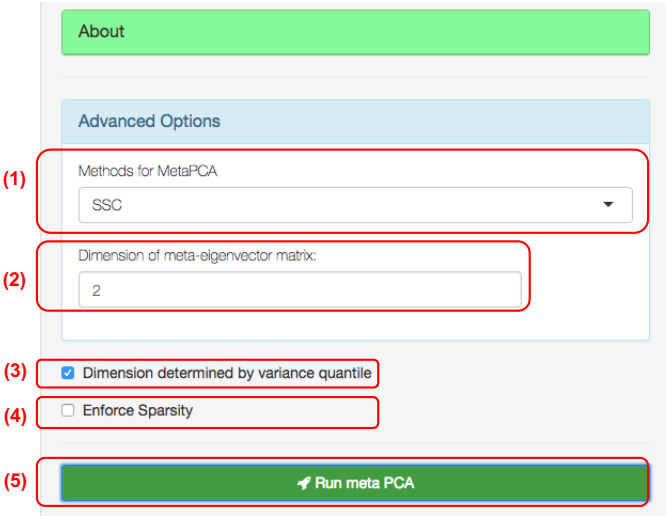


Figure 34: MetaPCA settings

5.7.1 Procedure

The procedure on describing how to use metaPCA is described below. A complete list of options is available in Section 6.7.

Step 1 Specify parameters

There are very few parameters that need to be specified for MetaPCA, as in Figure 34. Advanced options are not suggested to be changed unless the users are familiar with the algorithm. There are two methods for MetaPCA (at position (1)). SSC represents MetaPCA via sum of squared cosine (SSC) maximization. SV represents MetaPCA via sum of variance decomposition (SV). Details of SSC and SV can be found in MetaPCA manuscript (Kim et al., 2017). SSC has better performance and is suggested. The dimension of meta-eigenvector matrix option (2) allows users to specify the dimension

of the output meta-eigenvector matrix. The checkbox of "dimension determined by variance quantile" is suggested to be checked (3). When checked, the dimension size of each study's eigenvector matrix (SSC) is determined by the pre-defined level of variance quantile 80%. If the checkbox of "sparsity encouraged" is checked (at position (4)), users can perform MetaPCA. After clicking on the "search for optimal tuning parameter" button, the optimum tuning parameter will be returned to the box "tuning parameter for sparsity," which may be time consuming.

Step 2 Perform MetaPCA

By clicking the "Run Meta PCA" button, the MetaPCA module will be performed.

5.7.2 Results

The input dataset is the same as the input for MetaDE module. Detailed descriptions of these studies can be found in Table 1.

The result of MetaPCA is shown in Figure 35, which shows nice separations between three groups. These figures and eigenvectors are saved to the MetaPCA folder.

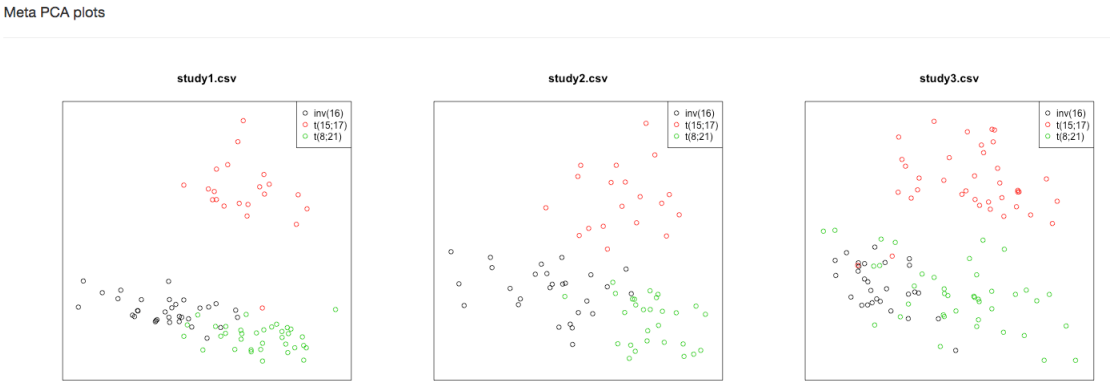


Figure 35: MetaPCA result. The x-axis (horizontal) is the first principal component, and the y-axis (vertical) is the second principal component. Each dot represents a sample in a study with the sample label marked to the top right of the figure.

6 Complete list of options

MetaPreprocess

Complete List of Options:

1. Upload expression data:
   - Header: should be checked if the input file includes a header.
   - Separator: indicates what type of separator is used for the data matrix.
   - Quote for String: how is the data matrix quoted.
   - Log transforming data: if you want to perform log transformation of your data, check yes.
   - Use existing datasets: if you want to load a dataset previously uploaded, you can choose from the checklist.
2. Annotation/impute/Replicate:
   - Annotation: possible ID type can be Gene Symbol (default), Probe ID, reference sequence ID, entrez ID.
   - Impute: if selected, missing value imputation will be performed by K-Nearest Neighbor (KNN) algorithm.

- Replicate Handling: if selected and if the same gene symbol maps to multiple probes, the probe with the largest interquartile range (IQR) will be selected as a representative for this gene.

3. Saved Data, Merging and Filtering Datasets:
   - Mean: the percentage of genes being filtered out based on the mean expression levels (e.g. 0.3 represent 30%).
   - Variance: the percentage of genes being filtered out based on the variance of expression (e.g. 0.3 represent 30%).
   - Study Name: dataset name after merging. This name will appear in the list of saved data table.
   - Merge from Selected Datasets: perform filtering and merging.
4. Danger zone:
   - Delete Selected Data: the selected data will be deleted permanently if clicked, so please be cautious.

6.1 MetaQC

Complete List of Options:

1. Options
   - Perform gene filtering: If yes: cut lowest percentile by mean, cut lowest percentile by variance.
   - Use adjusted p-value for selecting DE genes
   - p-value cutoff for selecting DE genes
   - Use adjusted p-value for selecting pathways
   - p-value cutoff for selecting pathways
2. Advanced Option (**Optional):
   - Pathway min gene size
   - Pathway max gene size
   - Number of permutations
3. Run MetaQC Analysis

6.2 MetaDE

Complete List of Options:

1. Meta Method Type: Combining p-value, Combining effect size, Others.
2. Meta Method: Fisher, AW-Fisher, FEM, REM, Sum of Rank, Product of Rank, multi-class correlation
3. Mixed data type: selected if both count data and continuous data exist.
4. Response Type:
   - Two class comparison, Multi-class comparison, Continuous outcome, Survival outcome.
   - Label Attribute: select the label name of the outcome.
   - Control Label & Experimental Label: specify the case/control label for two-class comparison.
5. Individual Study Option:
   - Setting individual study method
   - Setting individual study paired option
6. Advanced Option (**Optional):

- Use complete options
- Parametric
- Covariate
- Alternative hypothesis

7. Run
8. Pathway Databases
9. Pathway Analysis Option:
    - Pathway enrichment method
    - Pathway min gene size
    - Pathway max gene size
10. Run Pathway Analysis

6.3 MetaPath

Complete List of Options:

1. mixed data types: whether the input data is a mixture of count data and continuous data.
2. Response Type:
    - Two class comparison, Multi-class comparison, Continuous outcome, Survival outcome.
    - Label Attribute: select the label name of the outcome.
    - Control Label & Experimental Label: specify the case/control label for two-class comparison.
3. Individual Study Option:
    - Setting individual study method
    - Setting individual study paired option
4. Advanced Option (**Optional):
    - Covariate
    - Alternative hypothesis
5. Pathway Databases
6. Pathway Analysis Option:
    - Software
    - Pathway enrichment method
    - Pathway min gene size
    - Pathway max gene size
7. Run Pathway Analysis
8. Pathway Clustering Diagnostics
9. Get Clustering Result

6.4 MetaNetwork

Complete List of Options:

1. Generate Network:
   - Case Name: specify case group label.
   - Control Name: specify control group label.
   - Number of Permutations: the number of permutations used for generating network.
   - Edge Cutoff: edge cut-off determines the proportion of edges to be kept in the network.
2. Search for basic modules:
   - Number to repeat: the number of repeats used for each initial seed modules.
   - MC steps: the maximum Monte Carlo steps for simulated annealing algorithm.
   - Jaccard cutoff: maximum pairwise Jaccard index allowed for basic modules. If two repeats from Monte Carlo simulation are very similar (with Jaccard index greater than the cutoff), only one repeat with stable configuration (low energy) will be kept in the analysis.
3. Assemble supermodules:
   - FDR cutoff: FDR cut-off to select basic modules for supermodule assembly.

Intermediate results

1. Generate Network
   - AdjacencyMatrices.Rdata is a list of adjacency matrices for case and control subjects in each study. The order is study1 case, study2 case, ..., studyS case, study1 control, study2 control, ..., studyS control.
   - CorrelationMatrices.Rdata is a list of correlation matrices for case and control subjects in each study.
   - AdjacencyMatricesPermutationP.Rdata is a list of adjacency matrices for permuted datasets in permutation P.
2. Search for basic modules
   - basic_modules_summary_forward_weight_w1.csv is a summary table of basic modules that are higher correlated in case, detected using w1.
   - basic_modules_summary_backward_weight_w1.csv is a summary table of basic modules that are higher correlated in control, detected using w1.
   - threshold_forward.csv is a table of number of basic modules higher correlated in case, detected under different w1 values and FDR cut-offs.
   - threshold_backward.csv is a table of number of basic modules higher correlated in control, detected under different w1 values and FDR cut-offs.
   - permutation_energy_forward_P.Rdata is a list of energies for basic modules that higher correlated in case, detected from permutation P.
   - permutation_energy_backward_P.Rdata is a list of energies for basic modules that higher correlated in control, detected from permutation P.
3. Assemble supermodules
   - module_assembly_summary_weight_w1.csv is summary table of supermodules using w1 weight.
   - CytoscapeFiles folder contains the input files for Cytoscape to visualize supermodules.

6.5 MetaPredict

Complete List of Options:

1. Model trainings:
    - Methods for MetaPredict: include Mean score, Fisher, Stouffer.
    - Max number of top scoring pairs (K)
    - Number of cores for parallel computing
    - TWO labels to cluster: labels for MetaPredict
    - Please select studies for training
    - Please select studies for testing
    - Number of top scoring pairs (K): Number of top scoring pairs (K) for prediction.

6.6 MetaClust

Complete List of Options:

1. Tune K (** optional)
    - Maximum of K: the maximum number of K that gap statistics will step through.
    - Top percentage by larger variance: Top percentage p% by larger variance means that we will use top p% larger variance genes to perform gap statistics.
    - Number of permutaitons: Number of permutation is number of bootstrap samples for gap statistics.
    - Select studies to be tuned: Studies to be tuned.
    - Tune K: start tuning K.
2. Tune Wbounds (** optional)
    - Number of clusters for tuning wbounds: number of clusters for tuning Wbounds.
    - Iterations: Iterations are number of bootstrap samples for gap statistics.
    - Minimum of wbounds: lower bound of the searching space of Wbounds.
    - Maximum of wbounds: upper bound of the searching space of Wbounds.
    - Step of of wbounds: stepsize of the searching space of Wbounds.
    - Tune wbounds: start tuning wbounds.
3. Run Meta Sparse K-means:
    - Number of clusters: number of clusters. Can be tuned from Tune K option.
    - Wbounds: control numbers of selected features. Can be tuned from Tune Wbounds option.
    - Methods for MetaClust: Exhaustive is suggested if the data is not large. Linear will perform smart search and get solution much faster than Exhaustive, but it may yield less accuracy. MCMC might by very time consuming.
    - Adjust sample size: adjust sample size effect.
    - Run meta sparse Kmeans: start tuning wbounds.

6.7 MetaPCA

Complete List of Options:

1. Common MetaPCA parameters:
   - Methods for MetaPCA: SSC represent MetaPCA via sum of squared cosine (SSC) maximization. SV represent MetaPCA via sum of variance decomposition (SV).
   - Dimension of meta-eigenvector matrix: dimension of the output meta-eigenvector matrix.
   - Dimension determined by variance quantile: the dimension size of each study's eigenvector matrix (SSC) is determined by the pre-defined level of variance quantile 80%.
2. If sparsity encouraged is selected, there are extra tuning parameter (λ) that may need to be tuned.
   - Min λ: lower bound of the searching space of λ.
   - Max λ: upper bound of the searching space of λ.
   - Step of of λ: stepsize of the searching space of λ.
   - Tuning parameter for sparsity: Tuning parameter for sparsity that will be used for sparse MetaPCA.

References

Balgobind, B. V., den Heuvel-Eibrink, M. M. V., Menezes, R. X. D., Reinhardt, D., Hollink, I. H. I. M., Arentsen-Peters, S. T. J. C. M., van Wering, E. R., Kaspers, G. J. L., Cloos, J., de Bont, E. S. J. M., Cayuela, J.-M., Baruchel, A., Meyer, C., Marschalek, R., Trka, J., Stary, J., Beverloo, H. B., Pieters, R., Zwaan, C. M., and den Boer, M. L. (2010). Evaluation of gene expression signatures predictive of cytogenetic and molecular subtypes of pediatric acute myeloid leukemia. *Haematologica*, 96(2):221–230.

Chang, L.-C., Lin, H.-M., Sibille, E., and Tseng, G. C. (2013). Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC bioinformatics*, 14(1):368.

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19(suppl 1):i84–i90.

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M. S., et al. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clinical cancer research*, 13(11):3207–3214.

Fang, Z., Zeng, X., Lin, C.-W., Ma, T., and Tseng, G. C. (2017). Comparative pathway integrator: a framework of meta-analytic integration of multiple transcriptomic studies for consensual and differential pathway analysis. *In preparation*.

Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., and Chory, J. (2006). Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–2827.

Huo, Z., Ding, Y., Liu, S., Oesterreich, S., and Tseng, G. (2016). Meta-analytic framework for sparse k-means to identify disease subtypes in multiple transcriptomic studies. *Journal of the American Statistical Association*, 111(513):27–42.

Huo, Z., Tang, S., Park, Y., and Tseng, G. (2017). P-value evaluation, variability index and biomarker categorization for adaptively weighted fisher's meta-analysis method in omics applications. *arXiv preprint arXiv:1708.05084*.

Ivshina, A. V., George, J., Senko, O., Mow, B., Putti, T. C., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H., et al. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer research*, 66(21):10292–10301.

Kang, D. D., Sibille, E., Kaminski, N., and Tseng, G. C. (2012). Metaqc: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. *Nucleic acids research*, 40(2):e15–e15.

Kim, S., Kang, D., Huo, Z., Park, Y., and Tseng, G. C. (2017). Meta-analytic principal component analysis in integrative omics application. *Bioinformatics*, 1:8.

Kim, S., Lin, C.-W., and Tseng, G. C. (2016). MetaKTSP: A Meta-Analytic Top Scoring Pair Method for Robust Cross-Study Validation of Omics Prediction Analysis. *Bioinformatics*, 32(13):1966–73.

Kohlmann, A., Kipps, T. J., Rassenti, L. Z., Downing, J. R., Shurtleff, S. A., Mills, K. I., Gilkes, A. F., Hofmann, W.-K., Basso, G., Dell'Orto, M. C., Foà, R., Chiaretti, S., Vos, J. D., Rauhut, S., Papenhausen, P. R., Hernández, J. M., Lumbreras, E., Yeoh, A. E., Koay, E. S., Li, R., min Liu, W., Williams, P. M., Wieczorek, L., and Haferlach, T. (2008). An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the microarray innovations in LEukemia study prephase. *British Journal of Haematology*, 142(5):802–807.

Lapointe, J., Li, C., Higgins, J. P., Van De Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., et al. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 101(3):811–816.

Li, J. and Tseng, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics*, 5(2A):994–1019.

Lu, S., Li, J., Song, C., Shen, K., and Tseng, G. C. (2009). Biomarker detection in the integration of multiple multi-class genomic studies. *Bioinformatics*, 26(3):333–340.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118.

Nanni, S., Priolo, C., Grasselli, A., D'Eletto, M., Merola, R., Moretti, F., Gallucci, M., De Carli, P., Sentinelli, S., Cianciulli, A. M., et al. (2006). Epithelial-restricted gene profile of primary cultures from human prostate tumors: a molecular approach to predict clinical behavior of prostate cancer. *Molecular cancer research*, 4(2):79–92.

Shen, K. and Tseng, G. C. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, 26(10):1316–1323.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209.

Song, C. and Tseng, G. C. (2014). Hypothesis setting and order statistic for robust genomic meta-analysis. *The annals of applied statistics*, 8(2):777.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M. (1949). The american soldier: adjustment during army life.(studies in social psychology in world war ii, vol. 1.).

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Tomlins, S. A., Mehra, R., Rhodes, D. R., Smith, L. R., Roulston, D., Helgeson, B. E., Cao, X., Wei, J. T., Rubin, M. A., Shah, R. B., et al. (2006). Tmprss2: Etv4 gene fusions define a third molecular subtype of prostate cancer. *Cancer research*, 66(7):3396–3400.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.

Tseng, G. C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research*, 40(9):3785–3799.

Varambally, S., Yu, J., Laxman, B., Rhodes, D. R., Mehra, R., Tomlins, S. A., Shah, R. B., Chandran, U., Monzon, F. A., Becich, M. J., et al. (2005). Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer cell*, 8(5):393–406.

Verhaak, R. G., Wouters, B. J., Erpelinck, C. A., Abbas, S., Beverloo, H. B., Lugthart, S., Löwenberg, B., Delwel, R., and Valk, P. J. (2009). Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *haematologica*, 94(1):131–134.

Wallace, T. A., Prueitt, R. L., Yi, M., Howe, T. M., Gillespie, J. W., Yfantis, H. G., Stephens, R. M., Caporaso, N. E., Loffredo, C. A., and Ambs, S. (2008). Tumor immunobiological differences in prostate cancer between african-american and european-american men. *Cancer research*, 68(3):927–936.

Wang, X., Kang, D. D., Shen, K., Song, C., Lu, S., Chang, L.-C., Liao, S. G., Huo, Z., Tang, S., Ding, Y., Kaminski, N., Sibille, E., Lin, Y., Li, J., and Tseng, G. C. (2012). An r package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. *Bioinformatics*, 28(19):2534–2536.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120.

Welsh, J. B., Sapinoso, L. M., Su, A. I., Kern, S. G., Wang-Rodriguez, J., Moskaluk, C. A., Frierson, H. F., and Hampton, G. M. (2001). Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer research*, 61(16):5974–5978.

Yu, Y. P., Landsittel, D., Jing, L., Nelson, J., Ren, B., Liu, L., McDonald, C., Thomas, R., Dhir, R., Finkelstein, S., et al. (2004). Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *Journal of clinical oncology*, 22(14):2790–2799.

Zhu, L., Ding, Y., Chen, C.-Y., Wang, L., Huo, Z., Kim, S., Sotiriou, C., Oesterreich, S., and Tseng, G. C. (2016). Metadcn: meta-analysis framework for differential co-expression network detection with an application in breast cancer. *Bioinformatics*, 33(8):1121–1129.