A tutorial for MataOmics

Contents

1 Introduction

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

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

2 Preliminaries

2.1 Citing MetaOmics

MetaOmics implements many meta-analytic methodologies by many different authors. Please cite appropriate papers if you used MeteOmics suit, by which the authors will receive professional credits for their work.

- MetaOmics suit itself can be cited as: Ma et al. MetaOmics: Comprehensive Analysis Pipeline and Browser-based Software Suite for Transcriptomic Meta-Analysis.
- 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.

- MetaPredict: 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.

- MetaNetwork: 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 How to start MetaOmics

The full instruction of how to install, start MetaOmics software suit is available at `https://github.com/metaOmics/metaOmics`.

2.2.1 Requirement

- R >= 3.3.1
- Shiny >= 0.13.2

2.2.2 How to install the metaOmics software

- First, 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 metaOmics software

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

  install.packages('shiny')

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

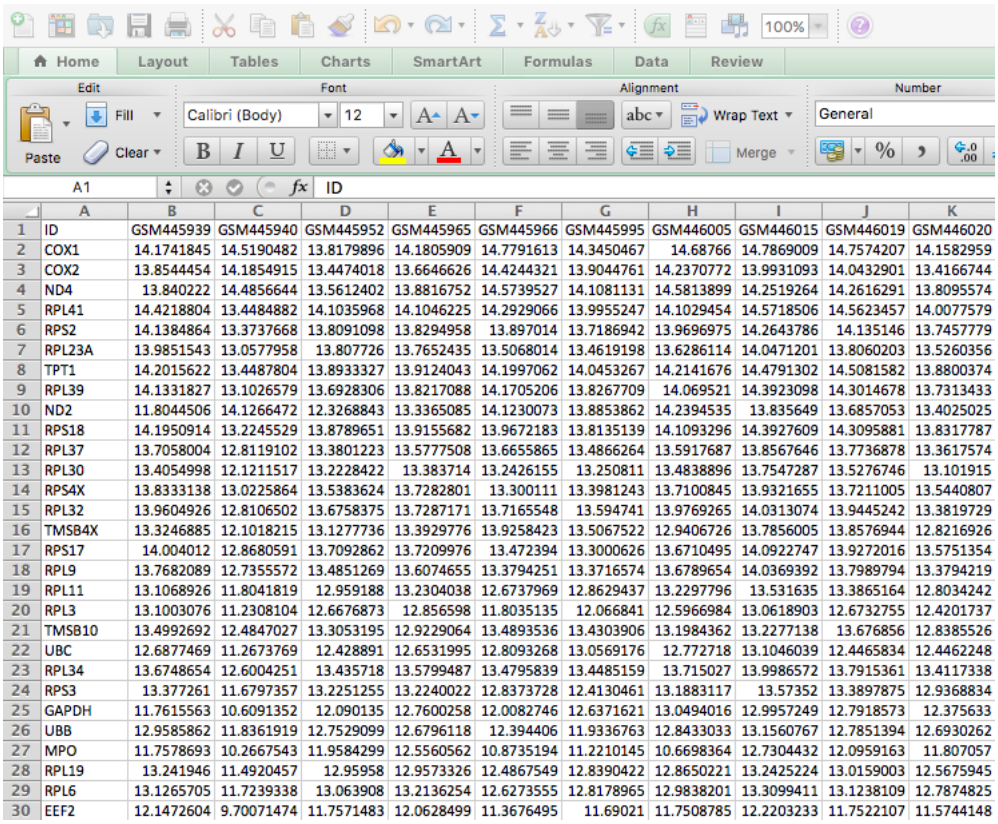
2.3 Question and bug report

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 data and count data.

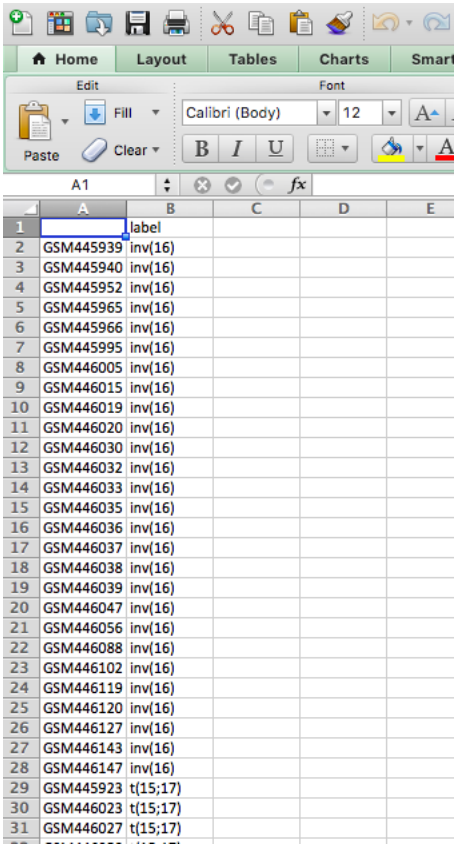


	A	B	C	D	E	F	G	H	I	J	K
1	ID	GSM445939	GSM445940	GSM445952	GSM445965	GSM445966	GSM445995	GSM446005	GSM446015	GSM446019	GSM446020
2	COX1	14.1741845	14.5190482	13.8179896	14.1805909	14.7791613	14.3450467	14.68766	14.7869009	14.7574207	14.1582959
3	COX2	13.8544454	14.1854915	13.4474018	13.6646626	14.4244321	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 input 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 (e.g. case/control labels).



	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)			
11	GSM446020	inv(16)			
12	GSM446030	inv(16)			
13	GSM446032	inv(16)			
14	GSM446033	inv(16)			
15	GSM446035	inv(16)			
16	GSM446036	inv(16)			
17	GSM446037	inv(16)			
18	GSM446038	inv(16)			
19	GSM446039	inv(16)			
20	GSM446047	inv(16)			
21	GSM446056	inv(16)			
22	GSM446088	inv(16)			
23	GSM446102	inv(16)			
24	GSM446119	inv(16)			
25	GSM446120	inv(16)			
26	GSM446127	inv(16)			
27	GSM446143	inv(16)			
28	GSM446147	inv(16)			
29	GSM445923	t(15;17)			
30	GSM446023	t(15;17)			
31	GSM446027	t(15;17)			

Figure 2: A example clinical data format

3.3 Example data with the MetaOmics software

We collected three multi-study examples as testing datasets for the MetaOmics software. Table 1 shows three acute myeloid leukemia (AML) gene expression profiles. Table 2 shows four breast cancer gene expression profiles, in which the first study contains both count data and fpkm data. Table 3 shows gene expression profiles from eight prostate cancer datasets. The leukemia datasets are used to demonstrate MetaProcess, MetaDE, MetaPath, MetaNetwork, MetaPredict, MetaClust and MetaPCA. The prostate cancer datasets are used to demonstrate MetaQC.

After starting MetaOmics, the first page is the MetaOmics setting page as shown in Figure 3. There are 4 tabs on top of the page (at position (1)): Setting, Preprocessing, Saved Data and Toolsets. The welcome page is below the 4 tabs. Further below, the first header is the session information. Why do we need session information? The second header is Directory for Saving Output Files

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

(at position (2)). By clicking "...", user 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 (at position (3)). The third header is Toolsets (at position (4)), where user can click to install desired modules if the "status" shows "not installed". If the packages are installed, there is a checked installed status. Otherwise, users can install individual package by clicking install blue button. The installation progress may take a few minutes for each module. There will be a notification icon after the installation. After the modules are installed to R, restart the MetaOmics software suit so that the shiny application interface is updated with installed modules. Position (5) shows the current active dataset, which will be introduced in Section 4.1 **??**.

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

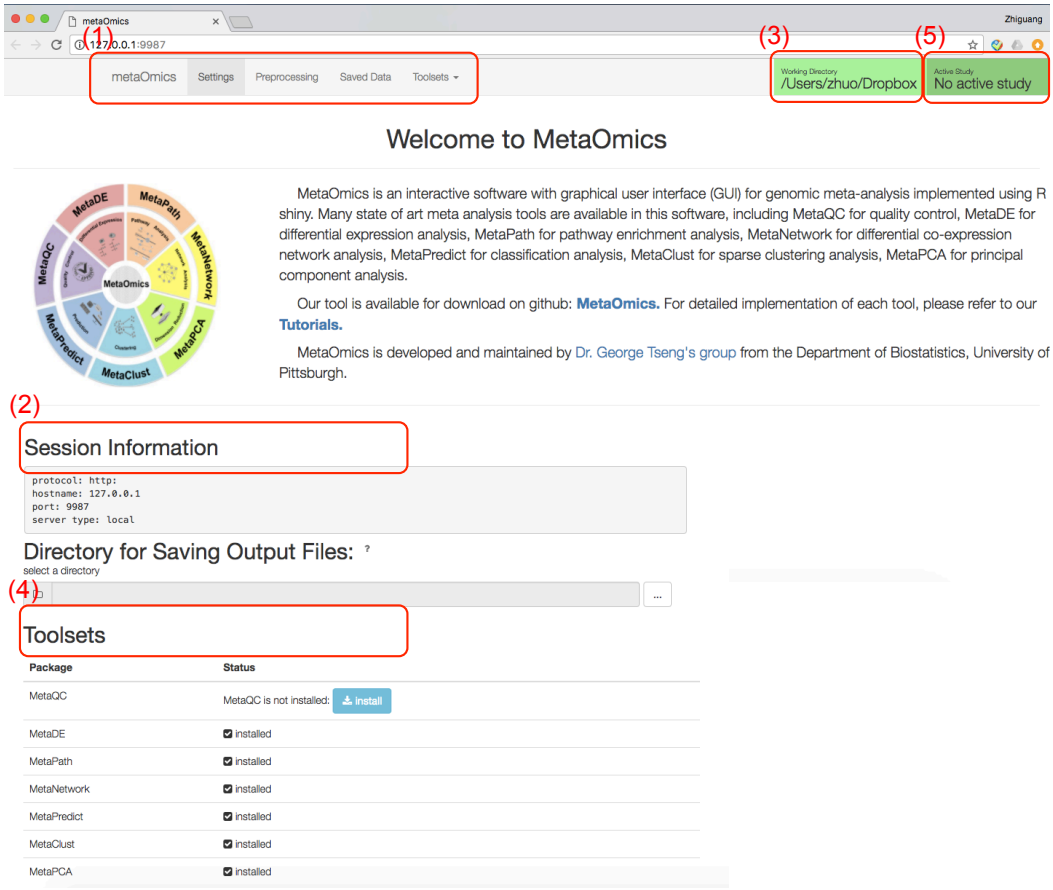


Figure 3: MetaOmics software suite GUI setting page

4 MetaPreprocess

In this subsection, we introduce how to upload your dataset into the MetaOmics suit such that each functional modules can be utilized. The R package for MetaPreprocess module can be found `https://github.com/metaOmics/preproc`.

4.1 Procedure

Step 1 Uploading data:

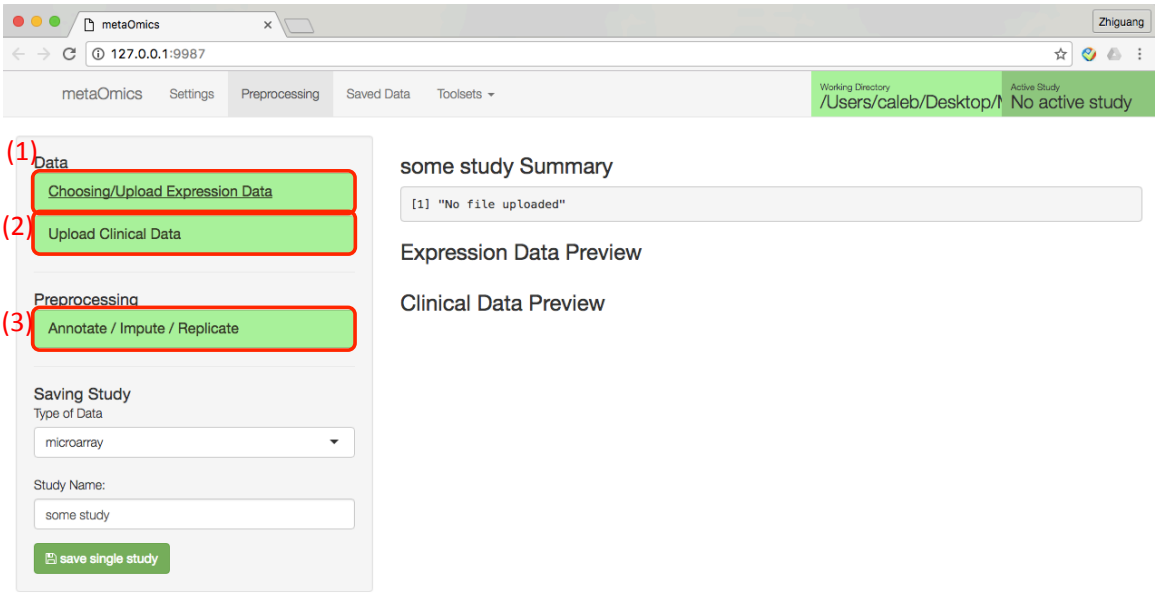


Figure 4: GUI Preprocessing page

After clicking the Preprocessing tab as in Figure 4, users can click on "Choosing/Upload Expression Data" tab to upload individual expression data files or choose the existing saved data file 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 their biological purposes. All MetaOmics modules except for MetaClust require external clinical labels. Three example datasets are available within MetaOmics folder "metaOmics/data/example/", but we will focus on the leukemia dataset ("metaOmics/data/example/") throughout this tutorial.

Step 2 Preprocessing:

The MetaOmics suit also provides handlers (at position (3) of Figure 4) 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 hand side of the page as Figure 5. 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 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 users can search their favorite genes.

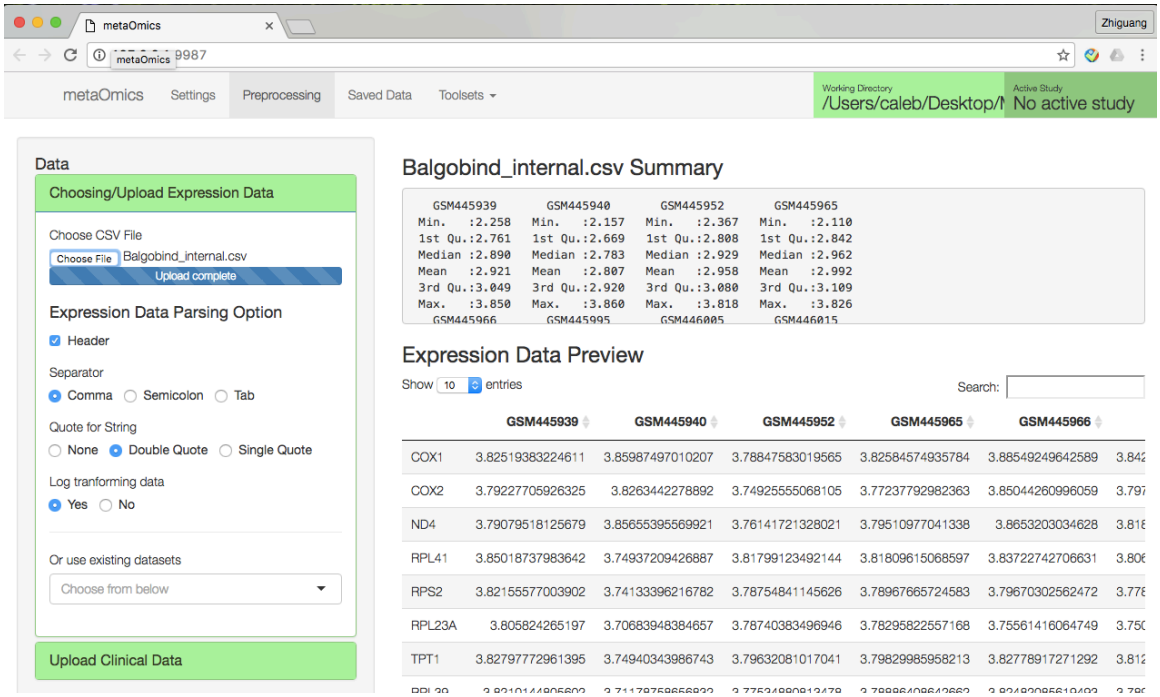


Figure 5: Uploading individual studies

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 **Save single study:** In the next step, specify the data type ("microarray" or "RNA-seq", continuous or discreate) and study name, click "save single study". To upload RNA-seq data, the count data file 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 for meta-analysis. All uploaded studies are now available in the "Saved Data" tab.

4.2 Saved Data tab

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

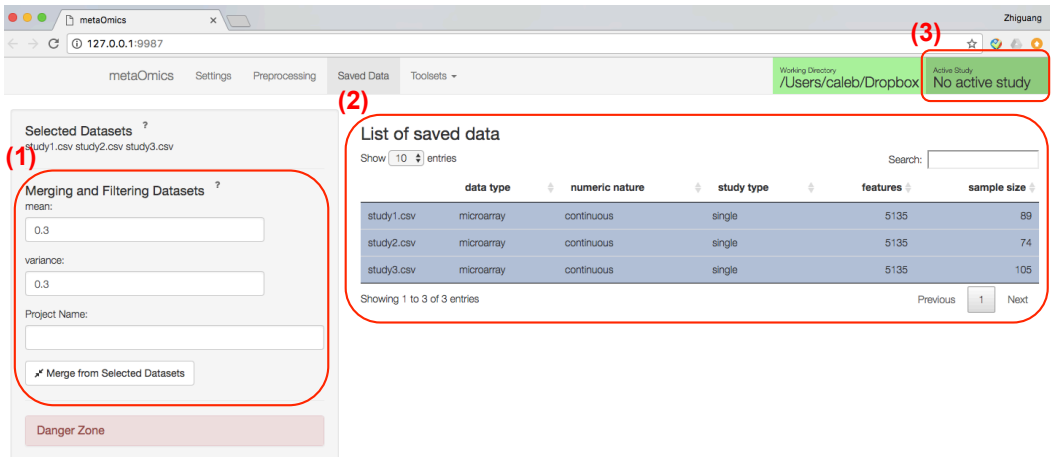


Figure 6: Merge from selected datasets

Step 1 **Merging and Filtering:** All saved datasets from the previous step will be found in Figure 6 (at position (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 30^{th} percentile or low variance (by default, variance lower than 30^{th} percentile)). Low expression genes can be non-expressed genes and low variance genes can be non-informative genes thus including them may contribute false positives. After specifying filtering criteria, enter Project Name and click on the "Merge from Selected Datasets" (at position (1)). A merged dataset

(study type = "multiple") will appear on the "List of saved data" panel (at position (3)). Creating multiple projects to include different sets of studies and genes from 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 selected dataset.

Step 2 Make active dataset:

The last thing to do before using meta-analytic toolsets is to select merged data and click on "Make your dataset Active Dataset" - A big green button in Figure 7. Then the merged data becomes active study and shows up on the top right corner. The active dataset serves as the input for all other MetaOmics modules.

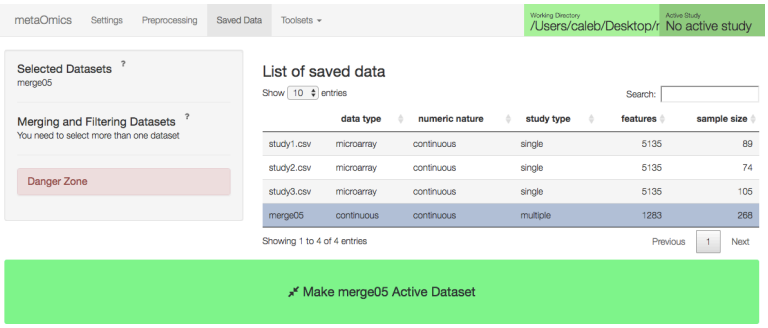


Figure 7: Make merged Dataset Active

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, if the same gene symbol maps to multiple probes, the probe with the largest inner quantile range (IQR) will be selected as a representative for this gene.

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.

5 Toolsets

After the MetaPreprocess, all the preparatory steps are done. It's time to apply seven metaOmics modules 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 these modules. For each module, a summary table to studies and sample sizes is shown on the top left corner. There is an "about" dropdown menu which contains brief introduction and tutorial. The "options" dropdown menu contains common options users can select or tune in the analysis. The "advanced options" section are more technical which we generally do not recommend users to change unless they are familiar with the methods. After applying these metaOmics modules, all result files will be automatically saved in the working directory. 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 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/metaOmics/MetaQC`). The test data used to demo the "MetaQC" package here is merged from eight prostate cancer studies, the details of these studies can be found in Table 3.

5.1.1 Procedure

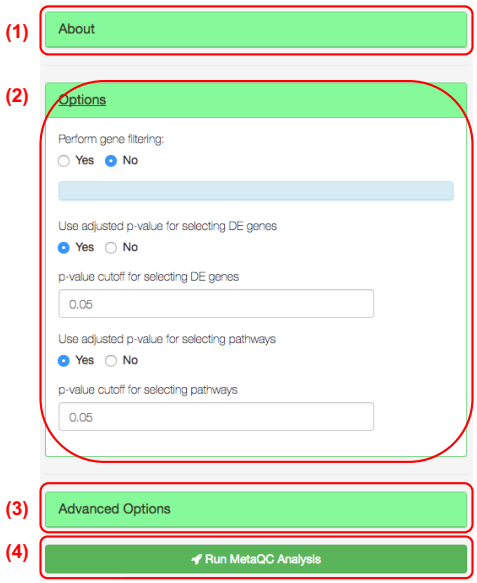


Figure 8: "MetaQC" options

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) in Figure 8), users can specify whether to

- 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.

- users need to specify the approach (either by raw p-value or FDR) and cutoff to select potentially DE genes needed in the computation of IQC, EQC, AQCg and CQCg.
- users need to specify the approach (either by raw p-value or FDR) 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 optional 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

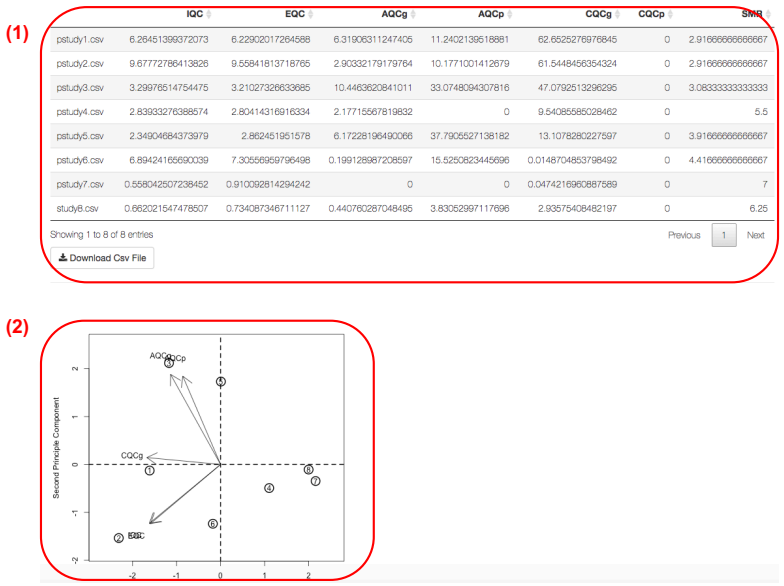


Figure 9: "MetaQC" Results

We applied MetaQC on the eight prostate cancer datasets with 50% filtering by mean and 50% filtering by variance. 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 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. The R package for MetaDE module can be found `https://github.com/metaOmics/MetaDE`.

5.2.1 Procedure

There are two major steps to implement the package: meta differential analysis and pathway analysis. As shown in Figure 10, 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.

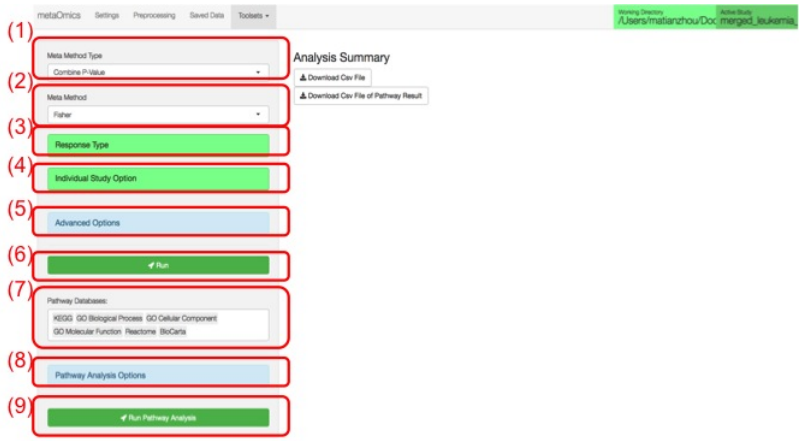


Figure 10: "MetaDE" options

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

5.2.2 Results

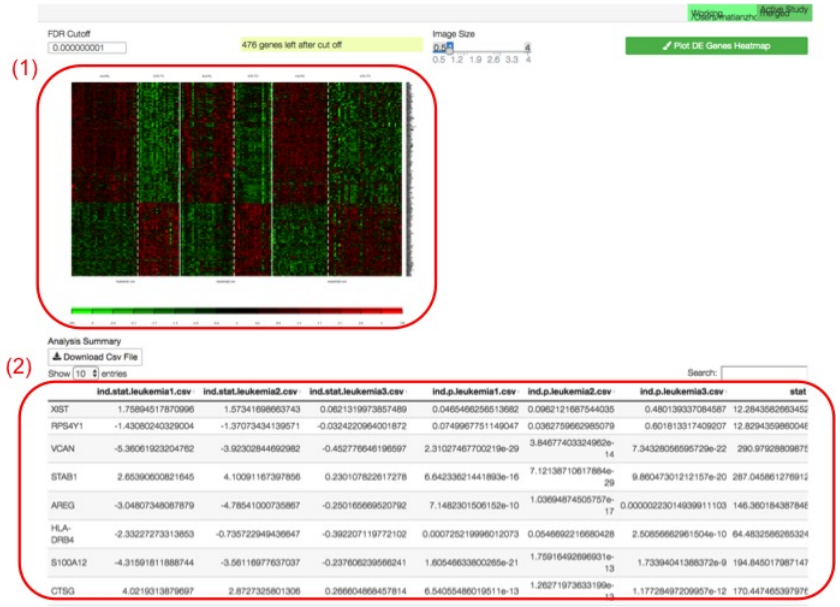


Figure 11: "MetaDE" Results (1)

Two main outputs from the first "meta differential analysis" step in the procedure are shown in Figure 11. 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.

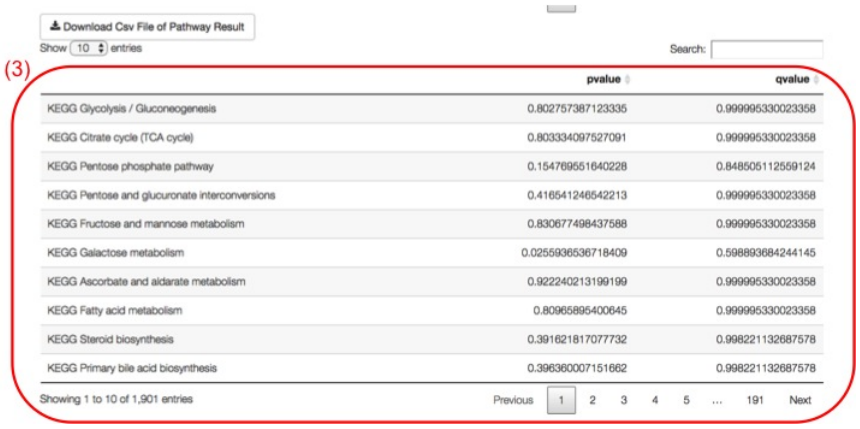


Figure 12: "MetaDE" Results (2)

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

5.3 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. The R package for MetaPath module can be found `https://github.com/metaOmic/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 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.

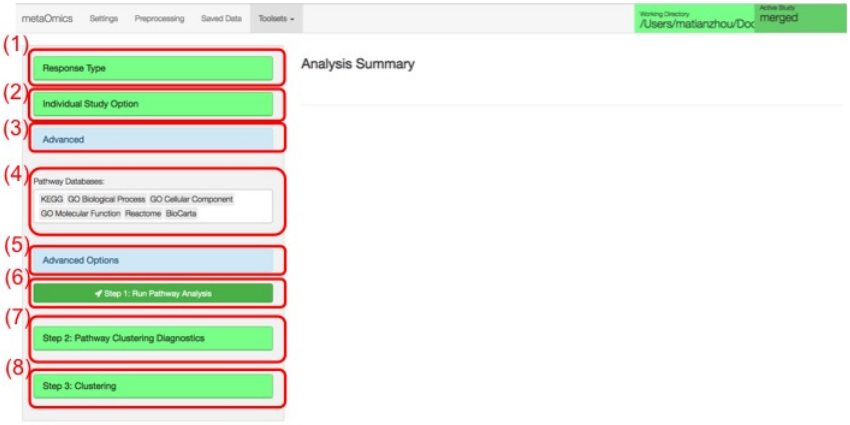


Figure 13: "MetaPath" options

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

5.3.2 Results

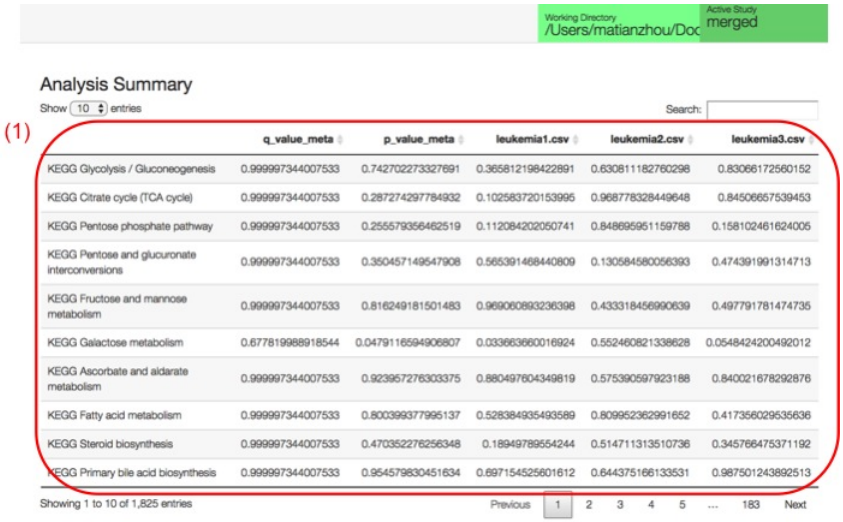


Figure 14: "MetaPath" Results (1)

After the first step is finished, (1) a summary table was generated as shown in Figure 14 (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.

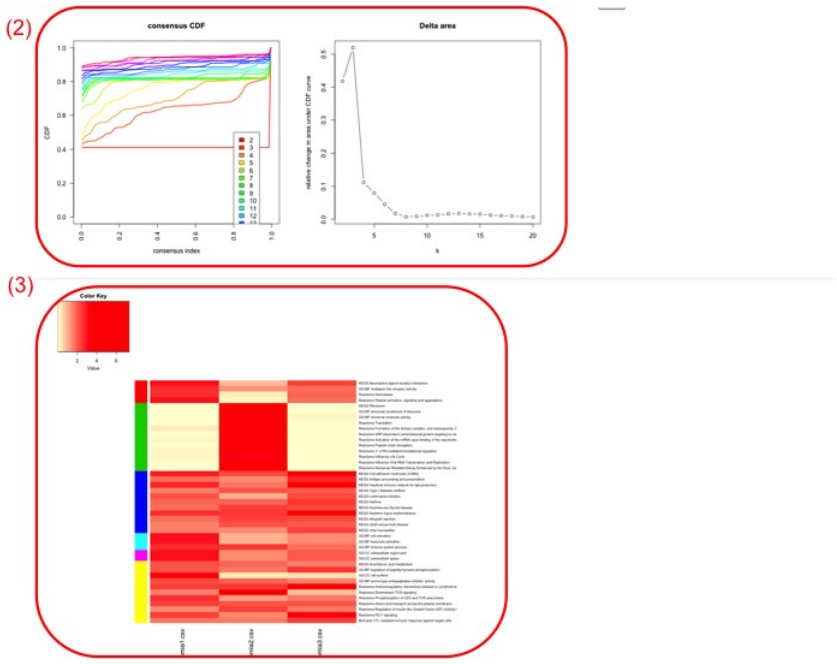
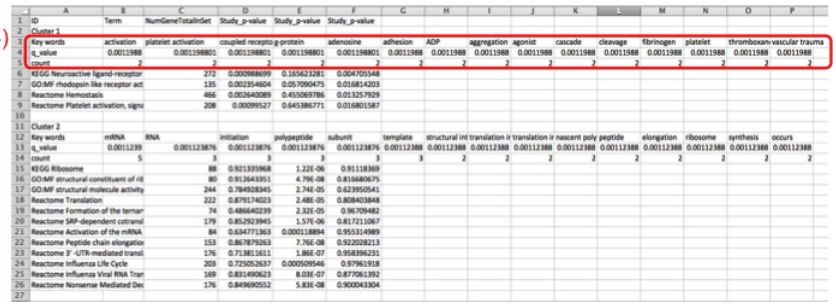


Figure 15: "MetaPath" Results (2)

After the "Pathway Cluster Diagnostics" step is finished, we will see (2) 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, 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	G	H	I	J	K	L	M	N	O	P
1	ID	Term	NumGeneTotalInSet	Study_p-value	Study_p-value	Study_p-value										
2	Cluster 1															
3	Key words	activation	platelet activation	coupled receptor	g-protein	adenosine	adhesion	ADP	aggregation	agonist	cascade	cleavage	fibrinogen	platelet	thromboxane	vascular trauma
4	q_value	0.0011988	0.001198801	0.001198801	0.001198801	0.001198801	0.0011988	0.0011988	0.0011988	0.0011988	0.0011988	0.0011988	0.0011988	0.0011988	0.0011988	0.0011988
5	count	2	2	2	2	2	2	2	2	2	2	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	template	structural int	translation ir	translation ir	nascent poly	peptide	elongation	ribosome	synthesis	occurs
13	q_value	0.0011239	0.001123876	0.001123876	0.001123876	0.001123876	0.00112388	0.00112388	0.00112388	0.00112388	0.00112388	0.00112388	0.00112388	0.00112388	0.00112388	0.00112388
14	count	5	3	3	3	3	3	2	2	2	2	2	2	2	2	2
15	KEGG Ribosome		88	0.921335968	1.22E-06	0.91118369										
16	GO:MF structural constituent of ri		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																

Figure 16: "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.

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/metaOmic/MetaNetwork`.

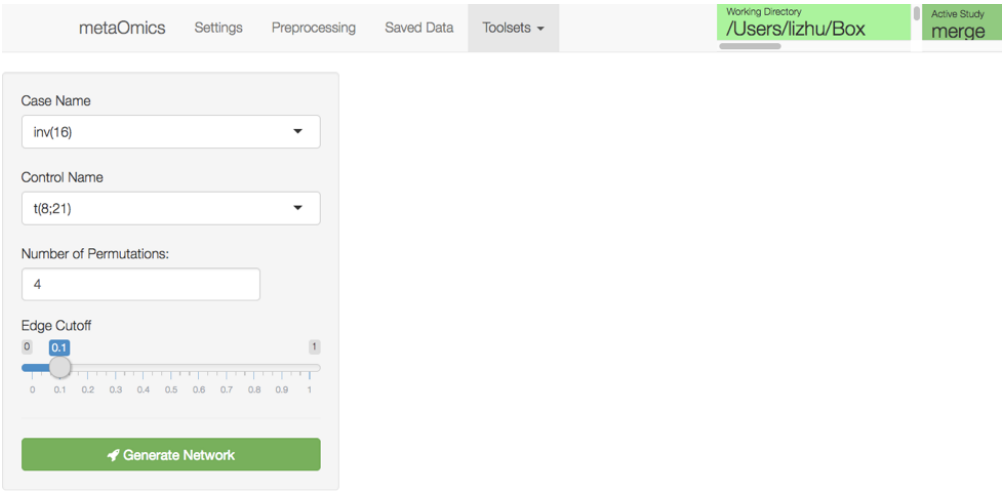


Figure 17: MetaNetwork homepage

MetaNetwork includes three steps to get differentially co-expressed networks: generate network, search for basic modules, and assemble supermodules. The left screen 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 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

After the generate network step is done, the control panel will be as Figure 18. The next step is to search for basic modules. Users need to specify the number of repeats used for each initial seed modules (Number to repeat), the maximum Monte Carlo steps for simulated annealing algorithm (MC Steps), and the maximum pairwise Jaccard index allowed for basic modules (Jaccard Cutoff). After clicking **Search for basic modules** button, screen will show message indicating the algorithm is running to search for basic modules.

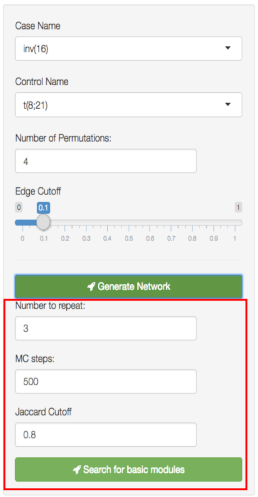


Figure 18: MetaNetwork control panel for search for basic modules

Search for basic modules will take minutes, 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 Figure 19.

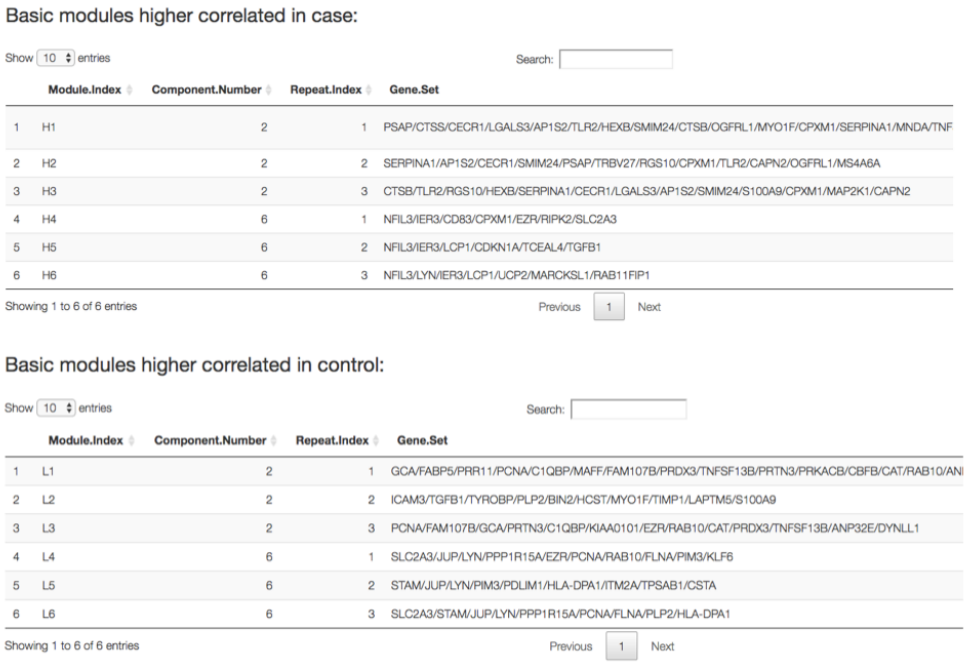


Figure 19: MetaNetwork output from search for basic modules step

Step 3 Assemble supermodules

After search for basic modules step is done, the control panel will be 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.

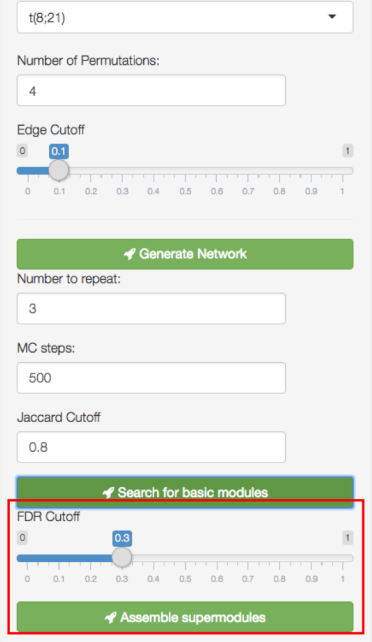


Figure 20: MetaNetwork control panel for assemble supermodules step

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

Generate Network

After the generate network step is done, no output will show up in the screen. Instead, a message box 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:

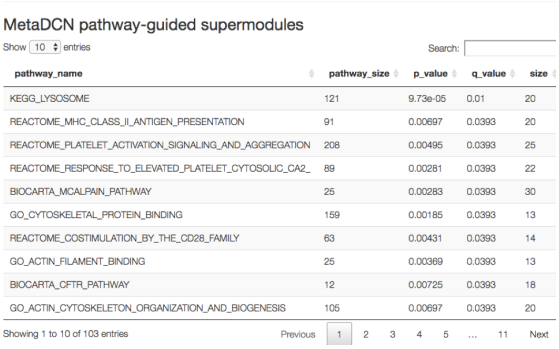
- basic_modules_summary_forward_weight_w1.csv is a summary table of basic modules that are higher correlated in case, detected using w1.
- basic_modules_summary_backward_weight_w1.csv is a summary table of basic modules that are higher correlated in control, detected using w1.
- threshold_forward.csv 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_ANTIGEN_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

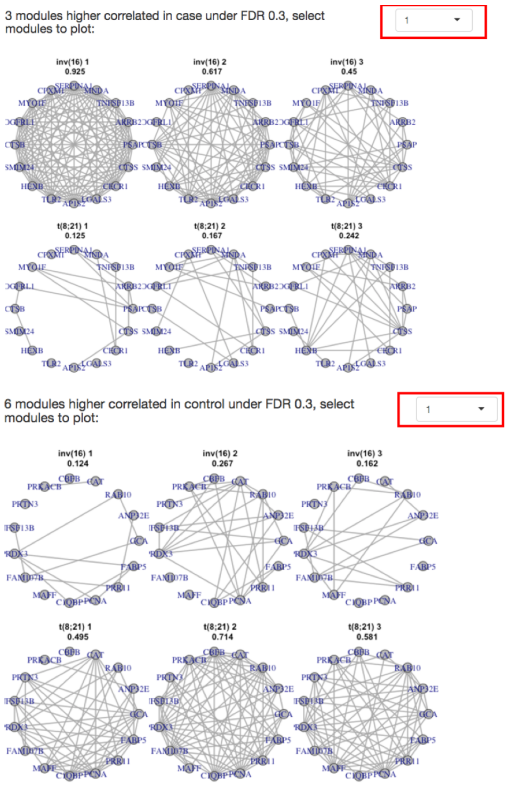


Figure 22: MetaNetwork select basic modules to plot

5.5 MetaPredict

Top scoring pairs is a robust algorithm for predicting gene expression profiles, which adopts nonparametric 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 `https://github.com/metaOmic/MetaPredict`.

After opening the MetaPredict page, as shown in Figure 23, 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) and "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 excuting tabs ("Train model" and "Predict").

5.5.1 Procedure

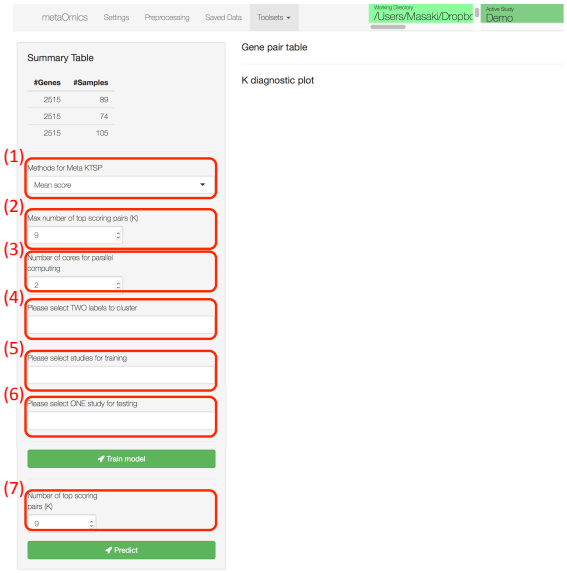


Figure 23: Homepage of MetaPredict

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) 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. Our interface will pop up all labels that are available. Then, select the dataset as training data and testing respectively, and click the "Train model" tab 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 output 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

tab “Predict” (Figure 24). Finally, a confusion matrix is output to show the prediction results ((1) in Figure 24).

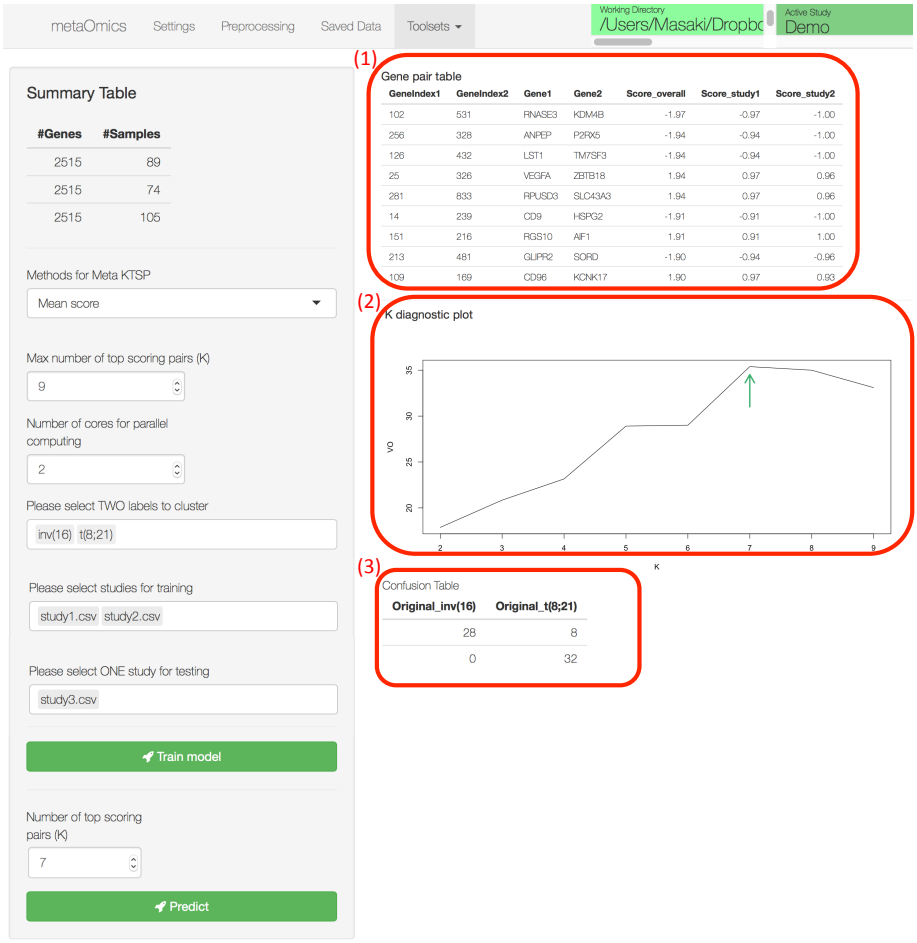


Figure 24: Results for MetaPredict.

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

A confusion matrix is output to show the prediction results ((1) in Figure 24). The prediction results are also saved in the folder.

5.6 MetaClust

By clicking toolsets and then metaClust, users are directed to metaClust home page as Figure 25. 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. The R package for MetaClust module can be found `https://github.com/metaOmic/MetaSparseKmeans`.

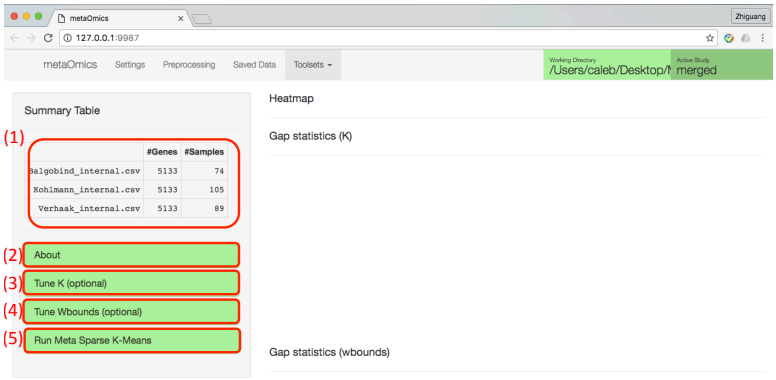


Figure 25: MetaClust home page

5.6.1 Procedure

Figure 25 shows the home page of MetaClust. On the top left panel users can see 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:

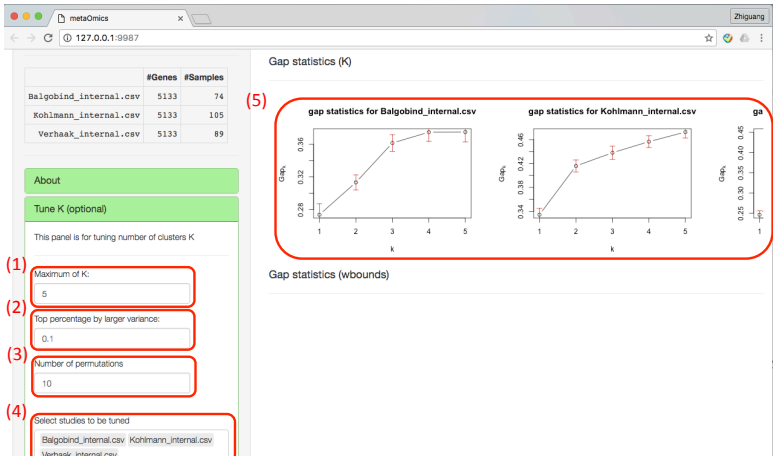


Figure 26: Tuning parameter selection for number of clusters

If the users are not sure what is number of clusters, they can start to use the Tune K panel as in Figure 26. 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 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.

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.

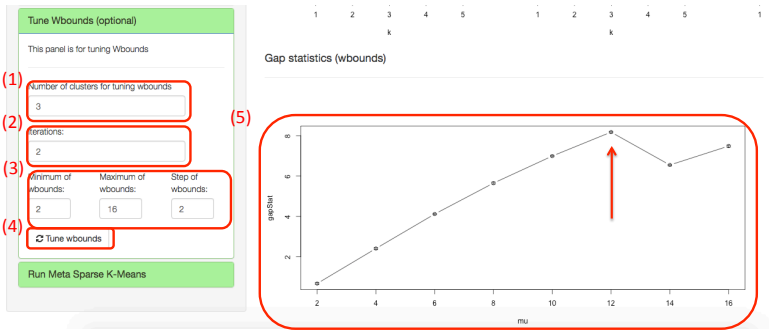


Figure 27: Wbound selection

Again, gap statistics will be used for tuning Wbounds. Users will specify number of clusters for tuning Wbounds (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 27 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 28.

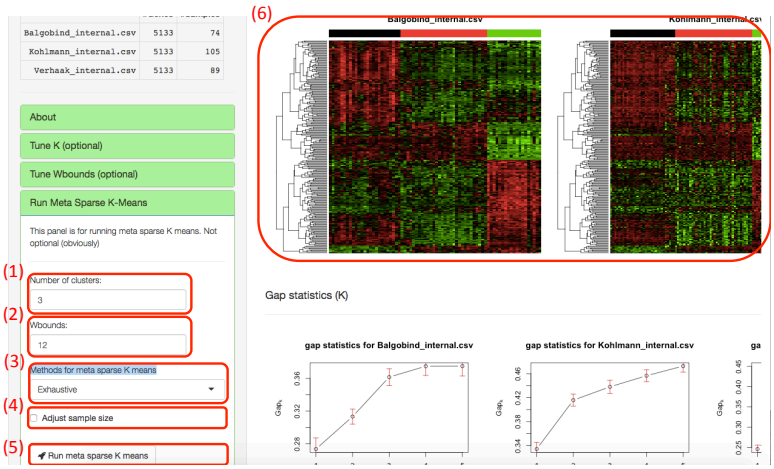


Figure 28: Result for MetaClust

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 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 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.

5.6.2 Results

The result is shown in Figure 28 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.

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 toolsets and then metaPCA, users are directed to metaPCA home page as Figure 29. The R package for MetaPCA module can be found `https://github.com/metaOmic/metaPCA`.

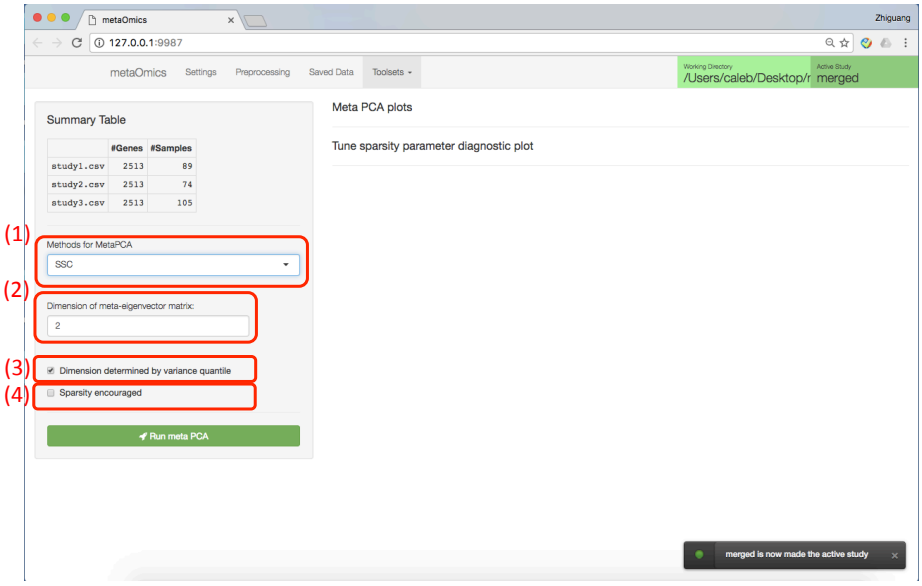


Figure 29: MetaPCA settings

5.7.1 Procedure

Step 1 Specify parameters

There are very few parameter to be specify in metaPCA, as in Figure 29. There are two methods for MetaPCA (at position (1)). SSC represent MetaPCA via sum of squared cosine (SSC) maximization. SV represent 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 (at position (2)) allows user to specify dimension of the output meta-eigenvector matrix. The checkbox of "dimension determined by variance quantile" is suggested to be selected (at position (3)), If it is selected, 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 selected (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, MetaPCA 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 predefined level of variance quantile 80%.

2. If sparsity encouraged is selected, there are extra tuning parameter (λ) that may need to be tuned.

    - Min λ: lower bound of the searching space of λ.
    - Max λ: upper bound of the searching space of λ.
    - Step of of λ: stepsize of the searching space of λ.
    - Tuning parameter for sparsity: Tuning parameter for sparsity that will be used for sparse metaPCA.

5.7.2 Results

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 from metaPCA are saved.

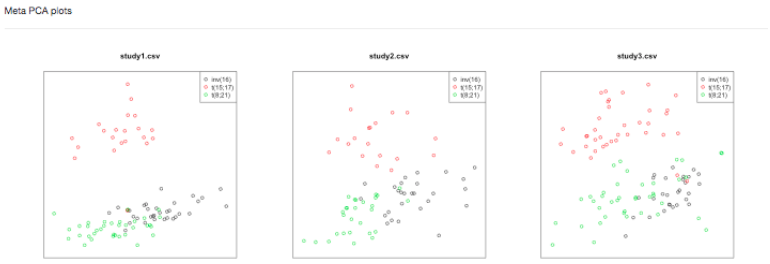


Figure 30: MetaPCA result

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.

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. (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.

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. (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.

Kohlmann, A., Kipps, T. J., Rassenti, L. Z., Downing, J. R., Shurtleff, S. A., Mills, K. I., Gilkes, A. F., Hofmann, W.-K., Basso, G., Dell'Orto, M. C., Foà, R., Chiaretti, S., Vos, J. D., Rauhut, S., Papenhausen, P. R., Hernández, J. M., Lumbreras, E., Yeoh, A. E., Koay, E. S., Li, R., min Liu, W., Williams,

P. M., Wieczorek, L., and Haferlach, T. (2008). An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the microarray innovations in LEukemia study prephase. *British Journal of Haematology*, 142(5):802–807.

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

Li, J., 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.

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.

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

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, 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*, page btw788.