

A Tutorial for MetaOmics

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 2 |
| 1.1 | Abbreviation terms | 2 |
| 2 | Preliminaries | 2 |
| 2.1 | Citing MetaOmics | 2 |
| 2.2 | How to start MetaOmics | 4 |
| 2.2.1 | Requirement | 4 |
| 2.2.2 | How to install the software | 4 |
| 2.2.3 | How to start the software | 4 |
| 2.3 | MetaOmics setting page | 4 |
| 2.4 | Question and bug report | 5 |
| 3 | Prepare data | 5 |
| 3.1 | Raw data | 5 |
| 3.2 | Clinical data | 6 |
| 3.3 | Example data in the MetaOmics software | 6 |
| 4 | MetaPreprocess | 8 |
| 4.1 | Preprocessing page | 8 |
| 4.2 | Saved Data page | 9 |
| 5 | Toolsets | 12 |
| 5.1 | MetaQC | 12 |
| 5.1.1 | Procedure | 12 |
| 5.1.2 | Results | 14 |
| 5.2 | MetaDE | 14 |
| 5.2.1 | Procedure of differential expression analysis | 15 |
| 5.2.2 | Results of differential expression analysis | 17 |
| 5.2.3 | Procedure of downstream pathway analysis | 17 |
| 5.2.4 | Result of downstream pathway analysis | 18 |
| 5.3 | MetaPath | 18 |
| 5.3.1 | Procedure | 18 |
| 5.3.2 | Results | 20 |
| 5.4 | MetaNetwork | 22 |
| 5.4.1 | Procedure | 23 |
| 5.4.2 | Results | 26 |
| 5.5 | MetaPredict | 28 |
| 5.5.1 | Procedure | 29 |
| 5.5.2 | Results | 30 |
| 5.6 | MetaClust | 30 |
| 5.6.1 | Procedure | 31 |
| 5.6.2 | Results | 34 |
| 5.7 | MetaPCA | 34 |
| 5.7.1 | Procedure | 34 |
| 5.7.2 | Results | 35 |

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.

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_tutorial.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
- Methods or tools:
 - AW-Fisher: Adaptively weighted Fisher’s method
 - CPI: Comparative pathway integrator
 - FEM: Fixed effects model
 - K-S 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

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*. *Accepted*.

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>.

2.2.1 Requirement

- R \geq 3.3.1
- Shiny \geq 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')
```

2.2.2 How to install the software

- At 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 command line:


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

2.2.3 How to start the software

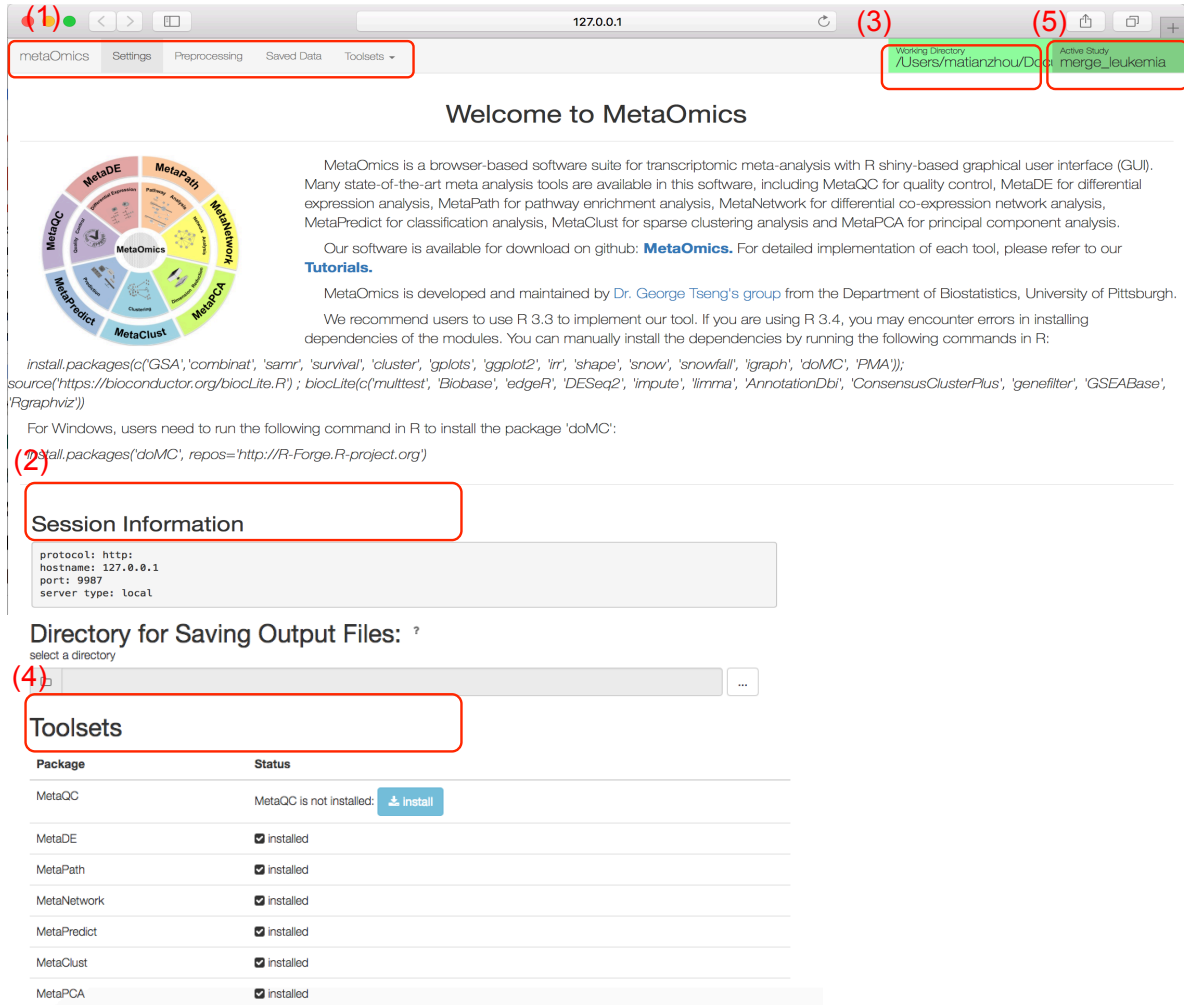
- In R (suppose the application directory is metaOmics),


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

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 the page (1) that will direct users to different pages of the software: Setting, Preprocessing, Saved Data and Toolsets. Below the tabs, there is a brief introduction of the software with additional information about the authors and maintainers. Further below, there are two headers: Session Information and Directory for Saving Output Files (2). By clicking “...”, users can set default working directory, in which all the meta-analysis results will be saved. Users can view their current working directory on the top right corner (3). There is one more section with the header Toolsets (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. There will be a notification icon at the bottom right corner after the installation. After the modules are installed, restart the MetaOmics software suite so that the Shiny application interface is updated with installed modules. (5) shows the current active dataset, which will be introduced in Section 4.1 **Step 2**.



2.4 Question and bug report

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

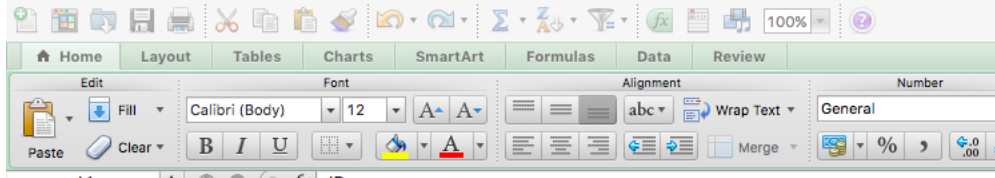
3 Prepare data

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

3.1 Raw data

Gene expression matrix should be prepared as the example in Figure 2, saved in tab-delimited “.txt” or comma-separated “.csv” files. The first column corresponds to the feature ID (e.g. gene symbol, probe id

or entrez ID) and the rest of columns are the expression data from all the samples. The first row contains the sample ID. Valid data types include continuous data (e.g. those from microarray or FPKM data from RNA-seq) and count data (e.g. those from RNA-seq).

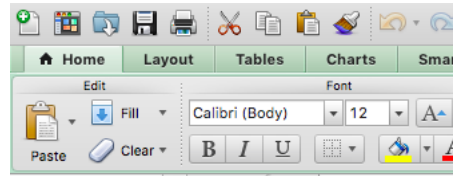


| | A | B | C | D | E | F | G | H | I | J | K |
|----|--------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| 1 | ID | GSM445939 | GSM445940 | GSM445952 | GSM445965 | GSM445966 | GSM445995 | GSM446005 | GSM446015 | 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: A example input data format

3.2 Clinical data

Clinical data should be prepared as 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-study data as the example datasets for the MetaOmics software. Table 1 summarizes the first example dataset including three studies on acute myeloid leukemia (AML). Table 2 summarizes the second example dataset including four studies on breast cancer, where the first study contains both count data and FPKM data (continuous). Table 3 summarizes the example dataset including eight studies on prostate cancer. The leukemia data is used to demonstrate MetaPreprocess, MetaDE, MetaPath, MetaNetwork, MetaPredict, MetaClust and MetaPCA. The prostate cancer data is used to demonstrate MetaQC.

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. Each 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 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 | # samples | # samples by label Normal/Primary | # genes |
|---------|--------------------------|-----------|--------------------------------------|---------|
| Study 1 | Welsh et al. (2001) | 34 | 9/25 | 8798 |
| Study 2 | Yu et al. (2004) | 146 | 81/65 | 8799 |
| Study 3 | Lapointe et al. (2004) | 103 | 41/62 | 13579 |
| Study 4 | Varambally et al. (2005) | 13 | 6/7 | 19738 |
| Study 5 | Singh et al. (2002) | 102 | 50/52 | 8799 |
| Study 6 | Wallace et al. (2008) | 89 | 20/69 | 12689 |
| Study 7 | Nanni et al. (2006) | 30 | 7/23 | 12689 |
| Study 8 | Tomlins et al. (2006) | 57 | 27/30 | 9703 |

4 MetaPreprocess

In this section, we introduce how to upload the datasets into the MetaOmics software suite to run any of the analytical modules. The R package for MetaPreprocess module can be found at <https://github.com/metaOmics/preproc>.

4.1 Preprocessing page

Step 1 Uploading data:

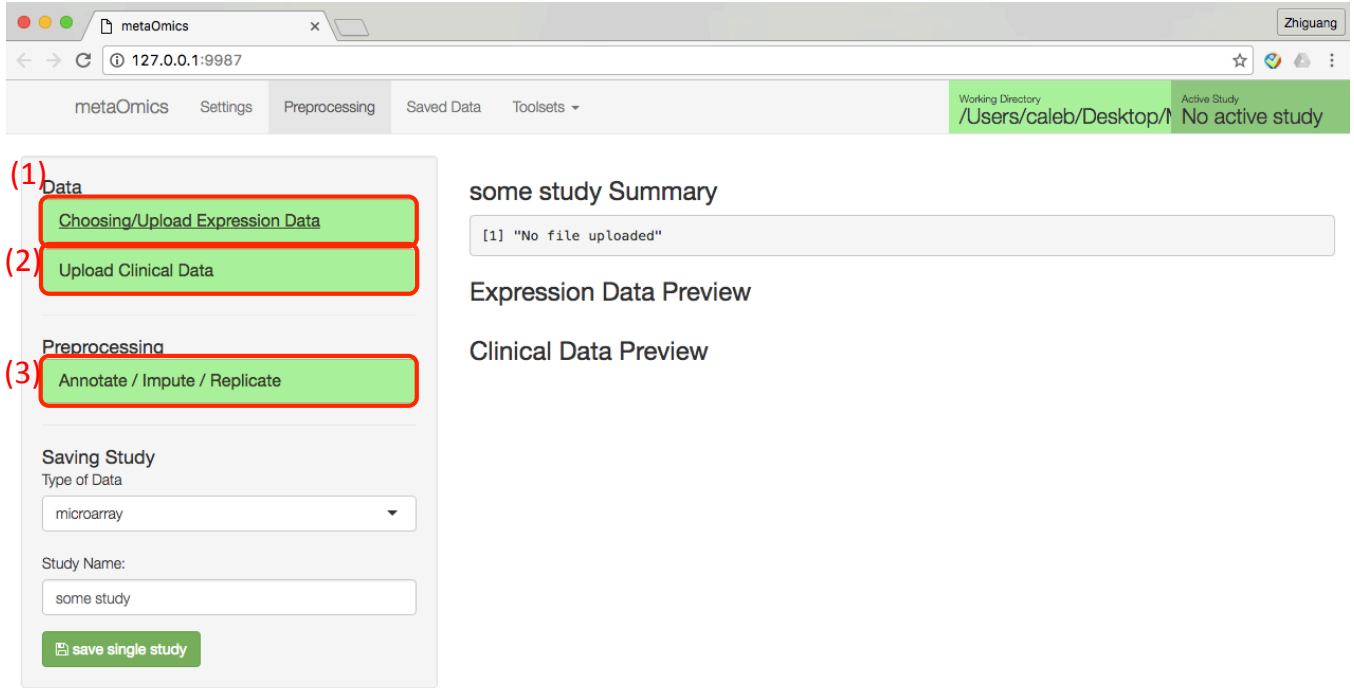


Figure 4: GUI Preprocessing page

In the Preprocessing page, users can click on “Choosing/Upload Expression Data” tab to upload individual expression data file (i.e. data file from each study) or choose the existing saved data file as in Figure 4 (1). The data should be prepared according to Section 3. Users may optionally upload Clinical Data (2), depending on their biological purposes. All MetaOmics modules except for MetaClust require external clinical labels. The three example datasets are available under MetaOmics folder “metaOmics/data/example/”, and we will mainly focus on the leukemia dataset (“metaOmics/data/example/”) throughout this tutorial.

Step 2 Preprocessing:

The MetaOmics software suite also provides handlers (3) for feature annotation, missing value imputation and multiple probe same genes. After the csv file for gene expression profile is specified, users can preview their data on the right side of the page. Several expression data parsing options (e.g. header, column separator, etc) are available on the left panel of Figure 5. For preprocessing, click on “Annotate/Impute/Replicate” to

1. annotate the probe ID/reference sequence ID/Entrez ID of individual dataset (choose Gene Symbol if the input data rows are already annotated).

2. impute missing value using knn method.
3. handle the multiple probes matching to the same gene issue.

A complete introduction of these options is available at the end of this subsection. 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.

The screenshot displays the metaOmics web application. The top navigation bar includes 'metaOmics', 'Settings', 'Preprocessing', 'Saved Data', and 'Toolsets'. The 'Working Directory' is '/Users/caleb/Desktop/' and the 'Active Study' is 'No active study'. The 'Data' section on the left has a 'Choosing/Upload Expression Data' tab. Under this tab, there is a 'Choose CSV File' section with a file named 'Balgobind_internal.csv' selected and an 'Upload complete' button. Below this is the 'Expression Data Parsing Option' section, which includes a 'Header' checkbox (checked), a 'Separator' dropdown (set to 'Comma'), a 'Quote for String' dropdown (set to 'Double Quote'), and a 'Log transforming data' checkbox (set to 'Yes'). There is also an option to 'Or use existing datasets' with a dropdown menu. At the bottom of this section is an 'Upload Clinical Data' button. To the right of the 'Data' section is the 'Balgobind_internal.csv Summary' table, which shows summary statistics for five datasets: GSM445939, GSM445940, GSM445952, GSM445965, and GSM445966. Below the summary is the 'Expression Data Preview' section, which shows a table of gene expression data for the same five datasets. The table has columns for gene names (COX1, COX2, ND4, RPL41, RPS2, RPL23A, TPT1) and rows for the five datasets. A search box is located at the top right of the preview section.

| Dataset | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|-----------|-------|---------|--------|-------|---------|-------|
| GSM445939 | 2.258 | 2.761 | 2.890 | 2.921 | 3.049 | 3.850 |
| GSM445940 | 2.157 | 2.669 | 2.783 | 2.807 | 2.920 | 3.860 |
| GSM445952 | 2.367 | 2.808 | 2.929 | 2.958 | 3.080 | 3.818 |
| GSM445965 | 2.110 | 2.842 | 2.962 | 2.992 | 3.109 | 3.826 |
| GSM445966 | 2.157 | 2.669 | 2.783 | 2.807 | 2.920 | 3.860 |

| | GSM445939 | GSM445940 | GSM445952 | GSM445965 | GSM445966 |
|--------|------------------|------------------|------------------|------------------|------------------|
| COX1 | 3.82519383224611 | 3.85987497010207 | 3.78847583019565 | 3.82584574935784 | 3.88549249642589 |
| COX2 | 3.79227705926325 | 3.8263442278892 | 3.74925555068105 | 3.77237792982363 | 3.85044260996059 |
| ND4 | 3.79079518125679 | 3.85655395569921 | 3.76141721328021 | 3.79510977041338 | 3.8653203034628 |
| RPL41 | 3.85018737983642 | 3.74937209426887 | 3.81799123492144 | 3.81809615068597 | 3.83722742706631 |
| RPS2 | 3.82155577003902 | 3.74133396216782 | 3.78754841145626 | 3.78967665724583 | 3.79670302562472 |
| RPL23A | 3.805824265197 | 3.70683948384657 | 3.78740383496946 | 3.78295822557168 | 3.75561416064749 |
| TPT1 | 3.82797772961395 | 3.74940343986743 | 3.79632081017041 | 3.79829985958213 | 3.82778917271292 |

Figure 5: Uploading individual studies

Step 3 Save single study: In the next step, specify the data type (“microarray” or “RNA-seq”, continuous or discrete) and study name, click “save single study”. To upload RNA-seq data, the count data and FPKM/TPM data should be uploaded separately and saved using different names.

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

4.2 Saved Data page

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

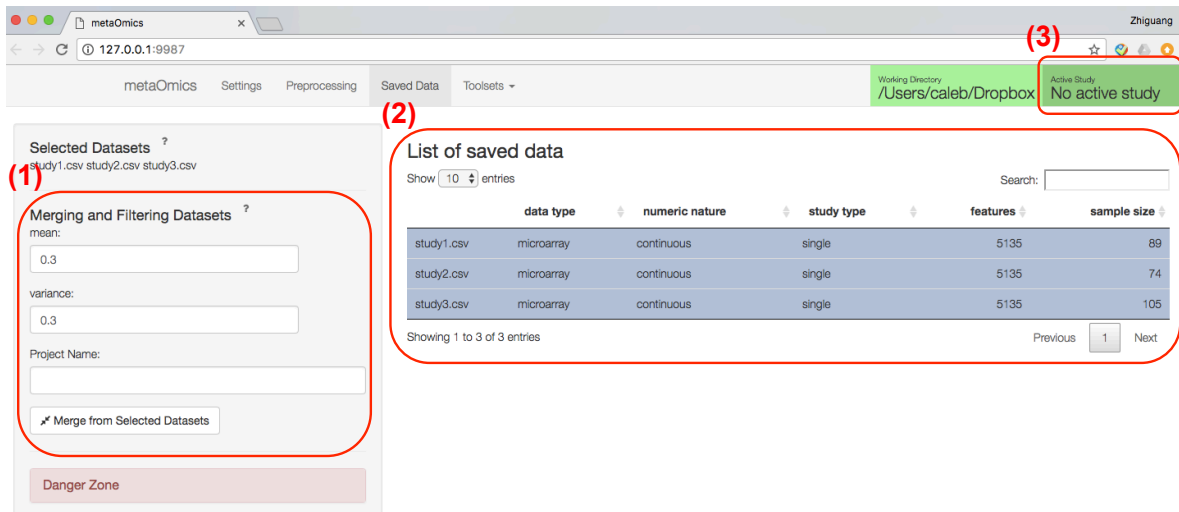


Figure 6: Merge from selected datasets

- Step 1 Merging and Filtering:** All saved datasets from the previous step will be found in Figure 6 (2). Users should select multiple datasets for further meta-analysis purpose. Users can filter out genes with low expression level (by default, mean expression lower than 30th percentile) or low variance (by default, variance lower than 30th percentile). Low expression genes can be non-expressed genes and low variance genes can be non-informative genes thus including them may contribute to false positives. After specifying filtering criteria, enter Project Name and click on the “Merge from Selected Datasets” (1). A merged dataset (study type = “multiple”) will appear on the “List of saved data” panel (3). 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 and implemented. If users want to delete any dataset, they can click 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 “Make your dataset Active Dataset” - A big green button in Figure 7. 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.

metaOmics

Settings

Preprocessing

Saved Data

Toolsets ▾

Working Directory

/Users/caleb/Desktop/r

Active Study

No active study

Selected Datasets ?

merge05

Merging and Filtering Datasets ?

You need to select more than one dataset

Danger Zone

List of saved data

Show 10 entries

Search:

| | data type | numeric nature | study type | features | sample size |
|------------|------------|----------------|------------|----------|-------------|
| study1.csv | microarray | continuous | single | 5135 | 89 |
| study2.csv | microarray | continuous | single | 5135 | 74 |
| study3.csv | microarray | continuous | single | 5135 | 105 |
| merge05 | continuous | continuous | multiple | 1283 | 268 |

Showing 1 to 4 of 4 entries

Previous 1 Next

✖ Make merge05 Active Dataset

Figure 7: Make merged Dataset Active

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 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.

5 Toolsets

After running the MetaPreprocess module, all the preparatory steps are done. It's time to apply the seven analytical modules in MetaOmics by clicking on the “Toolsets” tab and select the tool for your research question. In the next few subsections, we will introduce in details how to run each of these modules. For each module, a summary table of studies and sample sizes is shown on the top left corner. There is an “about” drop-down menu which contains a brief introduction and tutorial of the module. The “options” drop-down menu contain 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 to change unless they are familiar with the methods. After applying the modules, all result files will be automatically saved in the working directory which is specified in Section 2.3. For computationally demanding methods, the procedure may take minutes or up to hours depending on size of datasets. Users can keep track of the progress by checking the R console.

5.1 MetaQC

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: including 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. “MetaQC” package itself can be downloaded at (<https://github.com/metaOmics/MetaQC>).

5.1.1 Procedure

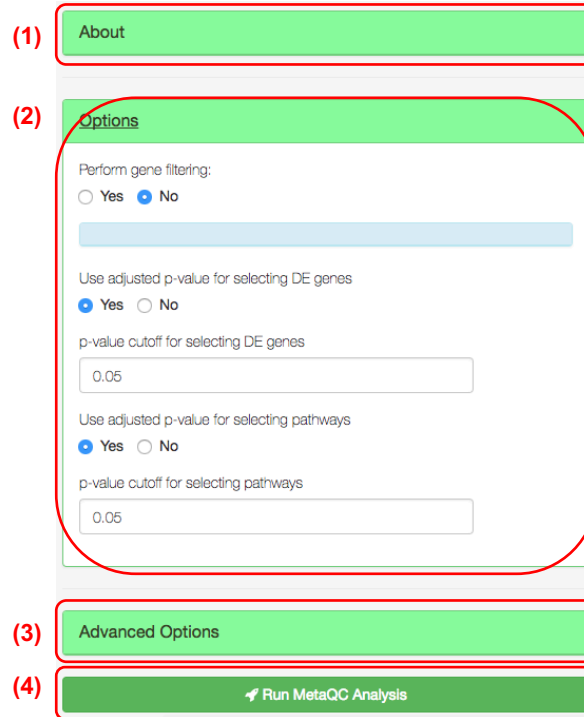


Figure 8: “MetaQC” options

There are three main options available for the “MetaQC” package as shown in Figure 8.

Step 1 Options: Under the drop-down menu “options” (2), users can

- perform gene filtering. Gene filtering is suggested to reduce computational cost. Once “Yes” is chosen for gene filtering, users are further asked to specify the filtering cutoffs by mean or by 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.

Step 2 Advanced options:

Under the drop-down menu “Advanced Options” ((3) in Figure 8), 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 detailed list of all options available for the package can be found at the end of the section. However, this is advanced options 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 (4) “Run MetaQC Analysis” to implement the tool.

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

5.1.2 Results

(1)

Show 10 entries Search:

| | IQC | EQC | AQCg | AQCp | CQCg | CQCp | SMR |
|--------|-------------------|-------------------|---------------------|------------------|-----------------------|------------------|------------------|
| p1.csv | 4.62972117515165 | 4.31884050217259 | 28.7506408159351 | 19.6553215724074 | 168.178161321905 | 29.1728506116653 | 1.83333333333333 |
| p2.csv | 9.49476305430643 | 9.76370441233525 | 22.8328911582152 | 18.5608278598915 | 161.156270084427 | 5.90587620132403 | 2 |
| p3.csv | 3.54459359448249 | 3.28895883530651 | 23.7550333901066 | 12.5641055466726 | 90.629894007444 | 0 | 4.08333333333333 |
| p4.csv | 4.24828698811111 | 4.23573392174607 | 4.1114791659993 | 8.25400998207925 | 18.7230783709056 | 0 | 4.75 |
| p5.csv | 0.894687963415149 | 0.445964759219677 | 15.2051724608639 | 64.5423701027504 | 45.0875412933291 | 0 | 4.91666666666667 |
| p7.csv | 0.813497306130717 | 0.682483123312291 | 0 | 0 | 0.389817980078202 | 0 | 7.25 |
| p6.csv | 8.16660524920825 | 7.9229350850283 | 0.00323872133424415 | 5.62346862034754 | 0.0000724622147555651 | 0 | 5.08333333333333 |
| p8.csv | 0.936649618643111 | 0.983217548518495 | 0.739905960624053 | 4.14217226240963 | 8.02899021611631 | 0 | 6.08333333333333 |

Showing 1 to 8 of 8 entries Previous 1 Next

Download Csv File

(2)

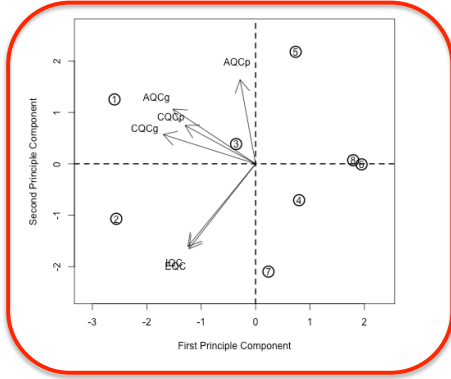


Figure 9: “MetaQC” Results

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 9, there are (1) a summary table of MetaQC results as well as (2) a PCA biplot generated. The table 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 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. Generally, studies with larger SMR values, and studies more off from the other studies and a majority of the measures are considered lower quality. In this case, the 7th and the 8th studies have relatively poorer quality. Both tabular summary and biplot are automatically saved to the working directory.

5.2 MetaDE

MetaDE package implements 12 major meta-analysis methods with 22 variations for differential expression analysis that fall into 3 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 class comparison, multi-class comparison, association with continuous or survival outcomes. The package allows the input of either microarray (continuous intensity) and/or RNA-seq data (count) for individual study analysis. The R package for MetaDE module can be found at <https://github.com/metaOmics/MetaDE>. After obtaining 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).

5.2.1 Procedure of differential expression analysis

Figure 10: “MetaDE” options

In Figure 10, Red boxes (1) - (7) are the options and steps to run the differential expression analysis. A detailed list of all options available for the package can be found at the end of this subsection.

Step 1 Choose the type of meta-analysis method: There are three types of meta-analysis to choose: combining p-values, combining effect size and others.

Step 2 Choose a meta-analysis method:

- For “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 “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 method (PR, SR and rankProd) and minMCC for multi-class meta-analysis. To choose from the overwhelmingly many meta-analysis methods, we follow Chang et al. (2013) and mark * for top performing methods AW-Fisher, REM (HO option) and rOP as recommendation 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 types of outcome (response) variable 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 group label name is needed.

Step 5 Choose Study Design for individual study:

- Individual data type can be either discrete (count) or continuous.
- Under drop-down menu “Setting Individual Study method”, user can specify individual study method 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

- Use complete options: 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 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 “Run” button to perform MetaDE analysis

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, Produce of Rank, multi-class correlation, Rank product.
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

5.2.2 Results of differential expression analysis

We used the leukemia data to demonstrate the MetaDE module. After merging the three datasets and filtering out 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. Two main outputs from the first “meta differential analysis” step in the procedure are shown in Figure 11. 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 scroll bar. In the heatmap, rows refer to DE genes selected, columns refer to samples, solid white lines are used to separate different studies and 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. Summary of meta analysis results is on bottom, including information of individual test statistics, individual study p-value, meta-analysis p-value, FDR, etc.

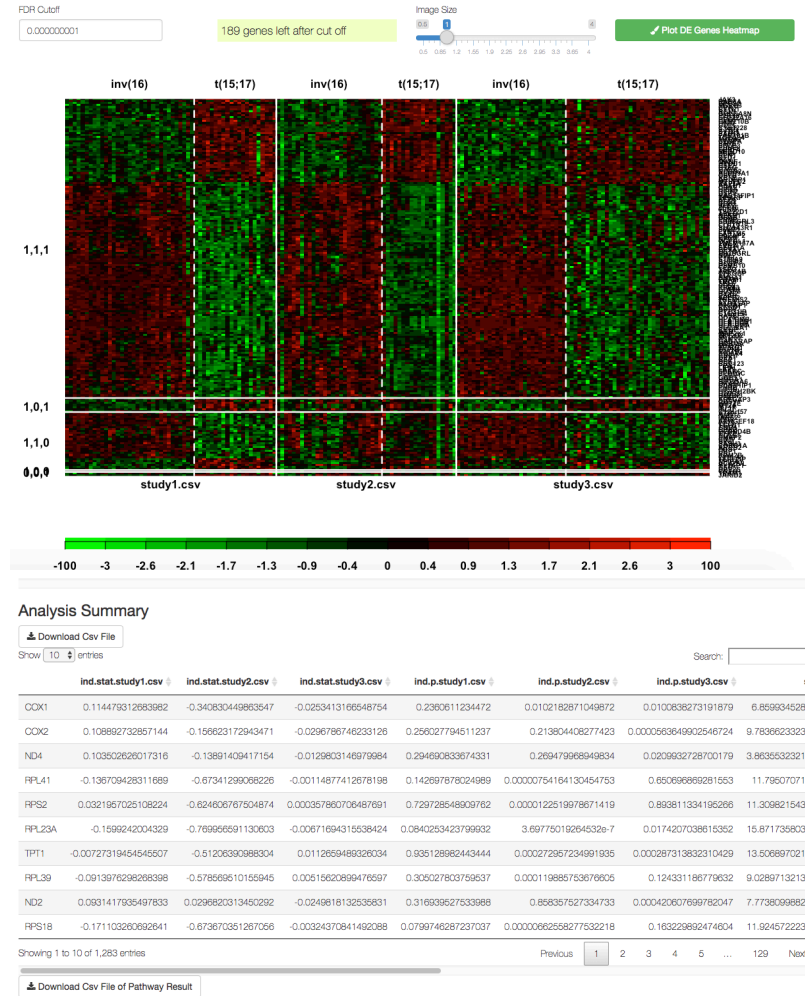


Figure 11: “MetaDE” Results

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 10. Procedure is outlined as below.

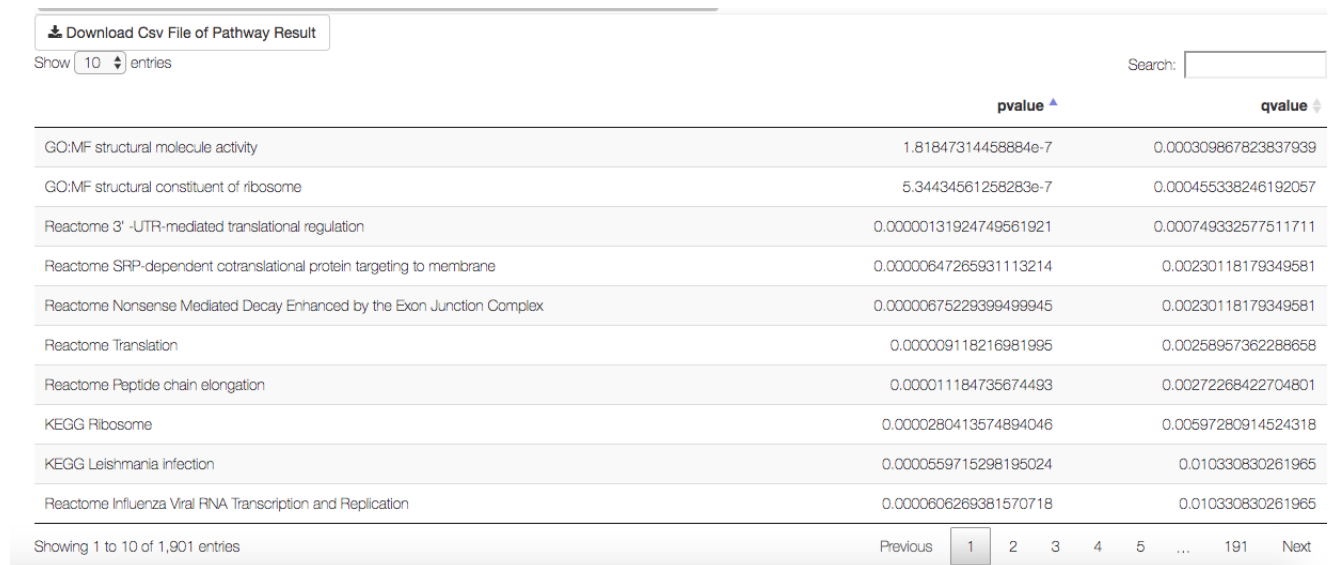
Step 1 Choose the pathway database: Users can select from 25 available pathway databases to perform the pathway enrichment analysis.

Step 2 Choose the pathway enrichment method and the pathway size range: In this step users can choose pathway enrichment options with Kolmogorov-Smirnov (KS) test as default option, or Fisher’s exact test by specifying number of input genes for pathway analysis. For Fisher’s exact test, the input genes can be obtained by either specifying a MetaDE p-value cutoff, or specifying number of top DE genes. 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 “Run Pathway analysis” button to perform pathway enrichment analysis.

5.2.4 Result of downstream pathway analysis

The result for downstream pathway analysis is shown in Figure 12. 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 “Download Csv File” on the top left of the summary table.





| | pvalue  | qvalue  |
|--|--|--|
| GO:MF structural molecule activity | 1.81847314458884e-7 | 0.000309867823837939 |
| GO:MF structural constituent of ribosome | 5.34434561258283e-7 | 0.000455338246192057 |
| Reactome 3' -UTR-mediated translational regulation | 0.00000131924749561921 | 0.000749332577511711 |
| Reactome SRP-dependent cotranslational protein targeting to membrane | 0.00000647265931113214 | 0.00230118179349581 |
| Reactome Nonsense Mediated Decay Enhanced by the Exon Junction Complex | 0.00000675229399499945 | 0.00230118179349581 |
| Reactome Translation | 0.000009118216981995 | 0.00258957362288658 |
| Reactome Peptide chain elongation | 0.000011184735674493 | 0.00272268422704801 |
| KEGG Ribosome | 0.0000280413574894046 | 0.00597280914524318 |
| KEGG Leishmania infection | 0.0000559715298195024 | 0.010330830261965 |
| Reactome Influenza Viral RNA Transcription and Replication | 0.0000606269381570718 | 0.010330830261965 |

Figure 12: Downstream pathway analysis based on MetaDE genes

5.3 MetaPath

In MetaDE package, following the detection of biomarkers, pathway analysis (a.k.a. gene set enrichment analysis) is usually performed for functional annotation and biological interpretation. Beyond that, the MetaPath module provides 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). It 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>.

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 13, there are nine major options that need to be specified to implement the package. Detailed list of all options available for the package can be found at the end of this subsection.

The screenshot shows a web interface for 'MetaPath' with the following elements and annotations:

- (1)** A radio button selection for 'mixed data types?' with 'No' selected and 'Yes' as an option.
- (2)** A green button labeled 'Response Type'.
- (3)** A green button labeled 'Individual Study Option'.
- (4)** A light blue button labeled 'Advanced Options'.
- (5)** A section titled 'Pathway Databases:' containing a list of databases: KEGG, GO Biological Process, GO Cellular Component, GO Molecular Function, Reactome, and BioCarta.
- (6)** A light blue button labeled 'Options'.
- (7)** A green button with a right arrow and the text 'Step 1: Run Pathway Analysis'.
- (8)** A green button labeled 'Step 2: Pathway Clustering Diagnostics'.
- (9)** A green button labeled 'Step 3: Clustering'.

Figure 13: “MetaPath” options

Step 1 Setup pathway analysis parameters: Users need to specify (1) whether the input gene expression profile is mixed 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 MetaPath method (either MAPE or CPI). By default, the “MAPE” approach is used, otherwise “CPI” approach can also be used. Other options include pathway enrichment method (the 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 enrichment p-value.

Step 2 Run Pathway Analysis: Once the above options are specified, users can click on (7) to “Run Pathway Analysis”.

Step 3 Pathway clustering diagnostics: From the previous step (**Step 2**), users can choose the top enriched pathways for further clustering. One can expand the drop-down menu and use FDR cutoff to choose top pathways and click on (8) “Pathway Clustering Diagnostics” to implement the second step.

Step 4 Pathway clustering with text mining: From the previous step (**Step 3**), 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. Note that you may not want to select too large K since you wish to have a certain amount of pathways in each cluster for the validity of text mining algorithm. We generally suggest users to specify K no larger than 7 for fewer than 100 pathways.

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

5.3.2 Results

Analysis Summary

Show 10 entries

Search:

| | q_value_meta | p_value_meta | study1.csv | study2.csv | study3.csv |
|--|-------------------|-------------------|-------------------|-------------------|-------------------|
| KEGG Glycolysis / Gluconeogenesis | 0.999952944432397 | 0.975713730375006 | 0.716589454064696 | 0.814873418361965 | 0.865769906161247 |
| KEGG Citrate cycle (TCA cycle) | 0.999952944432397 | 0.510889688669431 | 0.211530174577612 | 0.992916710335908 | 0.821404141054911 |
| KEGG Pentose phosphate pathway | 0.999952944432397 | 0.85144555453665 | 0.548102615399992 | 0.856164955367949 | 0.471951151317958 |
| KEGG Fructose and mannose metabolism | 0.999952944432397 | 0.433422300367834 | 0.816463757920819 | 0.752605531269185 | 0.170427359195984 |
| KEGG Galactose metabolism | 0.999952944432397 | 0.969760558119967 | 0.840628669648421 | 0.963989021054332 | 0.692998846745079 |
| KEGG Ascorbate and aldarate metabolism | 0.999952944432397 | 0.351690352701752 | 0.663648742104101 | 0.176746098320381 | 0.163240921127113 |
| KEGG Fatty acid metabolism | 0.999952944432397 | 0.785586400080126 | 0.584204394255307 | 0.403175881949791 | 0.750935078799842 |
| KEGG Steroid biosynthesis | 0.999952944432397 | 0.241772080559306 | 0.151118401862353 | 0.171251505076894 | 0.145612810184764 |
| KEGG Primary bile acid biosynthesis | 0.999952944432397 | 0.939258438255731 | 0.606195296926418 | 0.617063475433555 | 0.989348045492025 |
| KEGG Steroid hormone biosynthesis | 0.999952944432397 | 0.731719469083418 | 0.993283631448111 | 0.356962695144036 | 0.993038039433862 |

Showing 1 to 10 of 1,704 entries

Previous 1 2 3 4 5 ... 171 Next

Figure 14: “MetaPath” Results (1)

We used the leukemia data with the same filtering criteria as in MetaDE module to demonstrate MetaPath module and we also compare two phenotypes: inv(16) and t(15;17). Detailed descriptions of these studies

can be found in Table 1. After **Step 2** is finished, a summary table was generated as shown in Figure 14 (based on the CPI method). The “Analysis Summary” includes the analysis results of all pathways, including individual study association analysis p-value, meta pathway analysis p-value/FDR, etc. Users can search the gene name in the “Search” bar, and the full table is automatically saved in the working directory specified before.

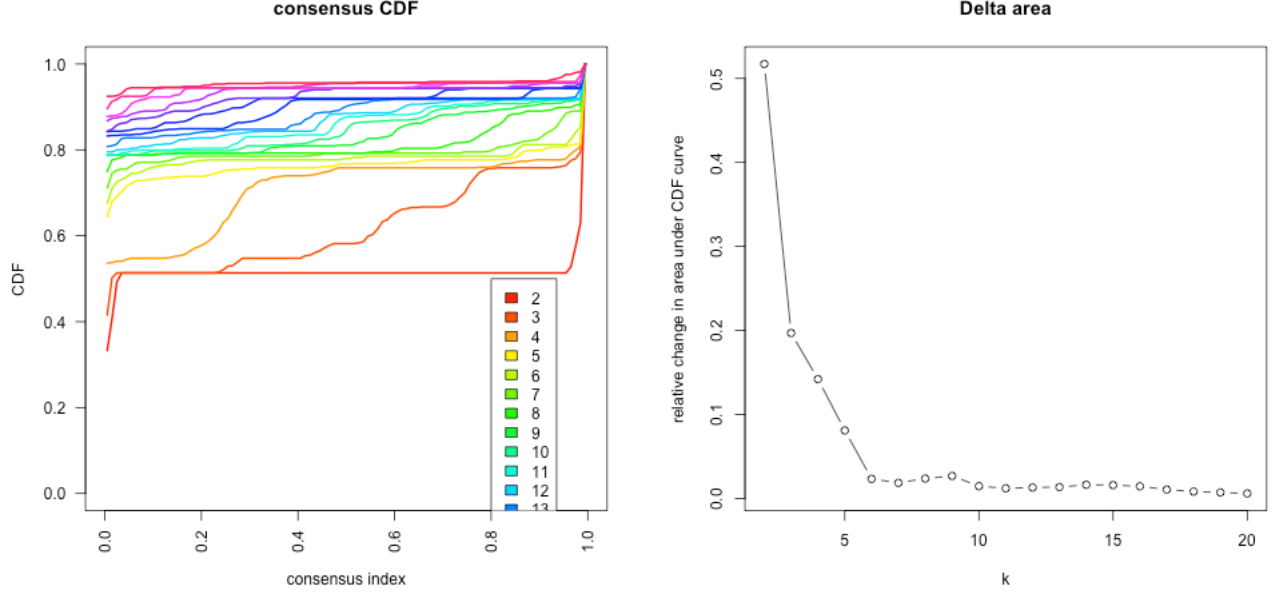


Figure 15: “MetaPath” Results (2)

After the “Pathway Cluster Diagnostics” step is finished, we will see two plots generated on the right panel (Figure 15): consensus CDF and Delta area plots, both from the “ConsensusClusterPlus” package. The 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 allows users to determine K at which there is no appreciable increase in CDF (which drops as the number of cluster increases). Both plots assist users in finding the optimal number of clusters K and you may refer to (Monti et al., 2003) for more detailed interpretation of the two plots. In the demo example, $K = 6$ is chosen since it locates at the elbow turning point (i.e. where the magnitude of incremental decrease in delta area diminishes).

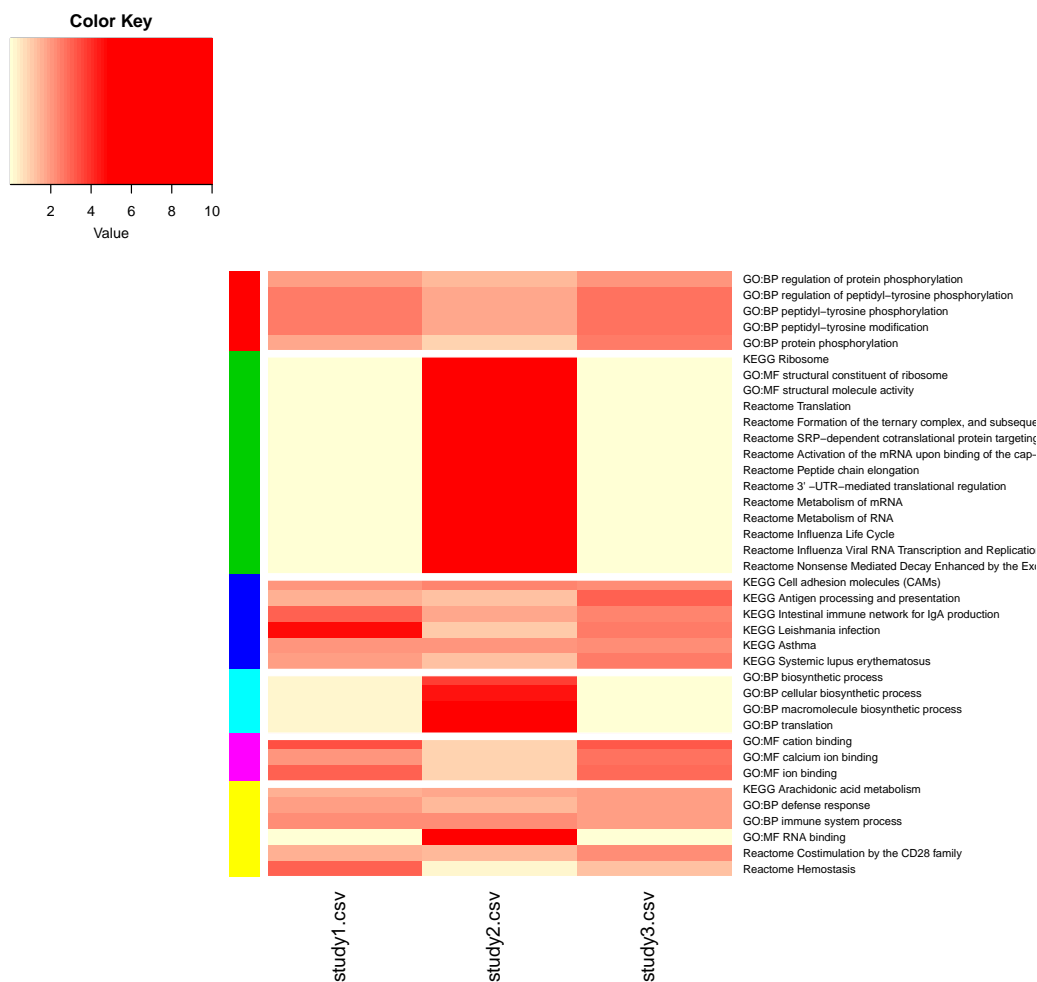


Figure 16: “MetaPath” Results (3)

The heatmap in Figure 16 shows the $-\log_{10}$ transformed p-value of enrichment analysis in each study from **Step 2**. Studies are on columns and the selected pathways are on rows, red means more enriched. The pathways are sorted by the pathway clusters as indicated by the colors on the left side of the heatmap. In addition, key words of each cluster of pathways are extracted and analyzed by a built-in text mining algorithm and 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 MetaNetwork home page as Figure 17. The R package for MetaNetwork module can be found <https://github.com/metaOmics/MetaNetwork>.

The screenshot shows the MetaNetwork homepage with a navigation bar at the top containing 'metaOmics', 'Settings', 'Preprocessing', 'Saved Data', and 'Toolsets'. On the right, there are two status boxes: 'Working Directory /Users/lizhu/Box' and 'Active Study merge'.

The main control panel on the left includes the following fields and controls:

- Case Name:** A dropdown menu with 'inv(16)' selected.
- Control Name:** A dropdown menu with 't(8;21)' selected.
- Number of Permutations:** A text input field containing the value '4'.
- Edge Cutoff:** A horizontal slider ranging from 0 to 1, with a blue marker positioned at 0.1.
- Generate Network:** A green button with a lightning bolt icon and the text 'Generate Network'.

Figure 17: MetaNetwork homepage

MetaNetwork includes three steps to get differentially co-expressed networks: generate network, search for basic modules, and assemble supermodules. The left side is the control panel of step 1. The control panel for next step will show up after the previous step is done.

5.4.1 Procedure

Step 1 Generate Network The first step of MetaNetwork is to generate 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 cut-off which determines the proportion of edges to be kept in the network. After clicking **Generate Network** button, screen will show message indicating the algorithm is running to generate network.

Step 2 Search for basic modules

The next step is to search basic modules. Advanced options (recommended not to change) include the number of repeats used for each initial seed modules ("Number of repeat"), the maximum Monte Carlo steps for simulated annealing algorithm (MC Steps), and the maximum pairwise Jaccard index allowed for basic modules (Jaccard Cutoff), as shown in Figure 18. After clicking **Search for basic modules** button, screen will show message indicating the algorithm is running to search for basic modules. This step is computationally demanding depending on gene size. After this step is done, the screen will show a table of basic modules highly connected in cases but lose connections in control and vice versa.

Case Name

inv(16)

Control Name

t(15;17)

Number of Permutations:

20

Edge Cutoff

00.10.20.30.40.50.60.70.80.901

Generate Network

Advanced Options

Number to repeat:

3

MC steps:

500

Jaccard Cutoff

0.8

Search for basic modules

Figure 18: MetaNetwork control panel for search for basic modules

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 19.

Basic modules higher correlated in case:

Show entries

Search:

| | Module.Index | Component.Number | Repeat.Index | Gene.Set |
|---|--------------|------------------|--------------|---|
| 1 | H1 | 2 | 1 | PSAP/CTSS/CECR1/LGALS3/AP1S2/TLR2/HEXB/SMIM24/CTSB/OGFRL1/MYO1F/CPXM1/SERPINA1/MNDA/TNF |
| 2 | H2 | 2 | 2 | SERPINA1/AP1S2/CECR1/SMIM24/PSAP/TRBV27/RGS10/CPXM1/TLR2/CAPN2/OGFRL1/MS4A6A |
| 3 | H3 | 2 | 3 | CTSB/TLR2/RGS10/HEXB/SERPINA1/CECR1/LGALS3/AP1S2/SMIM24/S100A9/CPXM1/MAP2K1/CAPN2 |
| 4 | H4 | 6 | 1 | NFIL3/IER3/CD83/CPXM1/EZR/RIK2/SLC2A3 |
| 5 | H5 | 6 | 2 | NFIL3/IER3/LCP1/CDKN1A/TCEAL4/TGFB1 |
| 6 | H6 | 6 | 3 | NFIL3/LYN/IER3/LCP1/UCP2/MARCKSL1/RAB11FIP1 |

Showing 1 to 6 of 6 entries

Previous Next

Basic modules higher correlated in control:

Show entries

Search:

| | Module.Index | Component.Number | Repeat.Index | Gene.Set |
|---|--------------|------------------|--------------|---|
| 1 | L1 | 2 | 1 | GCA/FABP5/PRR11/PCNA/C1QBP/MAFF/FAM107B/PRDX3/TNFSF13B/PRTN3/PRKACB/CBFB/CAT/RAB10/AN |
| 2 | L2 | 2 | 2 | ICAM3/TGFB1/TYROBP/PLP2/BIN2/HCST/MYO1F/TIMP1/LAPTM5/S100A9 |
| 3 | L3 | 2 | 3 | PCNA/FAM107B/GCA/PRTN3/C1QBP/KIAA0101/EZR/RAB10/CAT/PRDX3/TNFSF13B/ANP32E/DYNLL1 |
| 4 | L4 | 6 | 1 | SLC2A3/JUP/LYN/PPP1R15A/EZR/PCNA/RAB10/FLNA/PIM3/KLF6 |
| 5 | L5 | 6 | 2 | STAM/JUP/LYN/PIM3/PDLIM1/HLA-DPA1/ITM2A/TPSAB1/CSTA |
| 6 | L6 | 6 | 3 | SLC2A3/STAM/JUP/LYN/PPP1R15A/PCNA/FLNA/PLP2/HLA-DPA1 |

Showing 1 to 6 of 6 entries

Previous Next

Figure 19: MetaNetwork output from search for basic modules step

Step 3 Assemble supermodules

After search for basic modules step is done, the control panel becomes that in Figure 20. The last step is to assemble supermodules. Users can decide the FDR cut-off to select basic modules for supermodule assembly. After clicking **Assemble supermodules** button, 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.

t(8;21)

Number of Permutations:
4

Edge Cutoff
0 0.1 1

Generate Network

Number to repeat:
3

MC steps:
500

Jaccard Cutoff
0.8

Search for basic modules

FDR Cutoff
0 0.3 1

Assemble supermodules

Figure 20: MetaNetwork control panel for assemble supermodules step

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.
3. Assemble supermodules:
 - FDR cutoff: FDR cut-off to select basic modules for supermodule assembly.

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 compare two phenotypes: inv(16) and t(8;21). Detailed descriptions of these studies can be found in Table 1. In general, the MetaNetwork tool is time consuming for large datasets (for both network generation and search for basic modules steps). We generally suggest users to carefully restrict the number of genes (e.g. less than a thousand) for a test run before implementing 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.

Generate Network

After the generate network step is done, no output will show up in the screen. Instead, a message box will show up indicating several Rdata files are saved in the MetaNetwork folder, including:

- 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.

Search for basic modules

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 Figure 19. Meanwhile, several files will be saved in the MetaNetwork folder:

- 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.

Assemble supermodules

After supermodule assembly is done, screen will show a table of supermodules (Figure 21). Users can also select basic modules to plot (Figure 22). Meanwhile several files will be saved in the folder MetaNetwork:

- 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.

MetaDCN pathway-guided supermodules

Show

10

 entries

Search:

| pathway_name | pathway_size | p_value | q_value | size |
|--|--------------|----------|---------|------|
| KEGG_LYSOSOME | 121 | 9.73e-05 | 0.01 | 20 |
| REACTOME_MHC_CLASS_II_ANTIEN_PRESENTATION | 91 | 0.00697 | 0.0393 | 20 |
| REACTOME_PLATELET_ACTIVATION_SIGNALING_AND_AGGREGATION | 208 | 0.00495 | 0.0393 | 25 |
| REACTOME_RESPONSE_TO_ELEVATED_PLATELET_CYTOSOLIC_CA2_ | 89 | 0.00281 | 0.0393 | 22 |
| BIOCARTA_MCALPAIN_PATHWAY | 25 | 0.00283 | 0.0393 | 30 |
| GO_CYTOSKELETAL_PROTEIN_BINDING | 159 | 0.00185 | 0.0393 | 13 |
| REACTOME_COSTIMULATION_BY_THE_CD28_FAMILY | 63 | 0.00431 | 0.0393 | 14 |
| GO_ACTIN_FILAMENT_BINDING | 25 | 0.00369 | 0.0393 | 13 |
| BIOCARTA_CFTR_PATHWAY | 12 | 0.00725 | 0.0393 | 18 |
| GO_ACTIN_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS | 105 | 0.00697 | 0.0393 | 20 |

Showing 1 to 10 of 103 entries

Previous

1

2

3

4

5

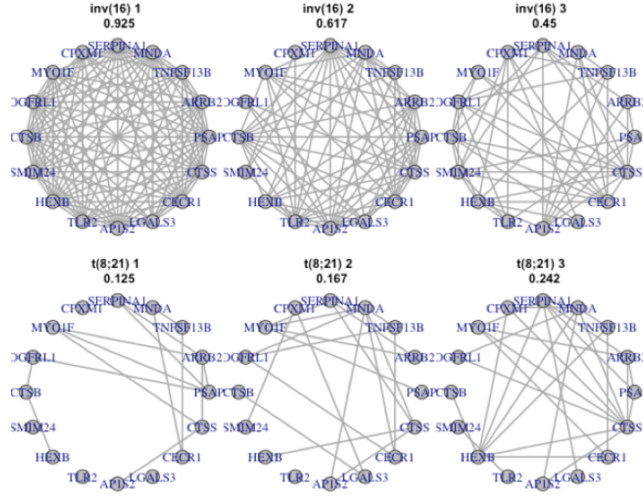
...

11

Next

Figure 21: MetaNetwork supermodules table

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



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

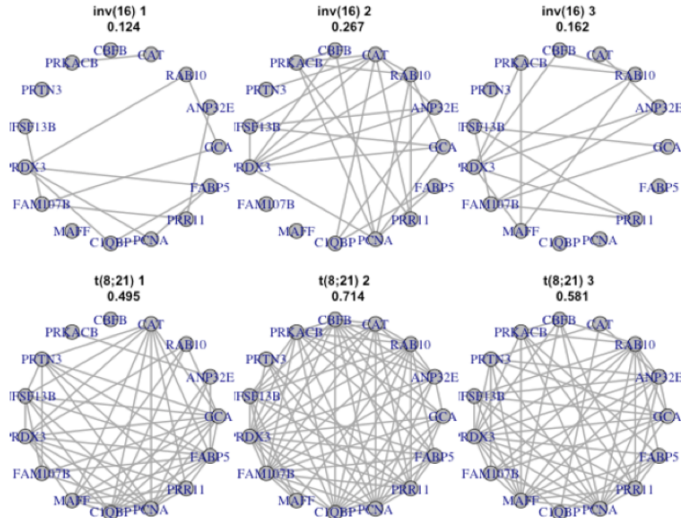


Figure 22: MetaNetwork select basic modules to plot

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 MetaPredict module can be found at <https://github.com/metaOmics/MetaPredict>.

The homepage for MetaPredict is shown in Figure 23. Under advanced options, there are 1 drop-down menu (“Methods for MetaPredict”) (1), three number entries (“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 suggested 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

Advanced Options

(1) Methods for MetaPredict
Mean score

(2) Max number of top scoring pairs (K)
29

(3) Number of cores for parallel computing
1

(4) Please select TWO labels to cluster

(5) Please select studies for training

(6) Please select ONE study for testing

Train model

(7) Number of top scoring pairs (K)
28

Predict

Figure 23: 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 23). 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” ((1) in Figure 24) which present the top K gene pairs statistics. A diagnostic plot ((2) in Figure 24) is outputted to assist users decide which K to use in the final prediction model. The suggested value is shown in the plot as green line, which is decided by 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 , then hit the button “Predict” (Figure 24). Finally, a confusion matrix is output to show the prediction results ((1) in Figure 24). The prediction results are also saved in the working directory.

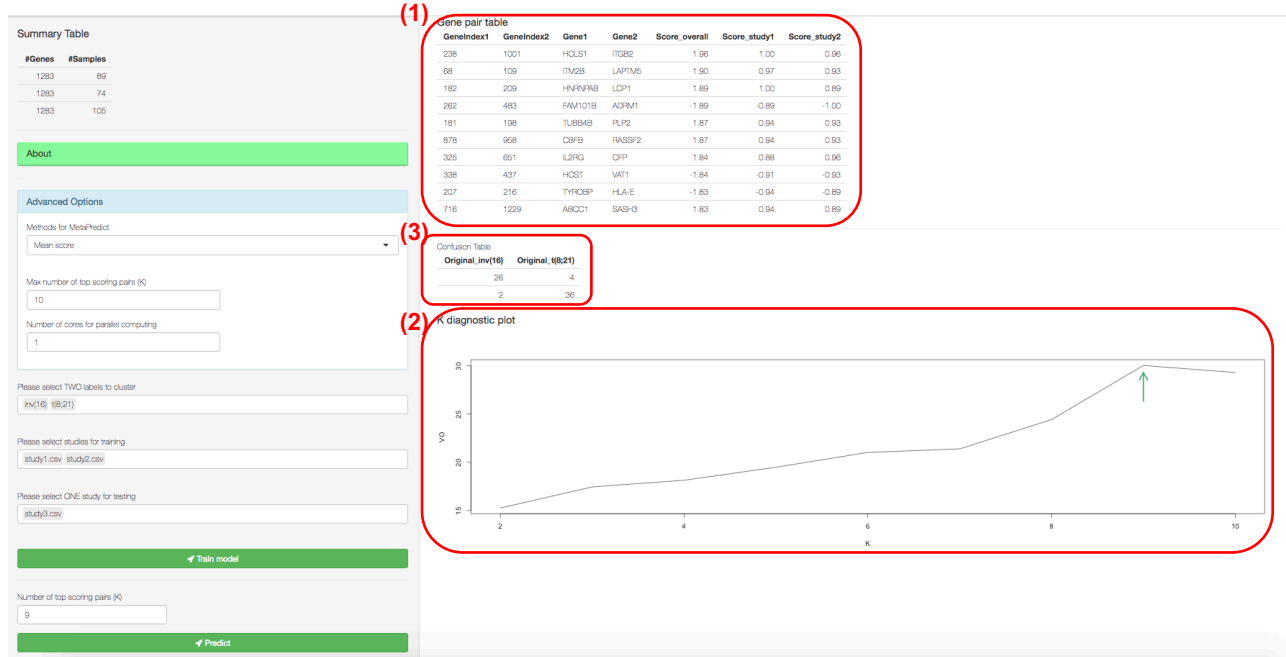


Figure 24: Results for MetaPredict.

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.

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(8;21). Detailed descriptions of these studies can be found in Table 1. A confusion matrix is output to show the prediction results ((1) in Figure 24). The prediction results are also saved in the working direction.

5.6 MetaClust

By clicking on the Toolset tab and then choose MetaClust, users are directed to MetaClust home page as in Figure 25. 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 an 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 MetaClust module can be found at <https://github.com/metaOmics/MetaSparseKmeans>.

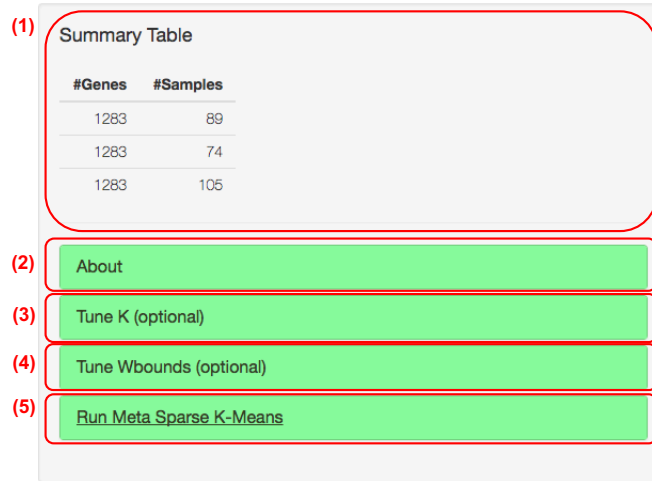


Figure 25: MetaClust home page

5.6.1 Procedure

Figure 25 shows the home page of MetaClust. On the top left panel users can see the data summary Table (1). Below there are four tabs. About tab (2) which includes basic introduction of MetaClust, Two tabs for Tune K as well as Wbounds, and the Run Meta Sparse K-Means tab. 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 are good K and Wbounds, please try Tune K (3) and Tune Wbounds (4) panel.

Step 1 Tune K :

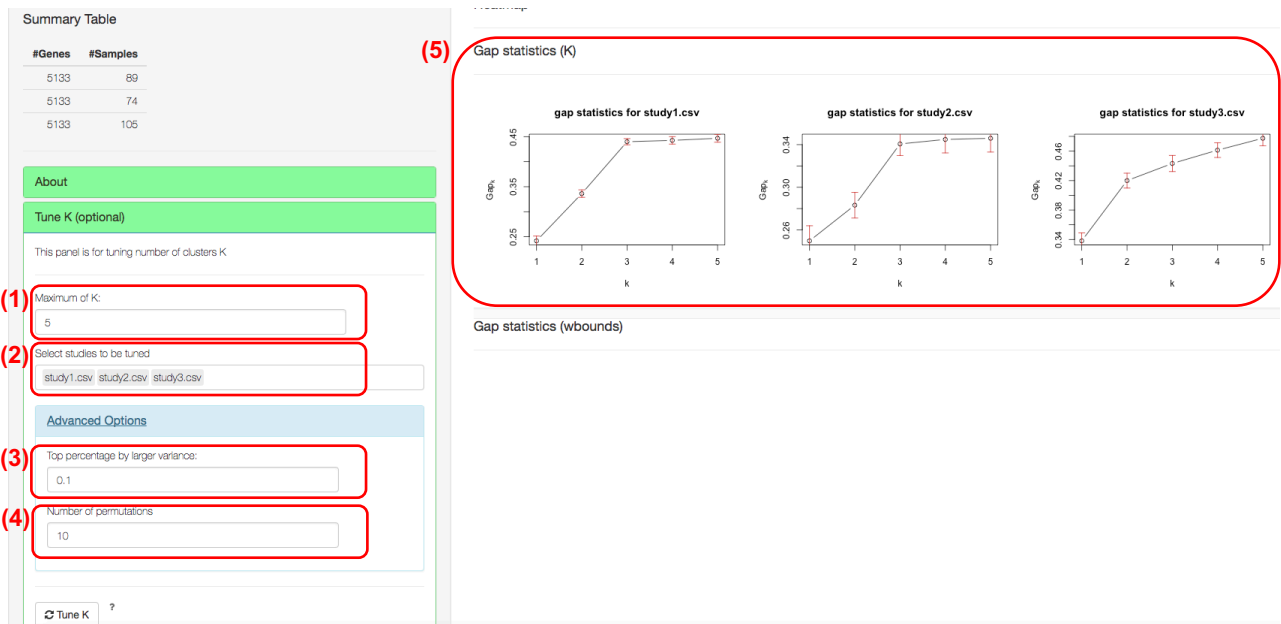


Figure 26: Tuning parameter selection for number of clusters

If the users are not sure what is the number of clusters, they can start to use the Tune K panel as in Figure 26. Gap statistics will be used to get optimal K for each individual study. Users need to specify maximum number of K (1), 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 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). Number of permutation 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 button "Tune K ", we will obtain gap statistics as in Figure 26. 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 metaClust. If the users are not sure what is a good Wbound, they can start to use the Tune Wbounds panel as in Figure 27.

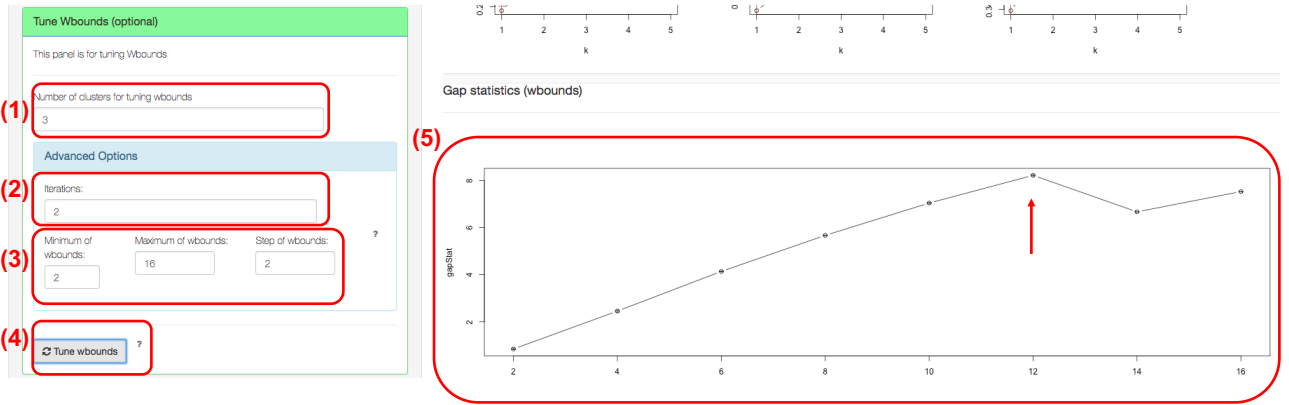


Figure 27: Wbound selection

Again, gap statistics will be used for tuning Wbounds. Users will specify 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 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, user can click on "Tune Wbounds" button (4). The results will be shown in Figure 27 (5). Wbound=12 is preferred since the corresponding gap statistics is maximized (where the red arrow indicates).

Step 3 Run MetaClust:

Under Run Meta Sparse K-Means panel, user can specify the number of clusters (1), Wbounds (2) and run MetaClust (5) as in Figure 28.

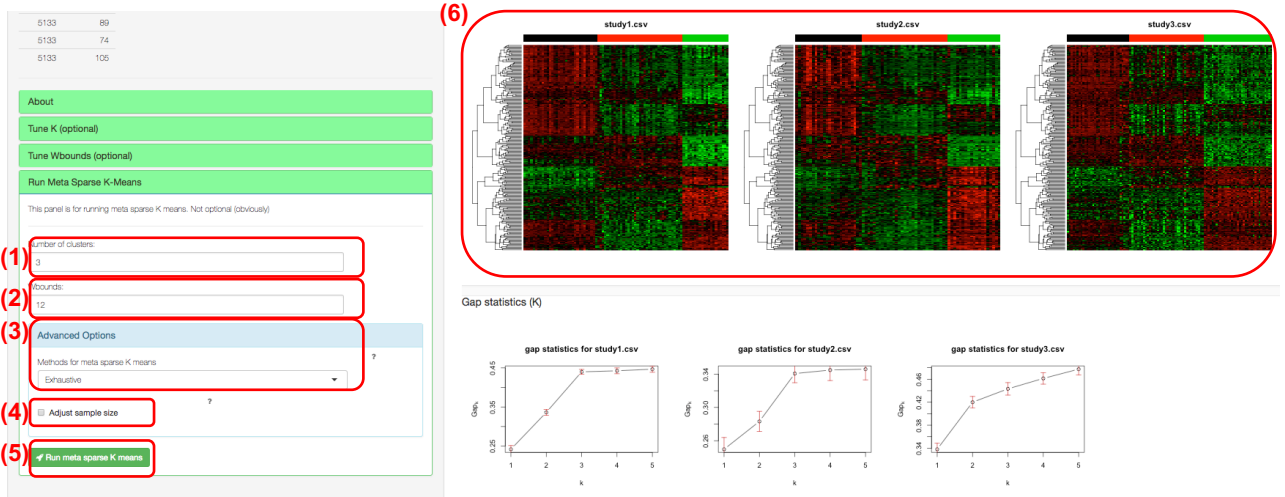


Figure 28: Result for MetaClust

In advanced options (for which users are not suggested to change if they are not familiar with the algorithm), There are three clustering matching methods (3): Exhaustive, linear, MCMC. 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 accurate results with very low probability. MCMC is suitable if many studies and clusters are provided. Adjust sample size checkbox (at position (5)) allows users to adjust sample size effect. After number of clusters and Wbounds are specified, users can click on Run meta sparse K means and obtain results as Figure 28.

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 be very time consuming.
- Adjust sample size: adjust sample size effect.
- Run meta sparse Kmeans: start tuning wbounds.

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 28 (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 on Toolsets tab and then choose MetaPCA, users are directed to MetaPCA home page as Figure 29. The R package for MetaPCA module can be found at <https://github.com/metaOmics/metaPCA>.

Figure 29: MetaPCA settings

5.7.1 Procedure

Step 1 Specify parameters

There are very few parameters need to be specified for MetaPCA, as in Figure 29. 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. SSC has better performance and is suggested. Dimension of meta-eigenvector matrix option (2) allows users to specify 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 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.

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 λ : stepsize of the searching space of λ .
- Tuning parameter for sparsity: Tuning parameter for sparsity that will be used for sparse MetaPCA.

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 30. For each study, only first two studies are visualized. The results show nice separations between three groups. These figures and eigenvectors are saved to the MetaPCA folder.

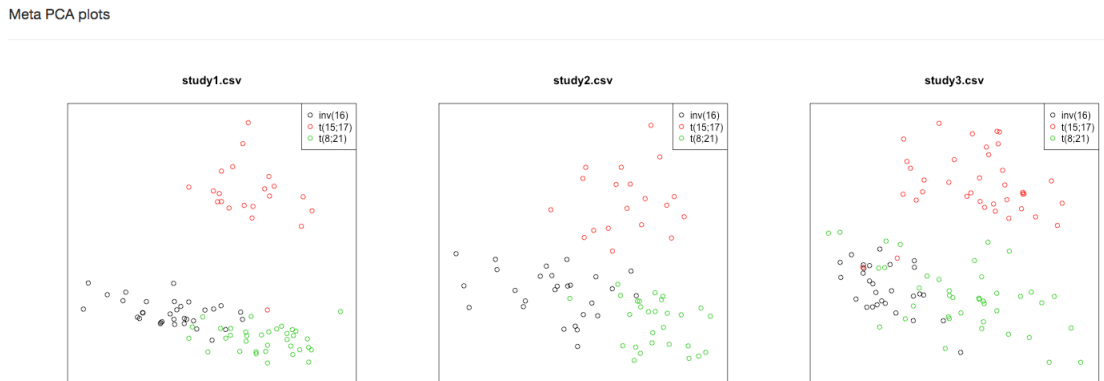


Figure 30: MetaPCA result

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*. *Accepted*.
- 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., Lumberras, E., Yeoh, A. E., Koay, E. S., Li, R., Lin, W., Williams, P. M., Wiecek, 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.

- 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.).
- 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: Etf4 gene fusions define a third molecular subtype of prostate cancer. *Cancer research*, 66(7):3396–3400.
- 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.