# A tutorial for metaOmic

## Contents

# 1 Introduction

MetaOmics is a GUI for meta-analysis implemented using R shiny. Current version includes MetaQC for quality control, MetaDE for differential expression analysis, MetaPath for pathway enrichment analysis, MetaClust for sparse clustering analysis, MetaPCA for principal component analysis, MetaKTSP for classification analysis, MetaDCN for differential co-expression network analysis, MetaLA for liquid association analysis.

In this tutorial, we will go through installation and usage step by step using real data examples.

The metaOmics suit software is publicly available at `https://github.com/metaOmic/metaOmics`. Individual R packages are also available on GitHub and the url will be introduced in each individual package section.

# 2 Preliminaries

## 2.1 Citing MetaOmics

MetaOmics implements many meta-analytic methodology by their authors. Please cite appropriate papers when you use result from MeteOmics suit, by which the authors will receive professional credit for their work.

- MetaOmics suit itself can be cited as:
- 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, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.
  - Li, J., Tseng, G. C., et al. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics*, 5(2A):994–1019.
  - 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.
  - and many more
- MetaPath:
  - Shen, K. and Tseng, G. C. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, 26(10):1316–1323.

  - Fang, Z., Zeng, X., Lin, C.-W., Ma, T., and Tseng, G. C. (2016). *Comparative Pathway Integrator: a framework of meta-analytic integration of multiple transcriptomic studies for consensual and differential pathway analysis*. PhD thesis, University of Pittsburgh.

- MetaClust: 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: Kim, S., Kang, D., Huo, Z., Park, Y., and Tseng, G. C. (submitted). Meta-analytic principal component analysis in integrative omics application.

- 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(March):btw115.

- 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*, page btw788.

## 2.2 Installation

The full instruction of how to install, start are available at `https://github.com/metaOmic/metaOmics`.

### 2.2.1 Requirement

- R >= 3.3.1
- Shiny >= 0.13.2

### 2.2.2 How to start the app

- First, clone the project
- git clone https://github.com/metaOmic/metaOmics
- in R (suppose the application directory is metaOmics),

  > install.packages('shiny')

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

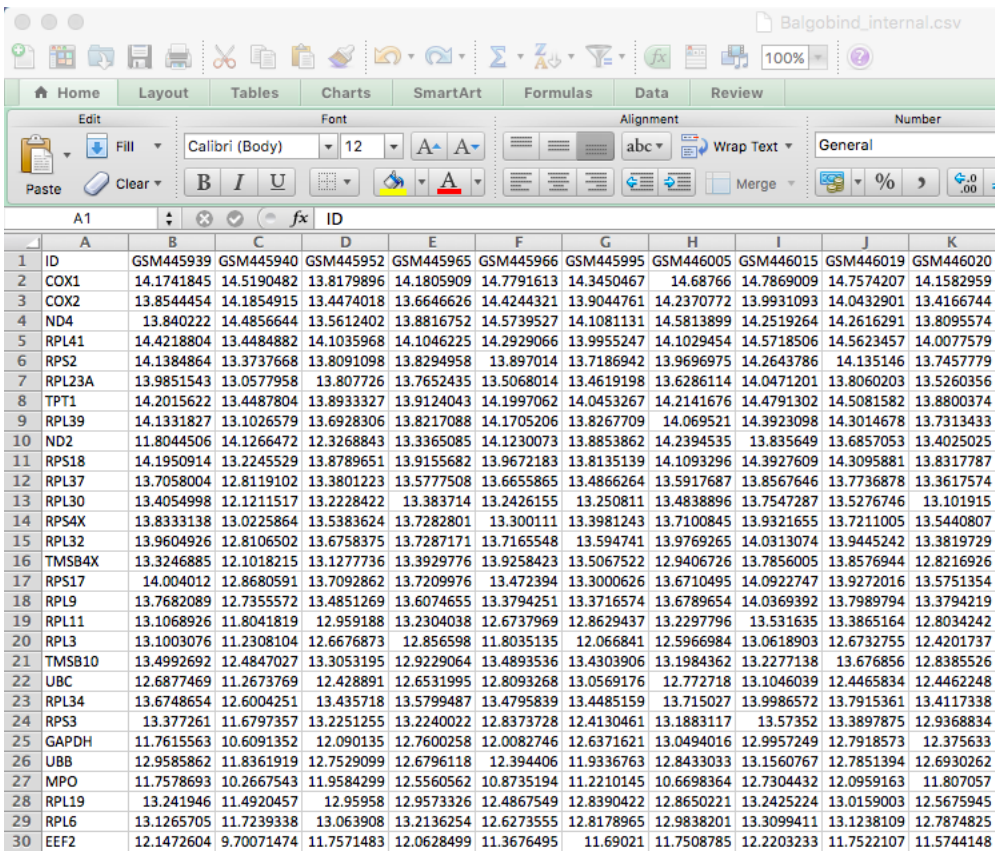
## 2.3 Question and bug report

Who should be responsible for maintaining the software?

# 3 Prepare data

## 3.1 Raw data

Data should be prepared as the example in Figure 1. First column should be feature ID (e.g. gene symbol) and the rest of the columns are samples. Note that the first column can also be other feature type (i.e. probe id, entrez ID). The first row is sample ID. Valid data type includes continuous, count.



| | 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 |
| 11 | RPS18 | 14.1950914 | 13.2245529 | 13.8789651 | 13.9155682 | 13.9672183 | 13.8135139 | 14.1093296 | 14.3927609 | 14.3095881 | 13.8317787 |
| 12 | RPL37 | 13.7058004 | 12.8119102 | 13.3801223 | 13.5777508 | 13.6655865 | 13.4866264 | 13.5917687 | 13.8567646 | 13.7736878 | 13.3617574 |
| 13 | RPL30 | 13.4054998 | 12.1211517 | 13.2228422 | 13.383714 | 13.2426155 | 13.250811 | 13.4838896 | 13.7547287 | 13.5276746 | 13.101915 |
| 14 | RPS4X | 13.8333138 | 13.0225864 | 13.5383624 | 13.7282801 | 13.300111 | 13.3981243 | 13.7100845 | 13.9321655 | 13.7211005 | 13.5440807 |
| 15 | RPL32 | 13.9604926 | 12.8106502 | 13.6758375 | 13.7287171 | 13.7165548 | 13.594741 | 13.9769265 | 14.0313074 | 13.9445242 | 13.3819729 |
| 16 | TMSB4X | 13.3246885 | 12.1018215 | 13.1277736 | 13.3929776 | 13.9258423 | 13.5067522 | 12.9406726 | 13.7856005 | 13.8576944 | 12.8216926 |
| 17 | RPS17 | 14.004012 | 12.8680591 | 13.7092862 | 13.7209976 | 13.472394 | 13.3000626 | 13.6710495 | 14.0922747 | 13.9272016 | 13.5751354 |
| 18 | RPL9 | 13.7682089 | 12.7355572 | 13.4851269 | 13.6074655 | 13.3794251 | 13.3716574 | 13.6789654 | 14.0369392 | 13.7989794 | 13.3794219 |
| 19 | RPL11 | 13.1068926 | 11.8041819 | 12.959188 | 13.2304038 | 12.6737969 | 12.8629437 | 13.2297796 | 13.531635 | 13.3865164 | 12.8034242 |
| 20 | RPL3 | 13.1003076 | 11.2308104 | 12.6676873 | 12.856598 | 11.8035135 | 12.066841 | 12.5966984 | 13.0618903 | 12.6732755 | 12.4201737 |
| 21 | TMSB10 | 13.4992692 | 12.4847027 | 13.3053195 | 12.9229064 | 13.4893536 | 13.4303906 | 13.1984362 | 13.2277138 | 13.676856 | 12.8385526 |
| 22 | UBC | 12.6877469 | 11.2673769 | 12.428891 | 12.6531995 | 12.8093268 | 13.0569176 | 12.772718 | 13.1046039 | 12.4465834 | 12.4462248 |
| 23 | RPL34 | 13.6748654 | 12.6004251 | 13.435718 | 13.5799487 | 13.4795839 | 13.4485159 | 13.715027 | 13.9986572 | 13.7915361 | 13.4117338 |
| 24 | RPS3 | 13.377261 | 11.6797357 | 13.2251255 | 13.2240022 | 12.8373728 | 12.4130461 | 13.1883117 | 13.57352 | 13.3897875 | 12.9368834 |
| 25 | GAPDH | 11.7615563 | 10.6091352 | 12.090135 | 12.7600258 | 12.0082746 | 12.6371621 | 13.0494016 | 12.9957249 | 12.7918573 | 12.375633 |
| 26 | UBB | 12.9585862 | 11.8361919 | 12.7529099 | 12.6796118 | 12.394406 | 11.9336763 | 12.8433033 | 13.1560767 | 12.7851394 | 12.6930262 |
| 27 | MPO | 11.7578693 | 10.2667543 | 11.9584299 | 12.5560562 | 10.8735194 | 11.2210145 | 10.6698364 | 12.7304432 | 12.0959163 | 11.807057 |
| 28 | RPL19 | 13.241946 | 11.4920457 | 12.95958 | 12.9573326 | 12.4867549 | 12.8390422 | 12.8650221 | 13.2425224 | 13.0159003 | 12.5675945 |
| 29 | RPL6 | 13.1265705 | 11.7239338 | 13.063908 | 13.2136254 | 12.6273555 | 12.8178965 | 12.9838201 | 13.3099411 | 13.1238109 | 12.7874825 |
| 30 | EEF2 | 12.1472604 | 9.70071474 | 11.7571483 | 12.0628499 | 11.3676495 | 11.69021 | 11.7508785 | 12.2203233 | 11.7522107 | 11.5744148 |

Figure 1: A example data format

## 3.2 Clinical data

Clinical data should be prepared as the example in Figure 2. First column should be sample ID and each row represents a sample. The rest of the columns are clinical information.

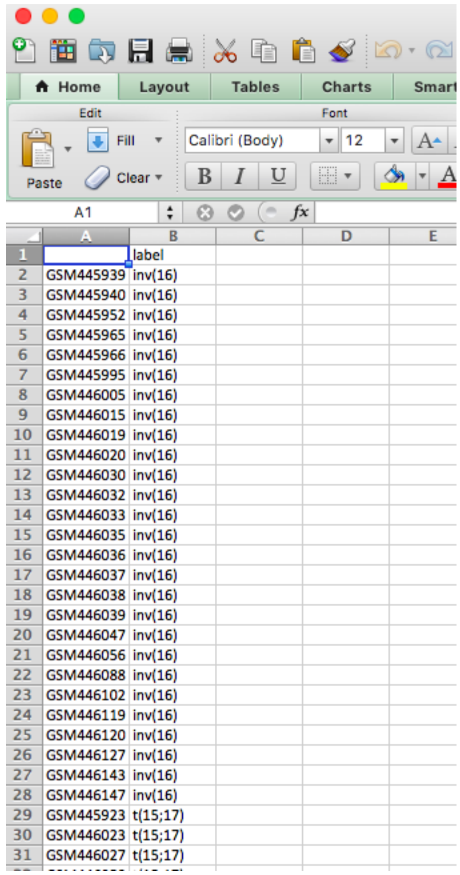



Figure 2: A example clinical data format

# 4 Toolsets

After starting metaOmics, the first page is the metaOmics setting page in Figure 3. There are 4 tabs on top of the page (at position (1)): Setting, Preprocessing, Saved Data and Toolsets. Below the 4 tabs, the first header is the session information. Why do we need session information? The second header is Directory for Saving Output Files (at position (2)). By clicking . . ., user can set default working directory, in which all the meta-analysis results will be saved. User can view their current working directory on the top right corner (at position (3)). The third header is Toolsets (at position (4)), here users can view if individual packages are installed. If the packages are installed, there is a checked installed status. Otherwise, users can install individual package by clicking install blue button. Position (5) shows the current active dataset, which will be introduced in Section 4.1.1 **Step 4:**

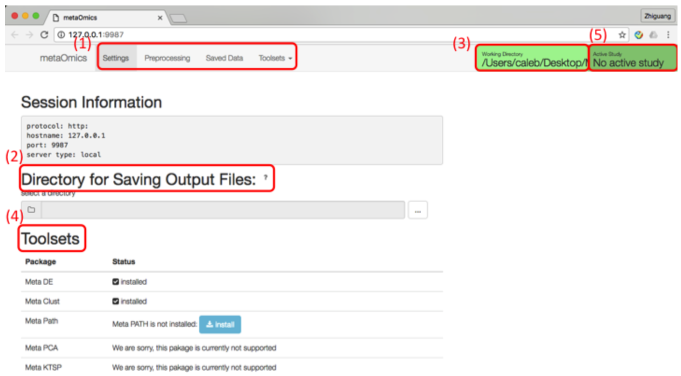



Figure 3: GUI setting page

## 4.1 Preprocessing

In this subsection, we will introduce how to upload your dataset into the MetaOmics suit such that the functional modules can process the uploaded datasets.

### 4.1.1 Procedure

**Step 1: Uploading data:**

If users go to the Preprocessing tab as in Figure 4, they are able to upload genomic data via the tab "Choosing/Upload Expression Data" as in Figure 5 (at position (1)). The data should be prepared according to Section 3. Users may optionally upload Clinical Data (at position (2)), depending on biological purpose. The all MetaOmics modules except for MetaClust require external clinical labels. The MetaOmics suit also provides handlers (at position (3)) for feature annotation, missing value imputation and multiple probe same genes. After uploading is complete, users can preview their data on the right hand side of the page as Figure 5.

**Step 2: Preprocessing:**

There are several expression data parsing option available on the left panel of Figure 5. A complete introduction of these options are available at the end of this subsection. The right hand side of Figure 5 shows the summary statistics of uploaded data and preview of the data matrix. There is a search box such that the user can search their favorite genes.

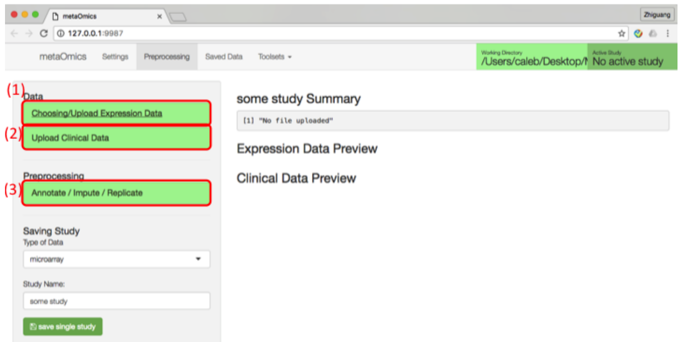



Figure 4: GUI Preprocessing page

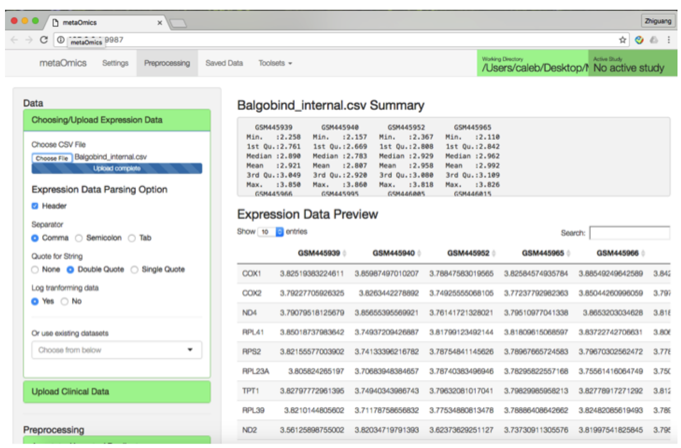



Figure 5: GUI Preprocessing page

After users upload clinical data (e.g. case control labels) and specify type of data and study name. They can click "save single study" button, single study will be saved.

**Step 3: Saved Data:**

After uploading multiple studies w/o clinical data, Users can turn to the

Saved Data tab. Users should select multiple datasets as Figure 6 (at position (2)).

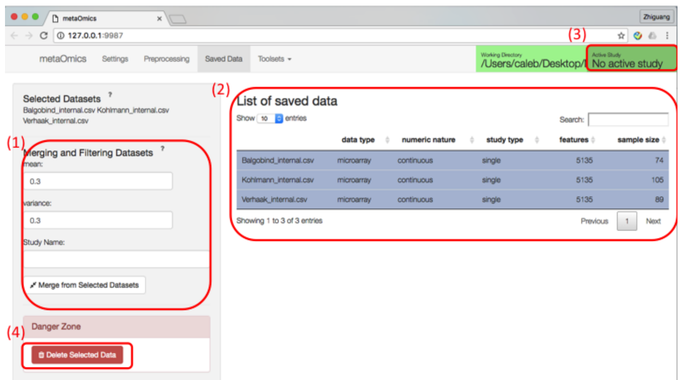



Figure 6: GUI Preprocessing page

Users can select filtering criteria, enter merged study name and click on the Merge from Selected Datasets (at position (1)). A merged dataset will appear on the "List of saved data" panel (at position (2)).

**Step 4: Make merged Dataset Active:**

The last thing users need to do before using meta-analytic toolsets is select merged data and click on "Make merged Active Dataset" - A big green button (at position (1)). Then the merged data becomes active study and shows up on the top right corner (at position (4)). The active dataset serves as the input for all other MetaOmics modules. If users want to delete a dataset, just click "Delete Selected Data" button (at position (3)) after selection the dataset.

**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, if the same gene symbol maps to multiple probes, the probe with the largest inner quantile range (IQR) will be selected.

3. Saved Data, Merging and Filtering Datasets:

    - Mean: the criteria such that how many percent of genes will be filtered out based on sum of mean ranks (e.g. 0.3 represent 30%).
    - Variance: after the Mean filtering, the criteria such that how many percent of genes will be filtered out based on sum of variance ranks (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 delete permanently if clicked, so please be cautious.

## 4.2 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 the Manuscript. 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. Detailed information can be found in the "MetaQC" package in the metaOmics software suite (`https://github.com/metaOmic/MetaQC`). The test data used to demo the "MetaQC" package here is merged from 8 prostate cancer studies, the details of these studies can be found in (cite MetaQC paper).

### 4.2.1 Procedure


Figure 7: "MetaQC" options

There are four main options available for the "MetaQC" package as shown in Figure 7. Users need to specify whether to (1) 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 like in merging step. In the demo example, the merged data have already had gene filtering, no further filtering is performed. Next, users need to specify (2) the approach (either by raw p-value or FDR) and cutoff to select potentially DE genes and enriched pathways needed in the computation of EQC, AQCg, CQCg, AQCp and CQCp. (3) "Advanced options" is optional and users are suggested not to modify the option setting in this section. 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. Once all the above options are specified, users can click on (4) "Run MetaQC Analysis" to implement the tool.

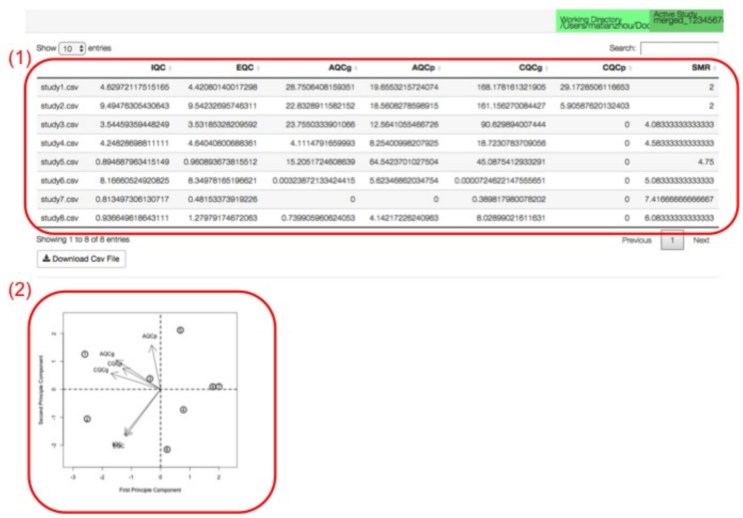
**Complete List of Options:**

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

#### 4.2.2 Results




| | IQC | EQC | AQCg | AQCp | CQCg | CQCp | SMR |
|---|---|---|---|---|---|---|---|
| study1.csv | 4.62972117515165 | 4.42080140017298 | 28.7506408159351 | 19.6553215724074 | 168.178161321905 | 29.1728506116653 | 2 |
| study2.csv | 9.49476305430643 | 9.54232695746311 | 22.8328911582152 | 18.5608278598915 | 161.156270084427 | 5.90587620132403 | 2 |
| study3.csv | 3.54459359448249 | 3.53185328209592 | 23.7550333901066 | 12.5641055466726 | 90.629894007444 | 0 | 4.08333333333333 |
| study4.csv | 4.24828698811111 | 4.64040800688361 | 4.1114791659993 | 8.25400998207925 | 18.7230783709056 | 0 | 4.58333333333333 |
| study5.csv | 0.894687963415149 | 0.960893673815512 | 15.2051724608639 | 64.5423701027504 | 45.0875412933291 | 0 | 4.75 |
| study6.csv | 8.16660524920825 | 8.34978165196621 | 0.00323872133424415 | 5.62346862034754 | 0.000072462214755651 | 0 | 5.08333333333333 |
| study7.csv | 0.813497306130717 | 0.48153373919226 | 0 | 0 | 0.389817980078202 | 0 | 7.41666666666667 |
| study8.csv | 0.936649618643111 | 1.27979174672063 | 0.739905960624053 | 4.14217226240963 | 8.02899021611631 | 0 | 6.08333333333333 |

Figure 8: "MetaQC" Results

As shown in Figure 8, 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.

### 4.3 MetaDE

MetaDE package implements 12 major meta-analysis methods for differential expression analysis falling into 3 main categories: combining p-values, combining effect sizes and others (e.g. combining ranks, etc.). Depending on the types of outcome, the package can perform two class comparison, multi-class comparison, association with continuous or survival outcome. The package allows the input of either microarray (continuous intensity) or RNA-seq data (count) for individual study analysis.

#### 4.3.1 Procedure

There are two major steps to implement the package: meta differential analysis and pathway analysis. As shown in Figure 9, there are 9 major options that need to be specified to implement the package: (1) - (6) are for the first step and (7) - (9) are for the second step. A detailed list of all options available for the package can be found at the end of this subsection. Individual MetaDE package is also available on GitHub at `https://github.com/metaOmic/MetaDE`.

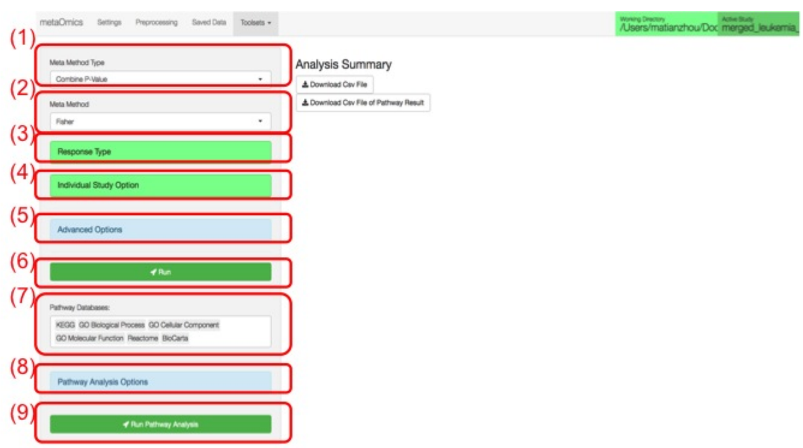



Figure 9: “MetaDE” options

**Step 1. Meta differential analysis:** This step includes the core strategies of the “MetaDE” package. Users first need to specify (1) “Meta Method Type” and (2) “Meta Method” correspondingly. There are three types to select from: combining p-value, combining effect size and others. “Fisher” and “AW-Fisher” meta methods are available for p-value combination, “Fixed Effect Model (FEM)” and “Random Effect Model (REM)” for effect size combination, and the other methods in the “Others” type. More meta-analysis methods are available if “complete option” is chosen from (5) “Advanced Options” section. Next, we need to specify the outcome of interest in (3) “Response Type”. For example, for differential expression analysis, two-class comparison is usually chosen. For two-class comparison, users need to specify the class label, and the level corresponding to the experimental and the control groups. Other outcome types such as continuous or survival data can also be chosen. In (4) "Individual study option", users can specify whether each of the study is a paired design, and for p-value combination method, one can select the differential analysis method to obtain p-values in each individual study (e.g. generally suggest LIMMA for microarray and edgeR for RNA-seq). "Advanced Options" is optional and users are suggested not to modify the option setting in this section. Once all the above options are specified, users can click on (6) "Run" to implement the first step.

**Step 2. Pathway analysis:** This step consists of a downstream pathway analysis for the meta differential analysis results from the first step. Users can select from 25 available pathway databases (7) to perform the pathway enrichment analysis. There are three main options for pathway analysis under (8) "Pathway Analysis Option": the enrichment method including the Fisher's exact test and KS test, the minimum as well as the 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: either by p-value cutoff or by number of top ranked genes. Once these options are set, users can click on (9) "Run Pathway Analysis" to implement the first step.

**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. 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.
4. Individual Study Option:
   - Setting individual study method
   - Setting individual study paired option
5. Advanced Option (**Optional):
   - Use complete options
   - Parametric
   - Covariate

- Alternative hypothesis

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

#### 4.3.2 Results

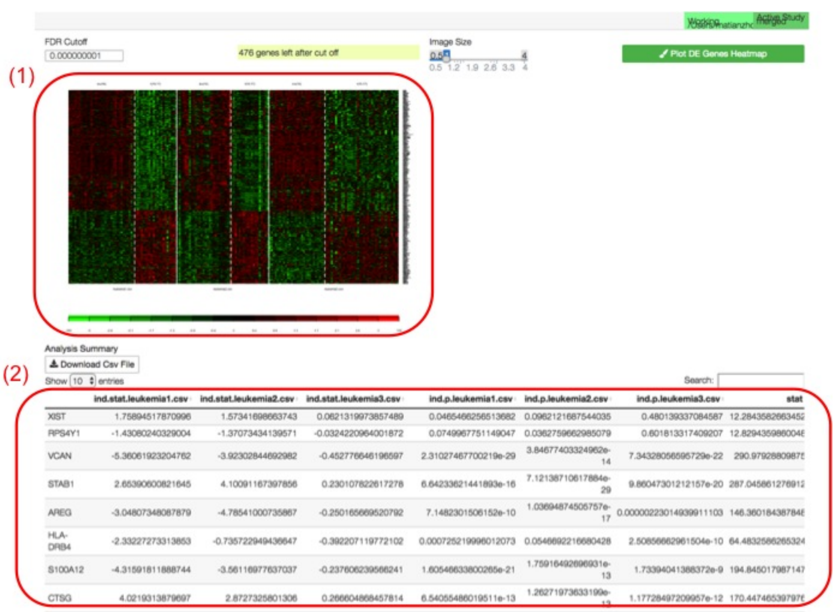
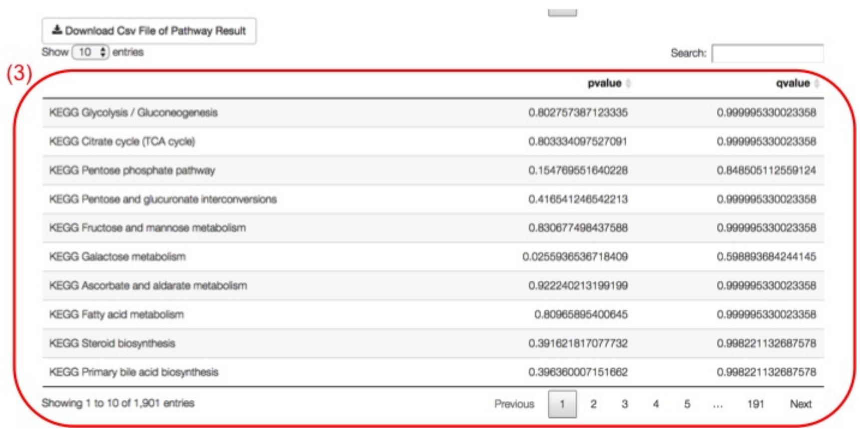



Figure 10: "MetaDE" Results (1)

Two main outputs from the first "meta differential analysis" step in the procedure are shown in Figure 10. The first is (2) a summary of meta analysis results, including information of individual test statistics, individual study p-value, meta-analysis p-value, FDR, etc. The second output is (1) a heatmap of DE genes drawn 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 sepate 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.

The (2) summary table might differ slightly for different meta-analysis methods, for example, AW-Fisher method will include additional columns of study-specific weights.




| | pvalue | qvalue |
|---|---|---|
| KEGG Glycolysis / Gluconeogenesis | 0.802757387123335 | 0.999995330023358 |
| KEGG Citrate cycle (TCA cycle) | 0.803334097527091 | 0.999995330023358 |
| KEGG Pentose phosphate pathway | 0.154769551640228 | 0.848505112559124 |
| KEGG Pentose and glucuronate interconversions | 0.416541246542213 | 0.999995330023358 |
| KEGG Fructose and mannose metabolism | 0.830677498437588 | 0.999995330023358 |
| KEGG Galactose metabolism | 0.0255936536718409 | 0.598893684244145 |
| KEGG Ascorbate and aldarate metabolism | 0.922240213199199 | 0.999995330023358 |
| KEGG Fatty acid metabolism | 0.80965895400645 | 0.999995330023358 |
| KEGG Steroid biosynthesis | 0.391621817077732 | 0.998221132687578 |
| KEGG Primary bile acid biosynthesis | 0.396360007151662 | 0.998221132687578 |

Figure 11: "MetaDE" Results (2)

For the second step "pathway analysis", there is (3) a tabular summary outputted, as shown in Figure 11. 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 two tabular results to the working directory by clicking on "Download Csv File" on the top left of the summary table.

## 4.4 MetaPath

Following the detection of biomarkers, pathway analysis (a.k.a. gene set enrichment analysis) is usually performed for functional annotation and biological interpretation. When there are multiple studies available on a related hypothesis, meta-analysis methods are necessary for joint pathway analysis. Two major approaches have been included in the MetaPath package to serve for this purpose: Comparative Pathway Integrator (CPI) and Meta-Analysis for Pathway Enrichment (MAPE) (Shen et al., 2010; Fang et al., 2017). Pathway clustering with statistically valid text mining is included in the package to reduce pathway redundancy to condense knowledge and increase interpretability of clustering results.

### 4.4.1 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 12, there are 8 major options that need to be specified to implement the package: (1) - (6) are for the first step, (7) for the second step and (8) for the third step. A detailed list of all options available for the package can be found at the end of this subsection. Individual MetaPath package is also available on GitHub at `https://github.com/metaOmic/MetaPath`.

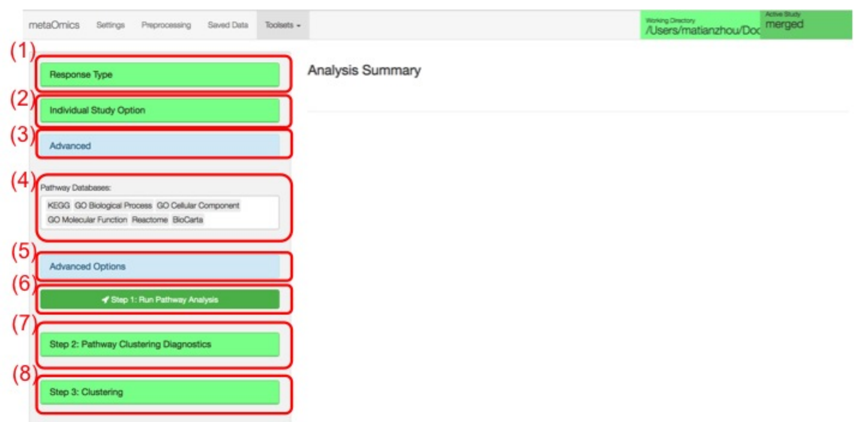



Figure 12: "MetaPath" options

**Step 1. Pathway analysis:** This step consists of a meta pathway analysis. Users need to specify (1) "Response type", (2) "Individual study option" and (3) "Advanced" as in MetaDE to perform the pathway enrichment analysis in the presence of multiple studies. Users can select from 25 available pathway databases (4) for the enrichment analysis. (5) "Advanced Options" is optional and users are suggested not to modify the option setting in this section. By default, the "CPI" approach is used, otherwise "MAPE" approach can also be used. Other options include pathway enrichment method (the Fisher's exact test or KS test), the minimum as well as the 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. Once these options are set, users can click on (6) "Run Pathway Analysis" to implement the first step.

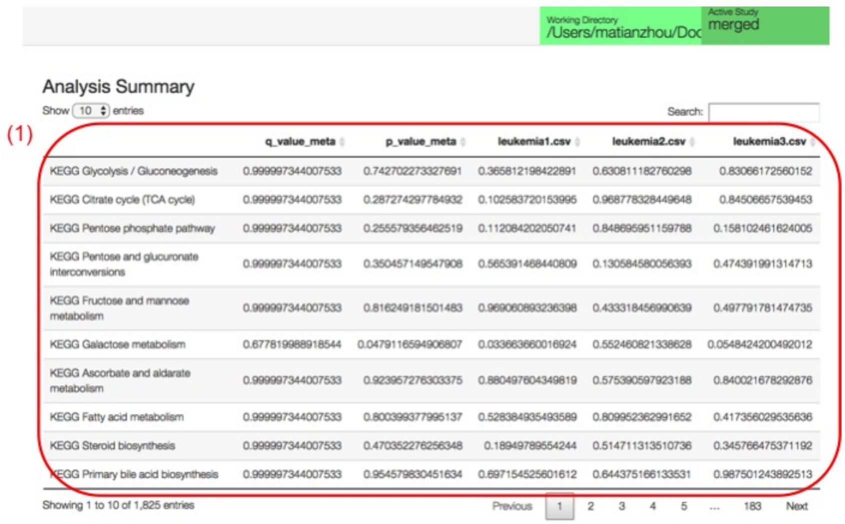
**Step 2. Pathway clustering diagnostics:** From the first step, users can choose the top enriched pathways for further clustering. One can expand the drop-down menu and use FDR cutoff to choose top pathways and click on (7) "Pathway clustering diagnostics" to implement the second step.

**Step 3. Pathway clustering with text mining:** From the second step, 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 (8) "Get clustering result" to implement the third step. Note that you may not want to select too large a 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. 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.
2. Individual Study Option:
    - Setting individual study method
    - Setting individual study paired option
3. Advanced Option (**Optional):
    - Covariate
    - Alternative hypothesis
4. Pathway Databases
5. Pathway Analysis Option:
    - Software
    - Pathway enrichment method
    - Pathway min gene size
    - Pathway max gene size
6. Step1: Run Pathway Analysis
7. Step2: Pathway Clustering Diagnostics
8. Step3: Get Clustering Result

### 4.4.2 Results




| | q_value_meta | p_value_meta | leukemia1.csv | leukemia2.csv | leukemia3.csv |
|---|---|---|---|---|---|
| KEGG Glycolysis / Gluconeogenesis | 0.999997344007533 | 0.742702273327691 | 0.365812198422891 | 0.630811182760298 | 0.83066172560152 |
| KEGG Citrate cycle (TCA cycle) | 0.999997344007533 | 0.287274297784932 | 0.102583720153995 | 0.968778328449648 | 0.84506657539453 |
| KEGG Pentose phosphate pathway | 0.999997344007533 | 0.255579356462519 | 0.112084202050741 | 0.848695951159788 | 0.158102461624005 |
| KEGG Pentose and glucuronate interconversions | 0.999997344007533 | 0.350457149547908 | 0.565391468440809 | 0.130584580056393 | 0.474391991314713 |
| KEGG Fructose and mannose metabolism | 0.999997344007533 | 0.816249181501483 | 0.969060893236398 | 0.433318456990639 | 0.497791781474735 |
| KEGG Galactose metabolism | 0.677819988918544 | 0.0479116594906807 | 0.033663660016924 | 0.552460821338628 | 0.0548424200492012 |
| KEGG Ascorbate and aldarate metabolism | 0.999997344007533 | 0.923957276303375 | 0.880497604349819 | 0.575390597923188 | 0.840021678292876 |
| KEGG Fatty acid metabolism | 0.999997344007533 | 0.800399377995137 | 0.528384935493589 | 0.809952362991652 | 0.417356029535636 |
| KEGG Steroid biosynthesis | 0.999997344007533 | 0.470352276256348 | 0.18949789554244 | 0.514711313510736 | 0.345766475371192 |
| KEGG Primary bile acid biosynthesis | 0.999997344007533 | 0.954579830451634 | 0.697154525601612 | 0.644375166133531 | 0.987501243892513 |

Showing 1 to 10 of 1,825 entries

Previous 1 2 3 4 5 … 183 Next

Figure 13: “MetaPath” Results (1)

After the first step is finished, (1) a summary table was generated as shown in Figure 13 (based on the default 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.

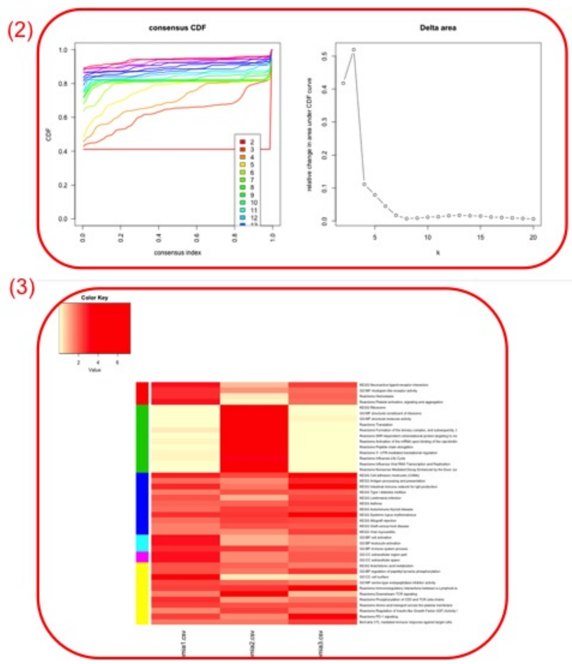
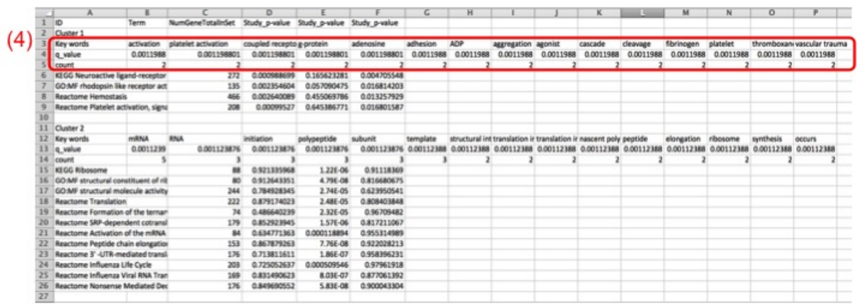



Figure 14: “MetaPath” Results (2)

After the “Pathway Cluster Diagnostics” step is finished, we will see (2) two plots generated on the right panel (Figure 14): 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, thus 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 the determine K at which there is no appreciable increase in CDF. 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, $K = 5$ have large enough CDF, is thus chosen (though $K = 7$ captures more, we only have 43 pathways here).



(4)

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | ID | Term | NumGeneTotalInSet | Study_p-value | Study_p-value | Study_p-value |
| 2 | Cluster 1 | | | | | |
| 3 | Key words | activation | platelet activation | coupled recepto | g-protein | adenosine |
| 4 | q_value | 0.0011988 | 0.001198801 | 0.001198801 | 0.001198801 | 0.001198801 |
| 5 | count | 2 | 2 | 2 | 2 | 2 |
| 6 | KEGG Neuroactive ligand-receptor | | 272 | 0.000988699 | 0.165623281 | 0.004705548 |
| 7 | GO:MF rhodopsin like receptor act | | 135 | 0.002354604 | 0.057090475 | 0.016814203 |
| 8 | Reactome Hemostasis | | 466 | 0.002640089 | 0.455069786 | 0.013257929 |
| 9 | Reactome Platelet activation, signa | | 208 | 0.00099527 | 0.645386771 | 0.016801587 |
| 10 | | | | | | |
| 11 | Cluster 2 | | | | | |
| 12 | Key words | mRNA | RNA | initiation | polypeptide | subunit |
| 13 | q_value | 0.0011239 | 0.001123876 | 0.001123876 | 0.001123876 | 0.001123876 |
| 14 | count | 5 | 3 | 3 | 3 | 3 |
| 15 | KEGG Ribosome | | 88 | 0.921335968 | 1.22E-06 | 0.91118369 |
| 16 | GO:MF structural constituent of rit | | 80 | 0.912643351 | 4.79E-08 | 0.816680675 |
| 17 | GO:MF structural molecule activity | | 244 | 0.784928345 | 2.74E-05 | 0.623950541 |
| 18 | Reactome Translation | | 222 | 0.879174023 | 2.48E-05 | 0.808403848 |
| 19 | Reactome Formation of the ternar | | 74 | 0.486640239 | 2.32E-05 | 0.96709482 |
| 20 | Reactome SRP-dependent cotransl | | 179 | 0.852923945 | 1.57E-06 | 0.817211067 |
| 21 | Reactome Activation of the mRNA | | 84 | 0.634771363 | 0.000118894 | 0.955314989 |
| 22 | Reactome Peptide chain elongatio | | 153 | 0.867879263 | 7.76E-08 | 0.922028213 |
| 23 | Reactome 3' -UTR-mediated transl | | 176 | 0.713811611 | 1.86E-07 | 0.958396231 |
| 24 | Reactome Influenza Life Cycle | | 203 | 0.725052637 | 0.000509546 | 0.97961918 |
| 25 | Reactome Influenza Viral RNA Tran | | 169 | 0.831490623 | 8.03E-07 | 0.877061392 |
| 26 | Reactome Nonsense Mediated Dec | | 176 | 0.849690552 | 5.83E-08 | 0.900043304 |
| 27 | | | | | | |

| | G | H | I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | adhesion | ADP | aggregation | agonist | cascade | cleavage | fibrinogen | platelet | thromboxan | vascular trauma |
| 4 | 0.0011988 | 0.0011988 | 0.0011988 | 0.0011988 | 0.0011988 | 0.0011988 | 0.0011988 | 0.0011988 | 0.0011988 | 0.0011988 |
| 5 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 12 | template | structural int | translation ir | translation ir | nascent poly | peptide | elongation | ribosome | synthesis | occurs |
| 13 | 0.00112388 | 0.00112388 | 0.00112388 | 0.00112388 | 0.00112388 | 0.00112388 | 0.00112388 | 0.00112388 | 0.00112388 | 0.00112388 |
| 14 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

Figure 15: “MetaPath” Results (3)

The heatmap in (3) shows the -log10 transformed p-value of enrichment analysis in each study from step 1. Studies are on columns and the selected pathways are on rows, red means more enriched. The pathways are sorted by the pathway cluster as indicated by the colors on the left side of the heatmap. In addition, one file named “Clustering_Summary.csv” is saved to the working directory and shows (4) a summary of the text mining algorithm. The most frequently appearing and enriched keywords of each cluster is highlighted in (4). All the results shown in the Browser is also automatically saved to the working directory.

## 4.5 MetaClust

By clicking toolsets and then metaClust, users are directed to metaClust home page as Figure 16. MetaClust (Huo et al., 2016) aims to perform sample clustering analysis combining multiple transcriptomic studies. By integrate information from multiple studies of similar biological purposes, MetaClust can identify an unified intrinsic gene sets among all studies, perform weighted clustering analysis using these common intrinsic gene sets, match the clustering pattern across studies to define disease subtype/cluster type. The resulting clustering from meta-analysis is more robust and accurate than single study analysis.

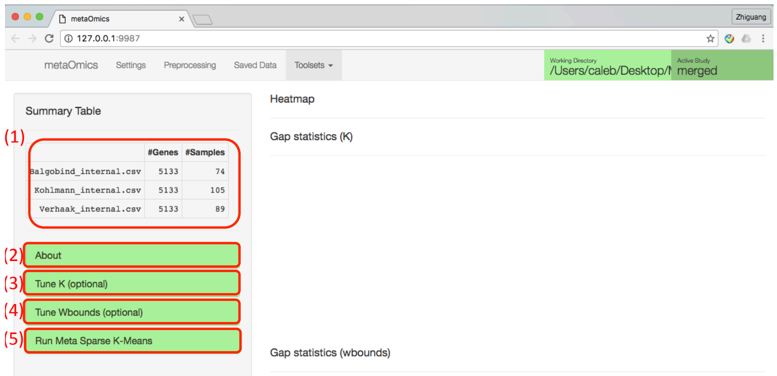



Figure 16: MetaClust home page

### 4.5.1 Procedure

Figure 16 shows the home page of MetaClust. On the top left panel users can see data summary Table (at position (1)). Below there are 4 tabs. About tab (at position (2)) includes basic introduction of metaClust. Starting with multiple studies, we could run MetaSparseKmeans (at position (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 (at position (3)) and Tune Wbounds (at position (4)) panel.

**Step 1: Tune K:**

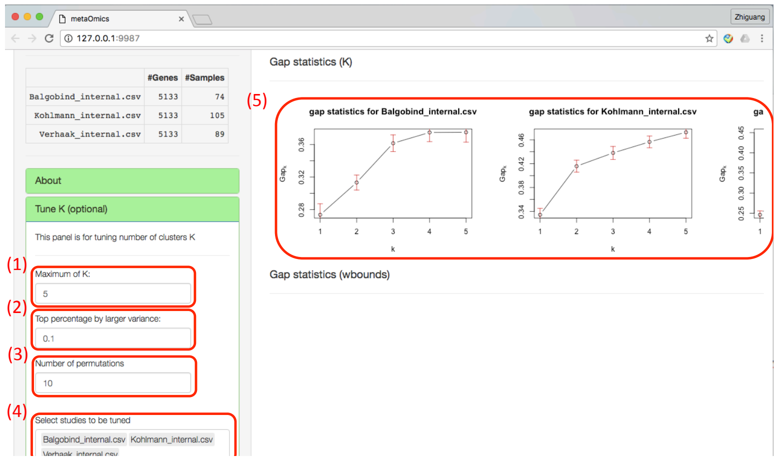



Figure 17: Tuning parameter selection for number of clusters

If the users are not sure what is number of clusters, they can start to use the Tune K panel as in Figure 17. Gap statistics will be used to get optimal K for each individual study. Users need to specify maximum number of K (at position (1)), which the algorithm will search number of studies from 1 to K. Top percentage p% by larger variance means that we will use top p% larger variance genes to perform gap statistics (at position (2)). Number of permutation is number of bootstrap samples for gap statistics (at position (3)). At least 50 bootstrap samples are suggested for a stable result of number of clusters. Studies to be tuned can be selected (at position (4)). By clicking button "Tune K", we will obtain gap statistics as in Figure 17. A good K is selected such that the $\text{Gap}_k$ is maximized or stabilized across all studies. From the figure, K=3 is preferred.

**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 18.

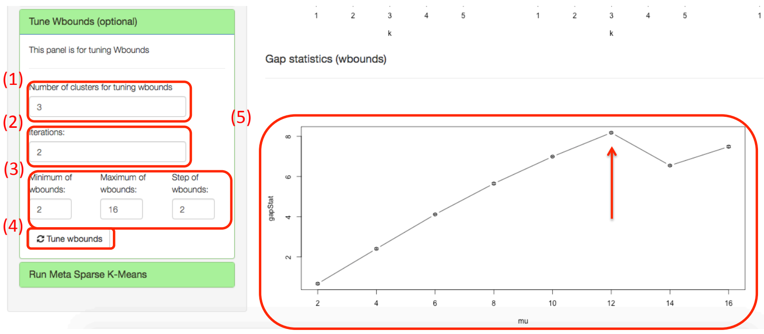



Figure 18: Wbound selection

Again, gap statistics will be used for tuning Wbounds. Users will specify number of clusters for tuning Wbounds (at position (1)), which could be obtained from the previous step. Iterations (at position (2)) is the same thing as 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 (at position (3)). After all these steps are set, user can click on "Tune Wbounds" button (at position (4)). The results will be shown in Figure 18 position (5). Wbound=12 is preferred since the corresponding gap statistics is maximized (where the red arrow indicates).

**Step 3: Run Meta Sparse K-Means:**

Under Run Meta Sparse K-Means panel, user can specify number of clusters (at position (1)), Wbounds (at position (2)) and run meta sparse K means (at position (5)), as in Figure 19.

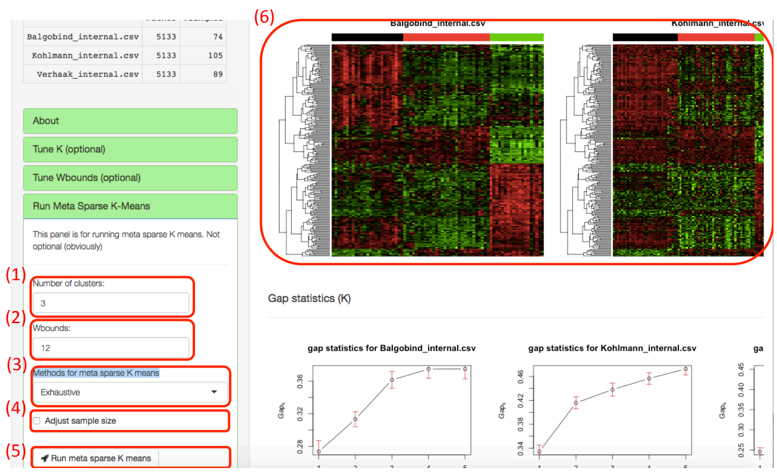


Figure 19: Result for MetaClust

There are three clustering matching methods (at position (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 accuracy. MCMC might by very time consuming. 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 19.

**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 meta sparse Kmeans: 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.

#### 4.5.2 Results

The result is shown in Figure 19 at position (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 metaOmics folder.

### 4.6 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 toolsets and then metaPCA, users are directed to metaPCA home page as Figure 20. On the top left panel users can see data summary Table. Below there are several options.

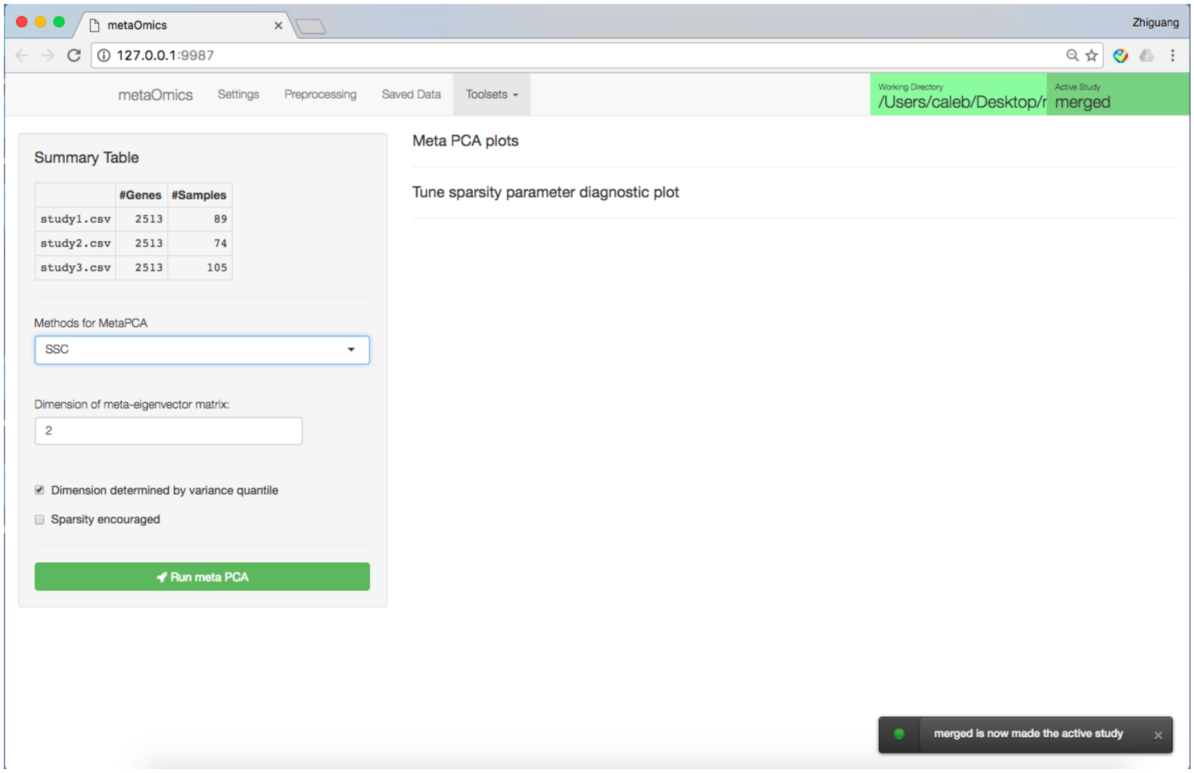



Figure 20: GUI Preprocessing page

### 4.6.1 Procedure

### 4.6.2 Methods for MetaPCA

- MetaPCA via sum of variance decomposition (SV)

  Let $X^{(m)}$ be an observed $p \times n^{(m)}$ data matrix of sample size $n^{(m)}$ and $p$ features for study $m$ $(1 \leq m \leq M)$. Denote by $S^{(m)}$ the maximum likelihood (ML) estimate of the $p \times p$ covariance matrix $\Omega^{(m)}$ of $X^{(m)}$. MetaPCA via sum of variance decomposition (SV) aims to solve the following eigen-value decomposition problem.

$$T^{SV} = \sum_{m=1}^{M} w^{(m)} S^{(m)}, \tag{1}$$

  where $w^{(m)}$ is the reciprocal of the largest eigenvalue of $S^{(m)}$. The common principal components $L$ are calculated from the eigen-decomposition of $T^{SV}$ : $L^T(T^{SV})L = \Lambda$ and $K$ top common PCs should be retained for down-stream analysis. Selection of the optimal $K$ will be described later in the the section of Parameter selection.

- MetaPCA via sum of squared cosine (SSC) maximization.

  the second MetaPCA framework motived by SSC criterion proceeds as below. The top $j^{(m)}$ eigenvectors are calculated from study $m$ to form eigenvector matrix $V^{(m)}$. We then perform eigen-decomposition on $T^{\mathrm{SSC}} = \sum_{m=1}^{M} V^{(m)} V^{(m)^T}$ and select the top $K$ eigenvectors to form the meta-analytic common eigen-space:

$$\Big( \sum_{m=1}^{M} V^{(m)} V^{(m)^T} \Big) B^{SSC} = \Lambda^* B^{SSC} \tag{2}$$

  where $V^{(m)}$ is a matrix consisting of $j^{(m)}$ leading eigenvectors, $\Lambda^*$ is a diagonal eigenvalue matrix, and $B^{\mathrm{SSC}} = (\beta_1^{\mathrm{SSC}}, \ldots, \beta_K^{\mathrm{SSC}})$ contains the top $K$ eigenvectors.

#### 4.6.3 Dimension of meta-eigenvector matrix

Dimension of meta-eigenvector matrix option allows user to specify dimension of the output meta-eigenvector matrix.

#### 4.6.4 Dimension determined by variance quantile

Logical value whether dimension size of each study's eigenvector matrix (SSC) is determined by the pre-defined level of variance quantile 80%.

#### 4.6.5 Sparsity encouraged

If the Sparsity encouraged checkbox is selected, we are able to tune the best tuning parameter $\lambda$ and perform sparse metaPCA. After clicking on search for optimal tuning parameter button, the optimum tuning parameter will be returned to the box "tuning parameter for sparsity"

#### 4.6.6 Run meta PCA

If Sparsity encouraged checkbox is selected, sparse meta PCA will be performed. Otherwise, meta PCA will be performed. The result is shown in the following figures.

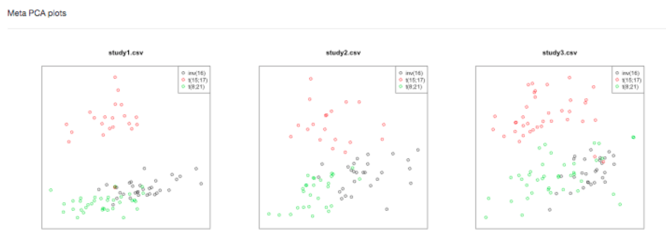



Figure 21: GUI Preprocessing page

# References

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.

Fang, Z., Zeng, X., Lin, C.-W., Ma, T., and Tseng, G. C. (2016). *Comparative Pathway Integrator: a framework of meta-analytic integration of multiple transcriptomic studies for consensual and differential pathway analysis*. PhD thesis, University of Pittsburgh.

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

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.

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. (submitted). Meta-analytic principal component analysis in integrative omics application.

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(March):btw115.

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

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

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*, page btw788.