A tutorial for metaOmic

Contents

1 Introduction **2**

2 Preliminaries **2**
- 2.1 Citing MetaOmics . . . 2
- 2.2 Installation . . . 3
  - 2.2.1 Requirement . . . 3
  - 2.2.2 How to start the app . . . 3
  - 2.2.3 How to start the documentation . . . 3
- 2.3 Question and bug report . . . 4

3 Prepare data **4**
- 3.1 Raw data . . . 4
- 3.2 Clinical data . . . 5

4 Toolsets **6**
- 4.1 MetaQC . . . 6
- 4.2 MetaDE . . . 6
  - 4.2.1 Meta analysis . . . 7
  - 4.2.2 Visualization . . . 11
  - 4.2.3 Downstream pathway analysis . . . 13
- 4.3 MetaPath . . . 15
  - 4.3.1 Run pathway analysis . . . 15
  - 4.3.2 Pathway clustering . . . 21
- 4.4 MetaClust . . . 21
  - 4.4.1 About . . . 22
  - 4.4.2 Tune K . . . 22
  - 4.4.3 Tune Wbounds . . . 23
  - 4.4.4 Run Meta Sparse K-Means . . . 24
- 4.5 MetaPCA . . . 25
- 4.6 MetaKTSP . . . 25
- 4.7 MetaDCN . . . 25
- 4.8 MetaLA . . . 25

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 a real example.

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.

– .

- 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: not published yet.
- MetaKTSP: not published yet.
- MetaDCN: not published yet.
- MetaLA: not published yet.

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)

2.2.3 How to start the documentation

- Install rmarkdown for R
- Inside 'doc' directory, start R console, and:
- in R

```
rmarkdown::run(shiny_args=list(port=9988, launch.browser=T))
```

- or in command line

```
R -e "rmarkdown::run(shiny_args=list(port=9988, launch.browser
    =T))"
```

If you run into an issue with something like 'pandoc version 1.12.3 or higher is required and was not found.?, just install pandoc manually. For example, on Mac, it would be 'brew install pandoc'. If you have Rstudio, you can also to get rstudio's pandoc environment. Go to rstudio console and find the system environment variable for 'RSTUDIO_PANDOC'

In R:

```
Sys.getenv("RSTUDIO_PANDOC")
```

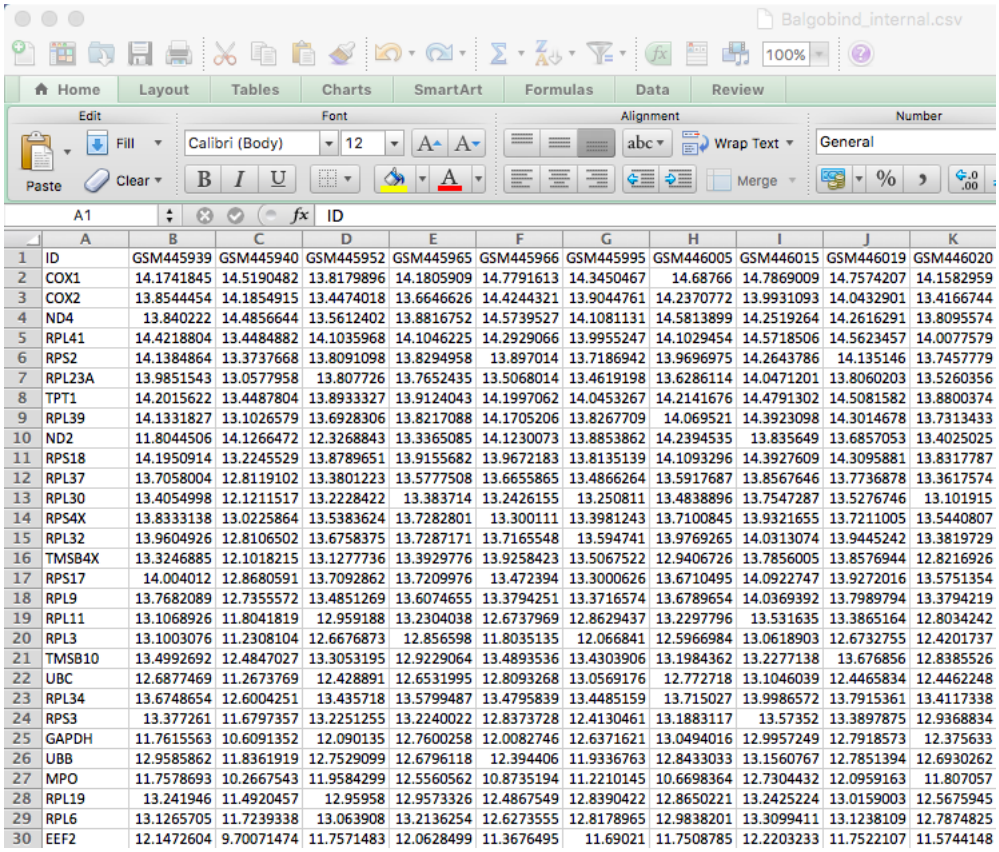
2.3 Question and bug report

Ask Anzhe what is the appropriate way to maintain the package?

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. The first row is sample ID. Valid data type includes continuous, count.

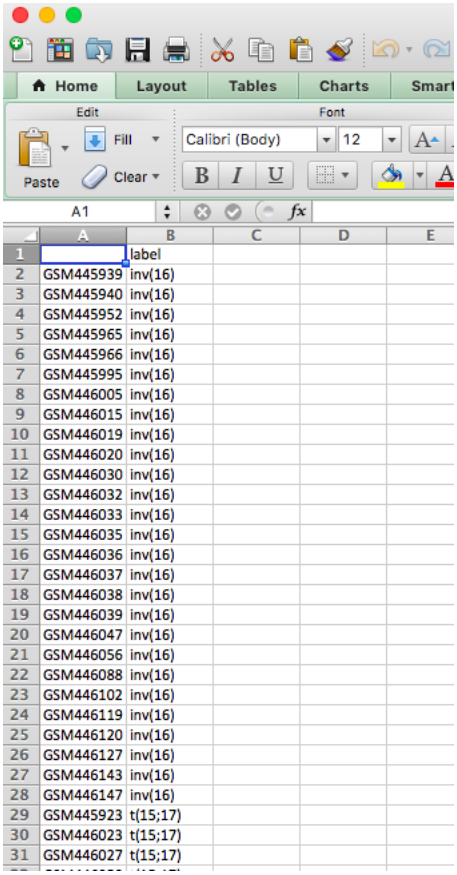


	A	B	C	D	E	F	G	H	I	J	K
1	ID	GSM445939	GSM445940	GSM445952	GSM445965	GSM445966	GSM445995	GSM446005	GSM446015	GSM446019	GSM446020
2	COX1	14.1741845	14.5190482	13.8179896	14.1805909	14.7791613	14.3450467	14.68766	14.7869009	14.7574207	14.1582959
3	COX2	13.8544454	14.1854915	13.4474018	13.6646626	14.4244321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



	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

4 Toolsets

4.1 MetaQC

4.2 MetaDE

MetaDE package implements 12 major meta-analysis methods for differential expression analysis, and now it allows the analysis of both microarray and RNA-seq data. In this tutorial, we will demonstrate the MetaDE pipeline step by step using two meta-analysis methods: Fisher's method and Adaptively weighted Fisher's method (AW-Fisher). Please refer to Fisher (1925) and Li et al. (2011) for details of these two methods. Individual MetaDE package is also available on GitHub at `https://github.com/metaOmic/MetaDE`.

4.2.1 Meta analysis

After opening the MetaDE page, as shown in Figure 3, there are 2 drop-down menus ("Meta Method Type" and "Meta Method") and 4 tabs on the left of the page ("Response Type", "Setting Individual Study Method", "Advanced Options" and "Run"). We generally suggest users not to change any parameter setting in the "Advanced Options" unless users know the underlying methodology well.

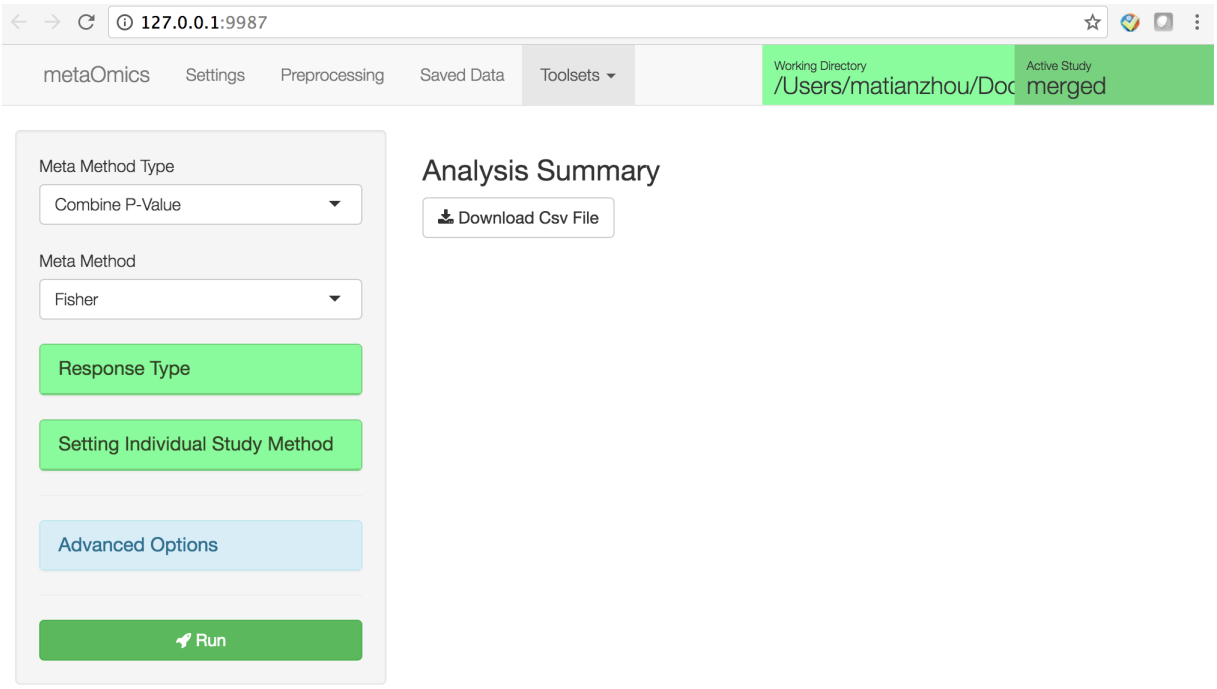


Figure 3: Homepage of MetaDE

For Fisher's method, we will choose "Combine P-value" and "Fisher" (Figure 4); and for AW Fisher's method, we will choose "Combine P-value" and "AW Fisher" (Figure 5).

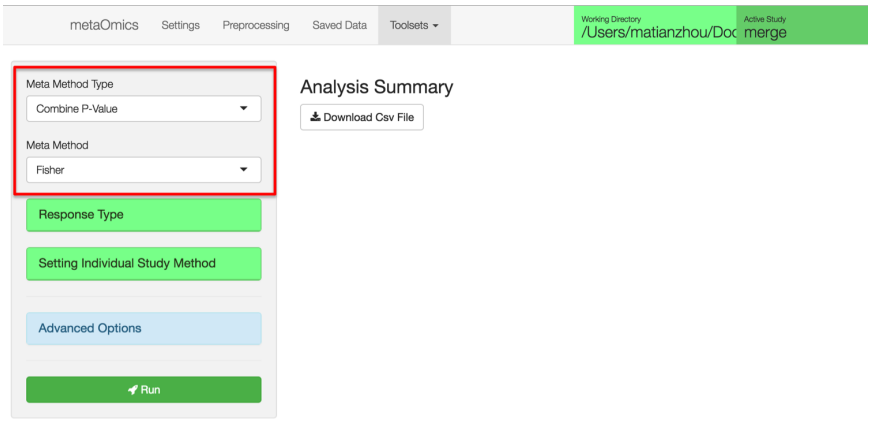


Figure 4: Fisher's method setting

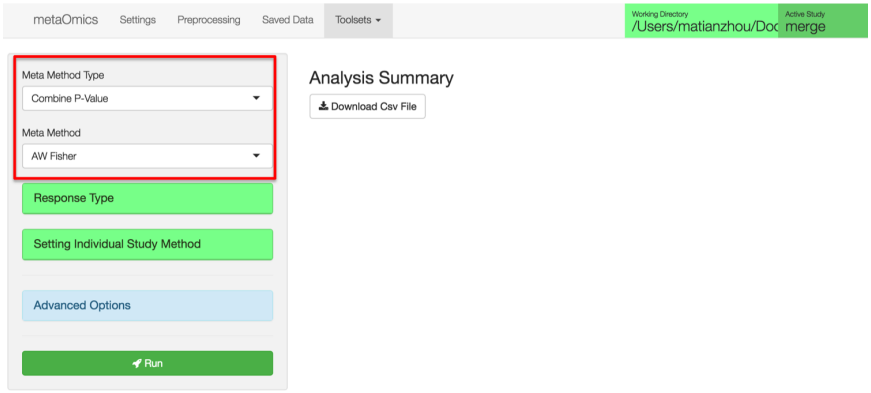


Figure 5: AW Fisher's method setting

Then, in the next step, we click on "Response Type", for two-class DE analysis, choose "Two Class Comparison", choose the group label name for the Label Attribute (from the column names of your clinical data). Then for the group label (a factor of at least two levels), choose a name for the "Control Label" and "Experimental Label", respectively (Figure 18).

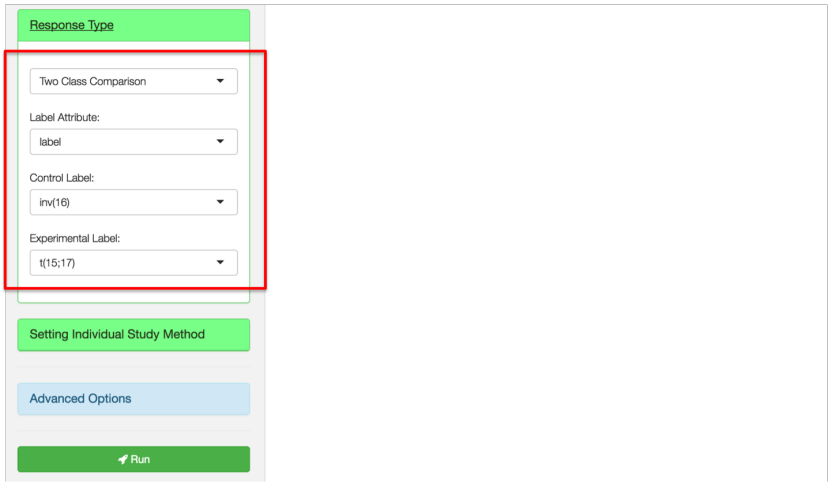


Figure 6: Response type setting

Next, we click on "Setting Individual Study Method" and choose "LIMMA" to perform DE analysis in each individual study (Figure 19). Available options include "LIMMA" and "SAM" for continuous data (e.g. microarray), "edgeR", "DESeq2" and "limmaVoom" for discrete data (e.g. RNA-seq count). Details of the above mentioned DE methods can be found in XXX (reference).

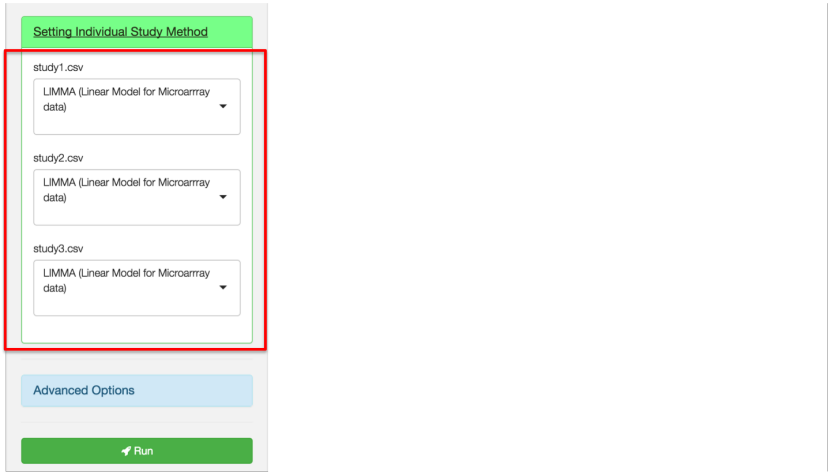


Figure 7: Individual study DE analysis method

** Optionally, we can click on "Advanced Options", choose "Parametric=yes" so parametric methods will be used for inference. We can choose whether to

adjust for any important "Covariates" (e.g. potential confounders) and we set the alternative hypothesis to be "abs", i.e. two-sided test (Figure 8).

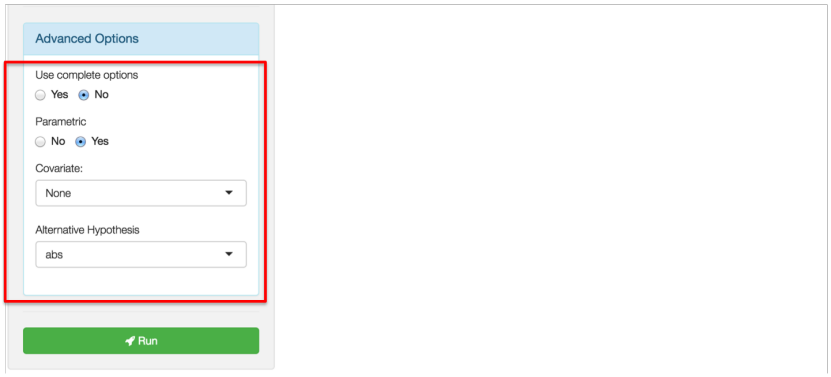


Figure 8: Advanced Options

After we click on "Run", we will see a summary table generated on the right of the page as shown in Figure 9. The "Analysis Summary" includes the analysis results of all genes, including individual study statistics/p-value, meta-analysis statistics/p-value/FDR, etc. Users can search the gene name in the "Search" bar, and download the full table as a csv file by clicking on "Download Csv File".

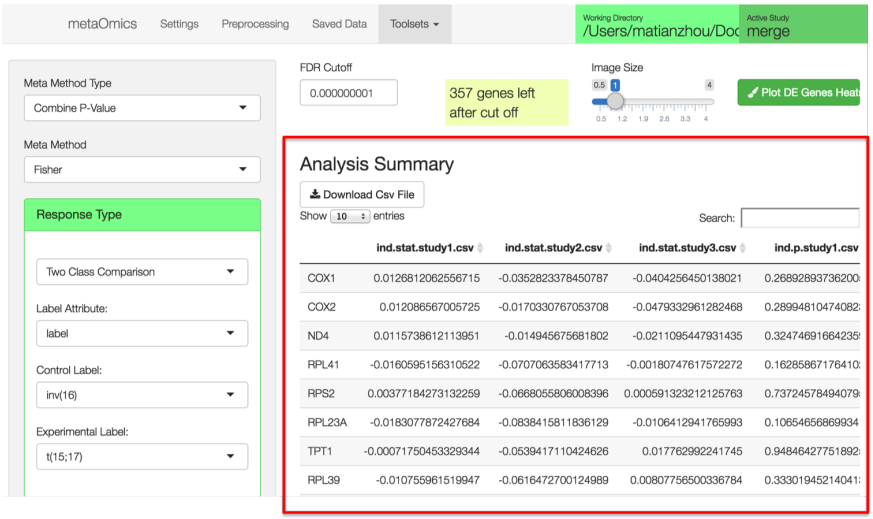


Figure 9: Summary table of Meta-analysis results

4.2.2 Visualization

In addition to tabular output, for better visualization, users can also plot heatmap of DE genes at specified FDR cutoff. Users can enter the "FDR Cutoff" (number of genes selected are shown interactively), then click on "Plot DE Genes Hetamap" to draw the heatmap. The "image size" can be adjusted by dragging the scroll bar (Figure 10).

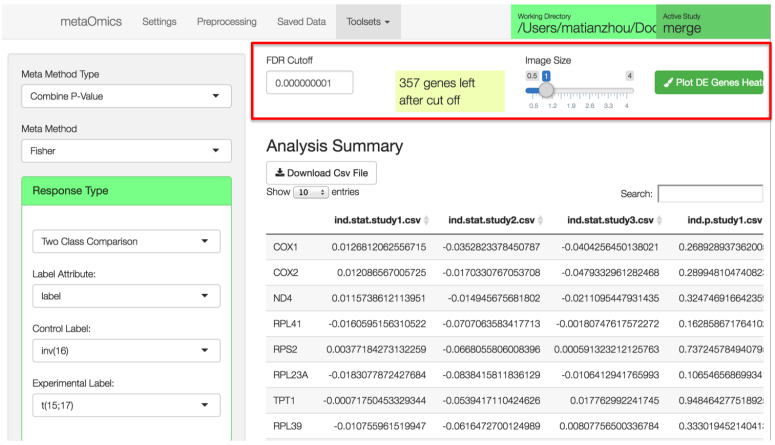


Figure 10: Plot heatmap setting

Shown in Figure 11 and Figure 12 are two heatmaps generated from Fisher and AW Fisher results. Note that one additional column of weight distribution across studies is added in the heatmap of AW-Fisher, and all genes in the heatmap are sorted based on their weight distribution.

Figure 11: Heatmap based on Fisher results

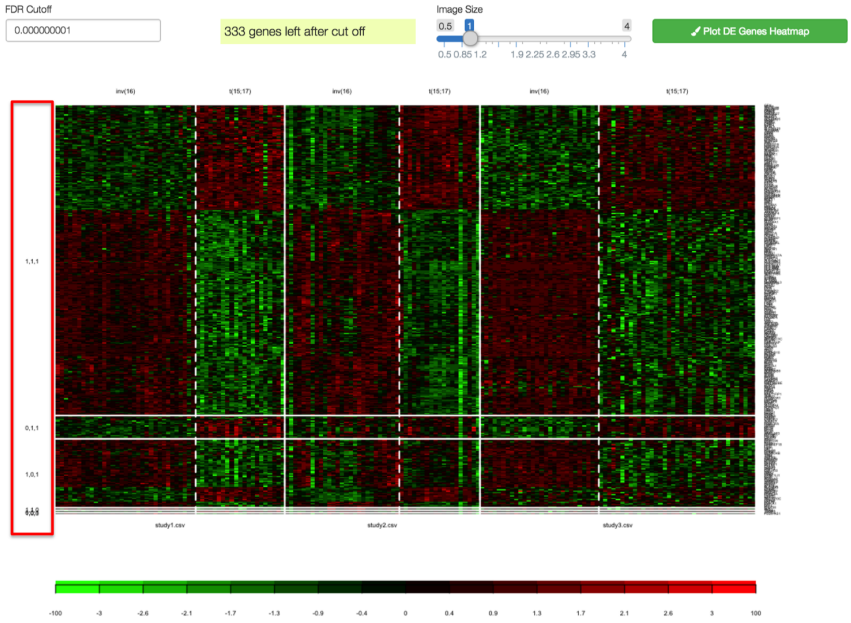


Figure 12: Heatmap based on AW Fisher results

4.2.3 Downstream pathway analysis

Upon getting the meta differential analysis results (users have to perform meta DE analysis first before pathway analysis), we will further perform downstream pathway analysis. As shown in Figure 21, we first need to choose the Pathway databases to be used as shown in the highlighted box. Next we will choose the pathway enrichment method, the default is the Kolmogorov-Smirnov test (KS test, Figure 14). Alternatively, one can choose to use Fisher's exact test, once this method is selected, you need to pick a DE gene set using a hard threshold, either by p-value cutoff or the number of top ranked genes (Figure 23). Lastly, we will specify the minimum/maximum pathway size of pathways we wish to include for functional analysis, then click "Run Pathway Analysis".

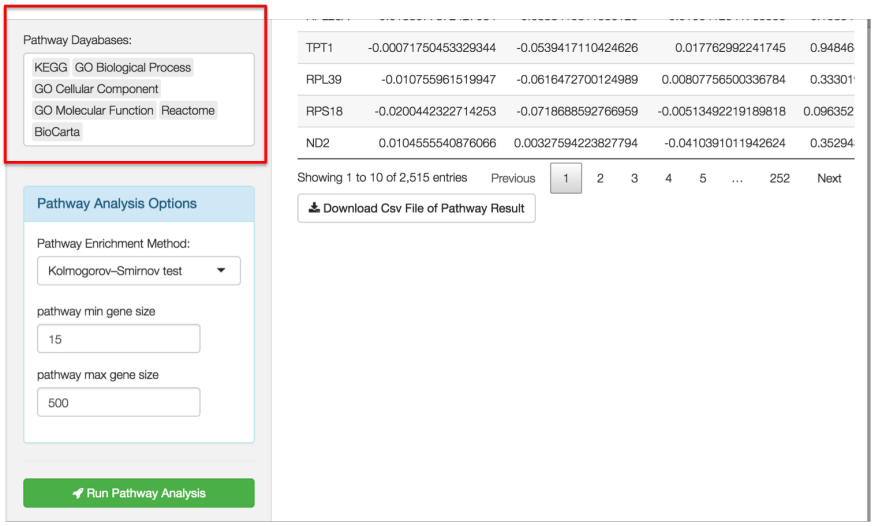


Figure 13: Selection of pathway database

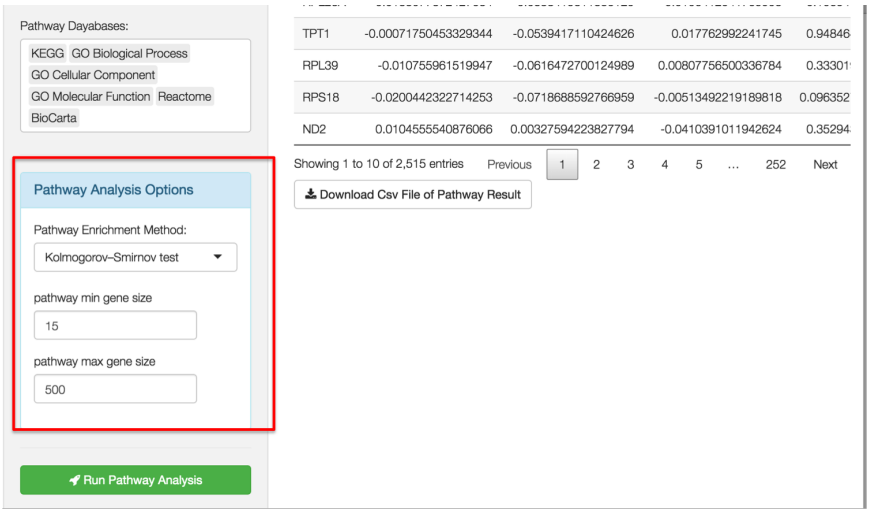


Figure 14: KS test setting

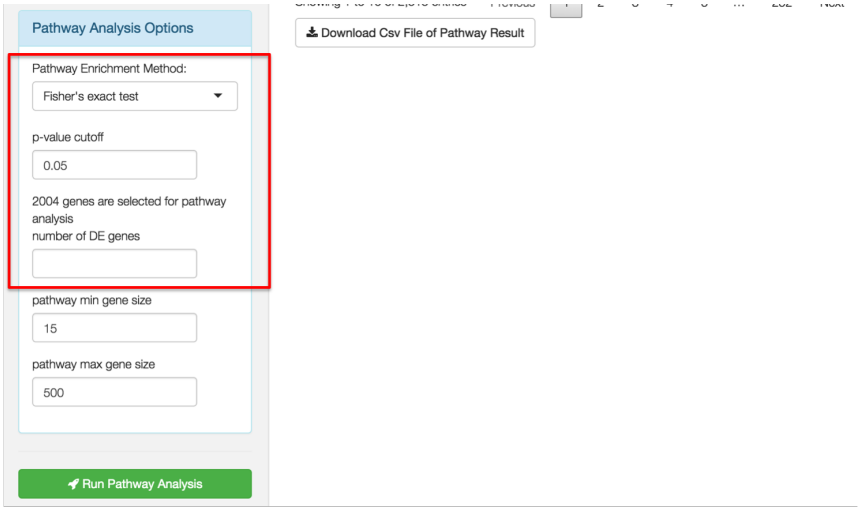


Figure 15: Fisher's exact test setting

We will then see a summary table of pathway analysis results generated on the right of the page as shown in Figure 16. Users can search the pathway name in the "Search" bar, and download the full table as a csv file by clicking on "Download Csv File of Pathway Result" as highlighted.

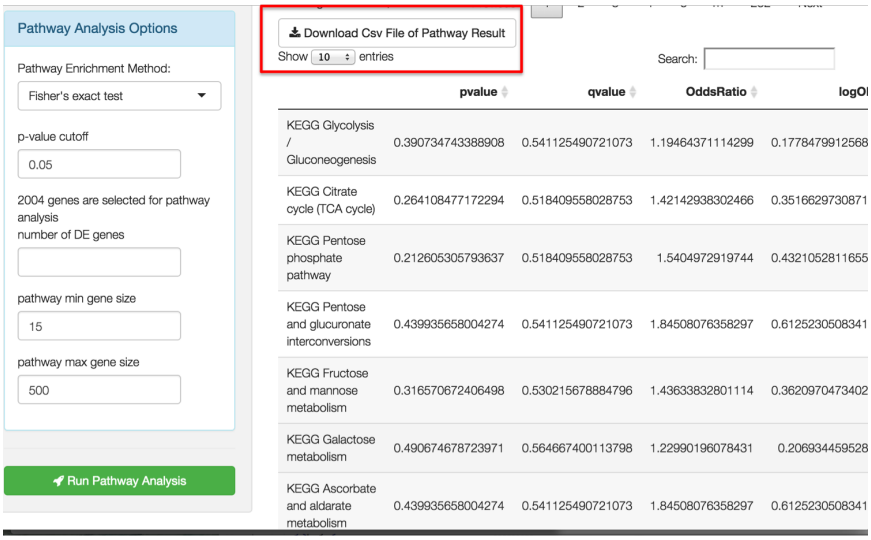


Figure 16: Summary of pathway analysis result

4.3 MetaPath

4.3.1 Run pathway analysis

After opening the MetaPath page, as shown in Figure 17, there are 3 main steps to implement MetaPath. We generally suggest users not to change any parameter setting in "Advanced" and "Advanced Options" unless users know the underlying methodology well.

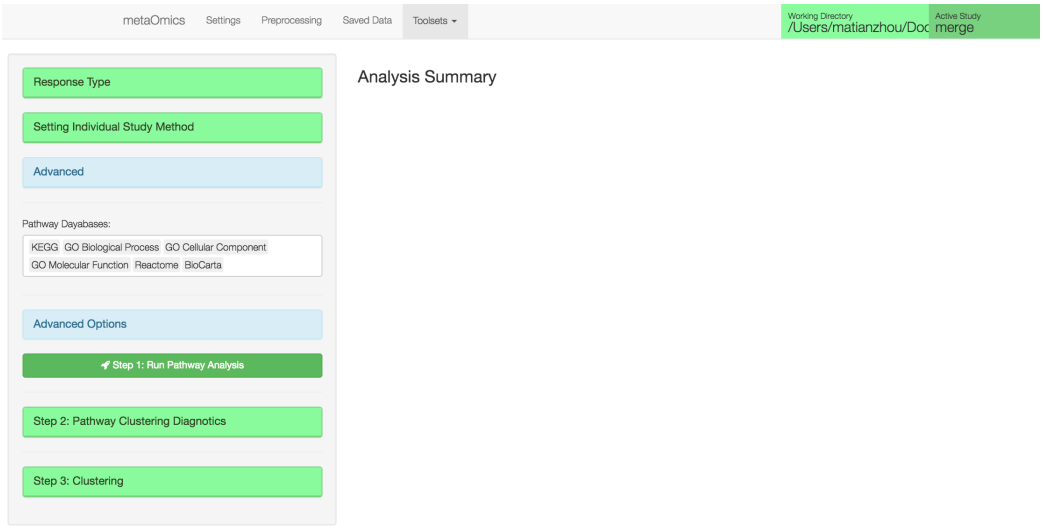


Figure 17: Homepage of MetaPath

To start, we need to perform DE (association) analysis in each individual study before performing pathway analysis for functional annotation. We click on "Response Type", for two-class DE analysis, choose "Two Class Comparison", choose the group label name for the Label Attribute (from the column names of your clinical data). Then for the group label (a factor of at least two levels), choose a name for the "Control Label" and "Experimental Label", respectively (Figure 18).

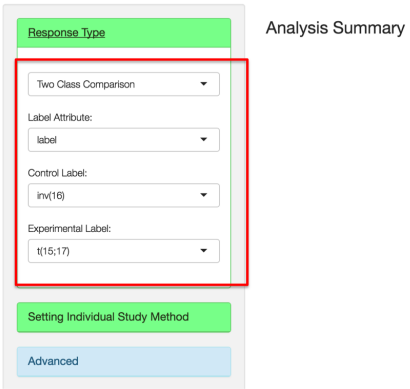


Figure 18: Response type setting

Then we click on "Setting Individual Study Method" and choose "LIMMA" to perform DE analysis in each individual study (Figure 19). Available options include "LIMMA" and "SAM" for continuous data (e.g. microarray), "edgeR", "DESeq2" and "limmaVoom" for discrete data (e.g. RNA-seq count). Details of the above mentioned DE methods can be found in XXX (reference).

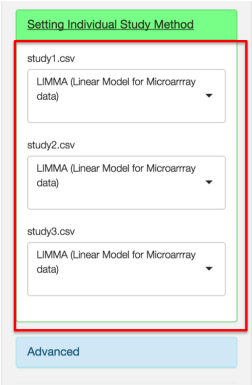


Figure 19: Individual study DE analysis method

Optionally, we can click on "Advanced" and choose whether to adjust for any important "Covariates" (e.g. potential confounders) and we set the alternative hypothesis to be "abs", i.e. two-sided test (Figure 20).

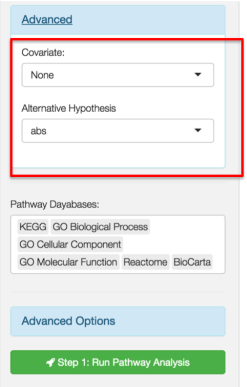


Figure 20: Advanced Options for single study association analysis

Next, we need to specify the parameters for pathway analysis. As shown in Figure 21, we first need to choose the Pathway databases to be used as shown in the highlighted box. Then optionally, we can click on the "Advanced Options" tab and choose the software to be used: CPI or MAPE. For CPI, we

can use the Kolmogorov-Smirnov test (Figure 22).CPI implements AW Fisher's method for meta p-value method in order to see both consensus and differential enrichment patterns. Alternatively, one can choose to use Fisher's exact test, once this method is selected, we need to pick the number of top ranked genes as the DE gene set for pathway analysis (Figure 23). For MAPE, the two tests are also available. There is one additional option of meta p-value method, available options include Fisher's method, maxP, minP, rOP and AW Fisher' method (Figure 24). When rOP is selected, additional option for order statistic of p-value (rth ordered p-value) will be available. Instead of closed-form distributions, users can use permutation of gene labels to get pathway p-value forthe Kolmogorov-Smirnov test (Figure 24). When "Permutation to get p-value" is choosen to be "YES", additional option for number of permutations will be available. However, we generally suggest users not to run permutation because it may take up to hours to compute. Lastly, we will specify the minimum/maximum pathway size of pathways we wish to include for functional analysis, then click "Step 1: Run Pathway Analysis".

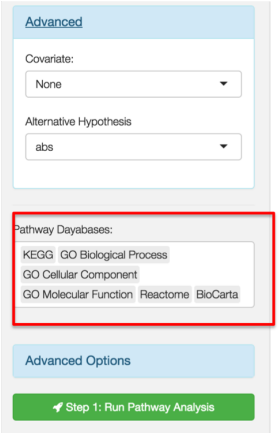


Figure 21: Selection of pathway database

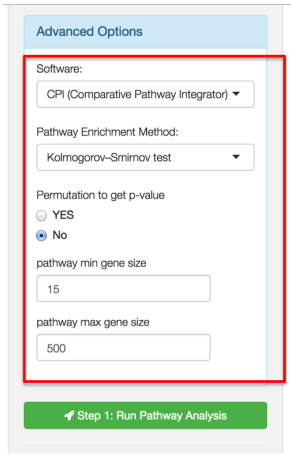


Figure 22: CPI setting

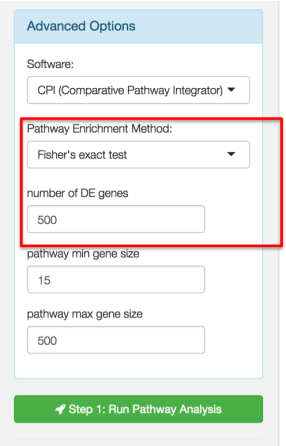


Figure 23: Fisher's Exact Test

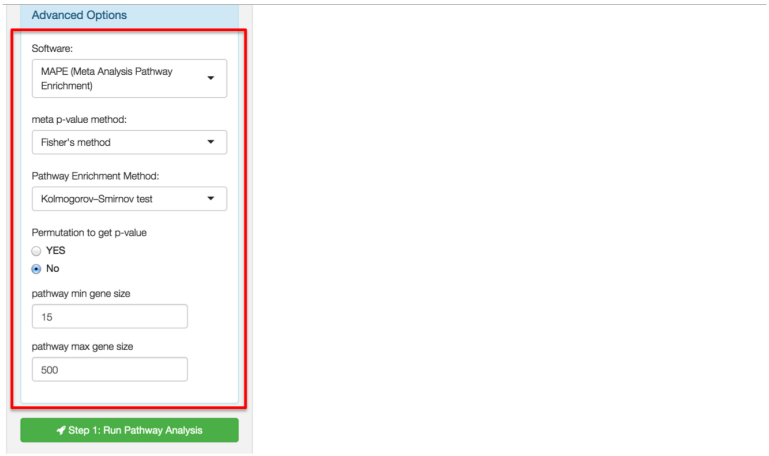


Figure 24: MAPE setting

We will see a summary table generated on the right of the page as shown in Figure 25 (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.

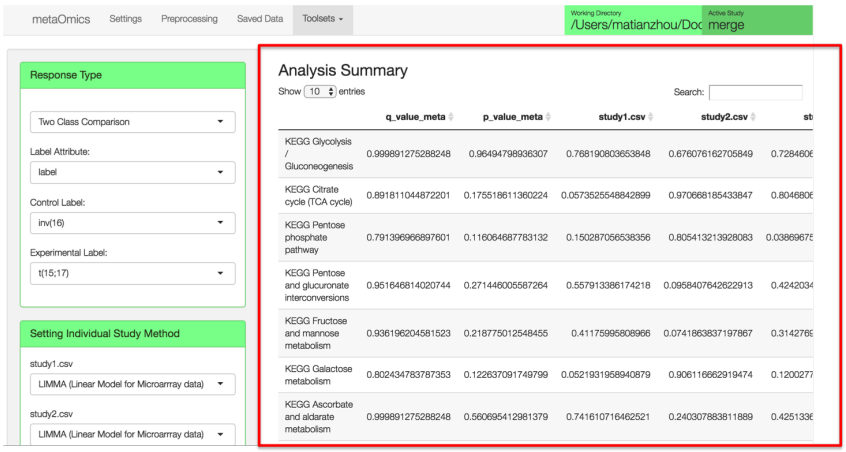


Figure 25: Summary table of CPI analysis results

4.3.2 Pathway clustering

In the next step, we performed clustering diagnostics of pathways at specified FDR cutoff (Figure 26). After clicking on "Pathway Cluster Diagnostics", we will see two plots generated on the right panel: consensus CDF and Delta area plots (Figure 27), we can use these two plots to determine the optimal number of clusters, to be used in the third clustering step.

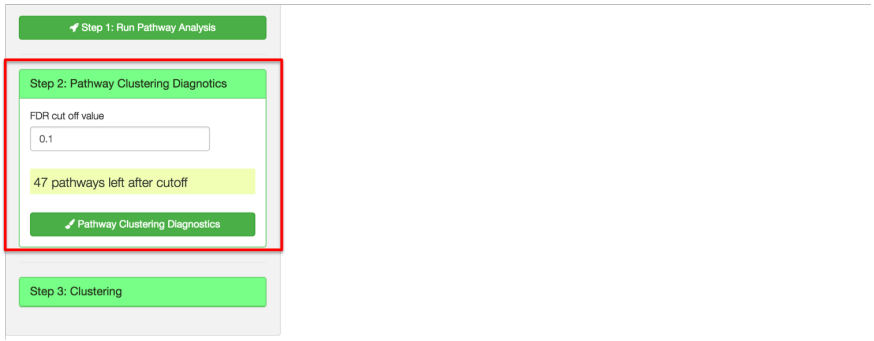


Figure 26: Pathway clustering diagnostics

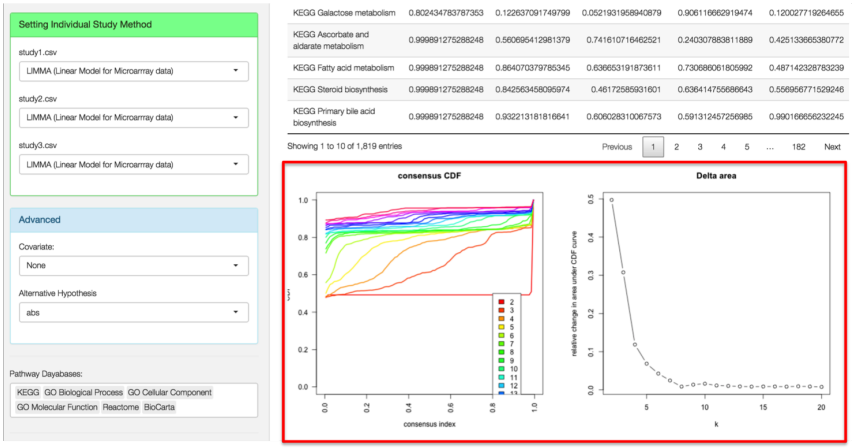


Figure 27: Diagnostic plots

4.4 MetaClust

By clicking toolsets and then metaClust, users are directed to metaClust home page as Figure 28.

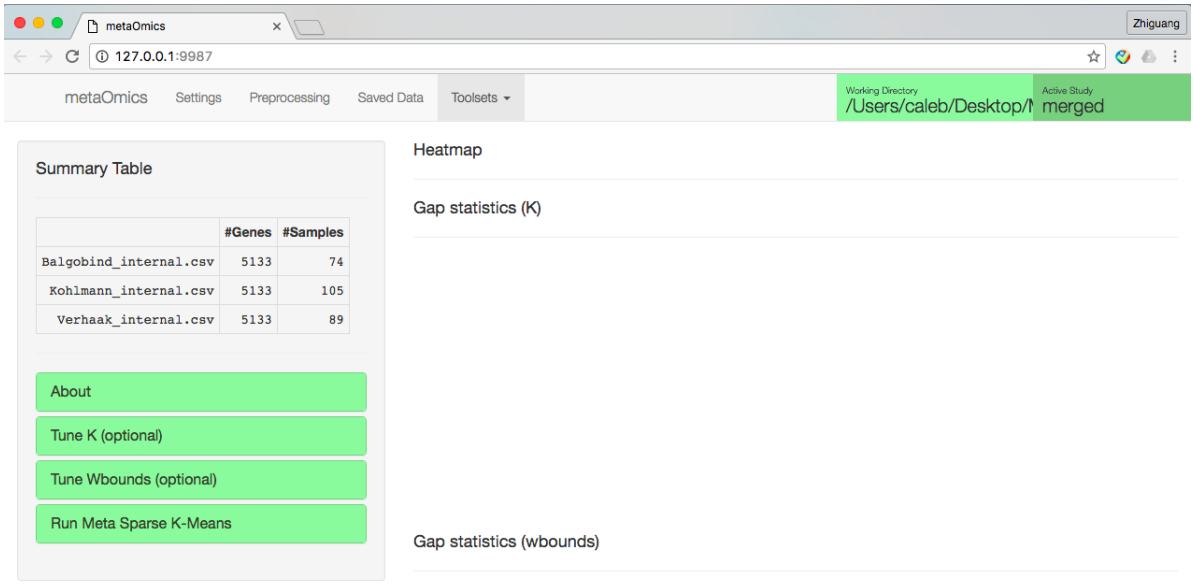


Figure 28: GUI Preprocessing page

On the top left panel users can see data summary Table. Below there are 4 tabs.

4.4.1 About

About tab includes basic introduction of metaClust. Starting with multiple studies, we could run MetaSparseKmeans 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 and Tune Wbounds panel.

4.4.2 Tune K

If the users are not sure what is number of clusters, they can start to use the Tune K panel as in Figure 29.

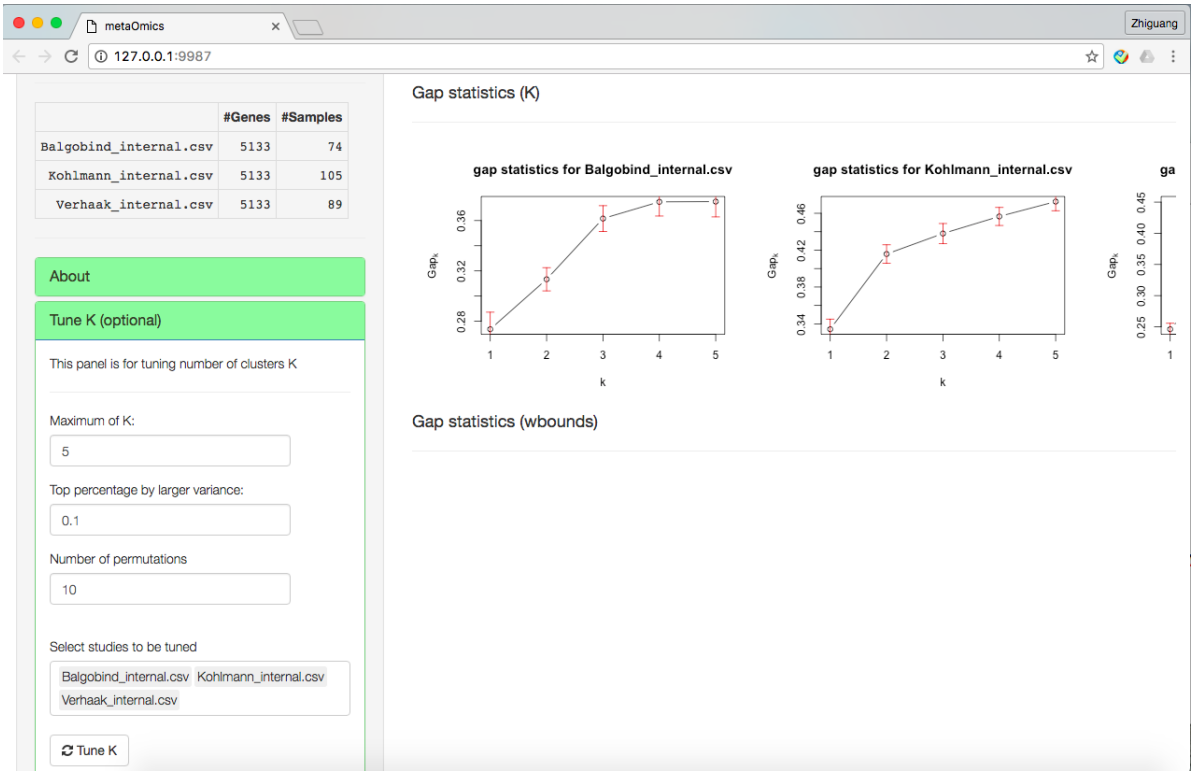


Figure 29: GUI Preprocessing page

Users will use gap statistics to get optimal K for each individual study. Users need to specify maximum number of K, 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. Number of permutation is number of bootstrap samples for gap statistics. After selecting studies to be tuned and clicking button "Tune K", we will obtain gap statistics plat as in Figure 29. A good K is selected such that the Gap_k is maximized or stablized. From the figure, K=3 is prefered.

4.4.3 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 30.

Figure 30: GUI Preprocessing page

Again, gap statistics will be used for tuning Wbounds. Users will specify number of clusters for tuning Wbounds, which could be obtained from the previous step. Iterations 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. After all these steps are set, user can click on "Tune Wbounds" button. The results will be shown in Figure 30. Wbound=12 is preferred since the corresponding gap statistics is maximized.

4.4.4 Run Meta Sparse K-Means

Under Run Meta Sparse K-Means panel, user can specify number of clusters, Wbounds and run meta sparse K means, as in Figure 31.

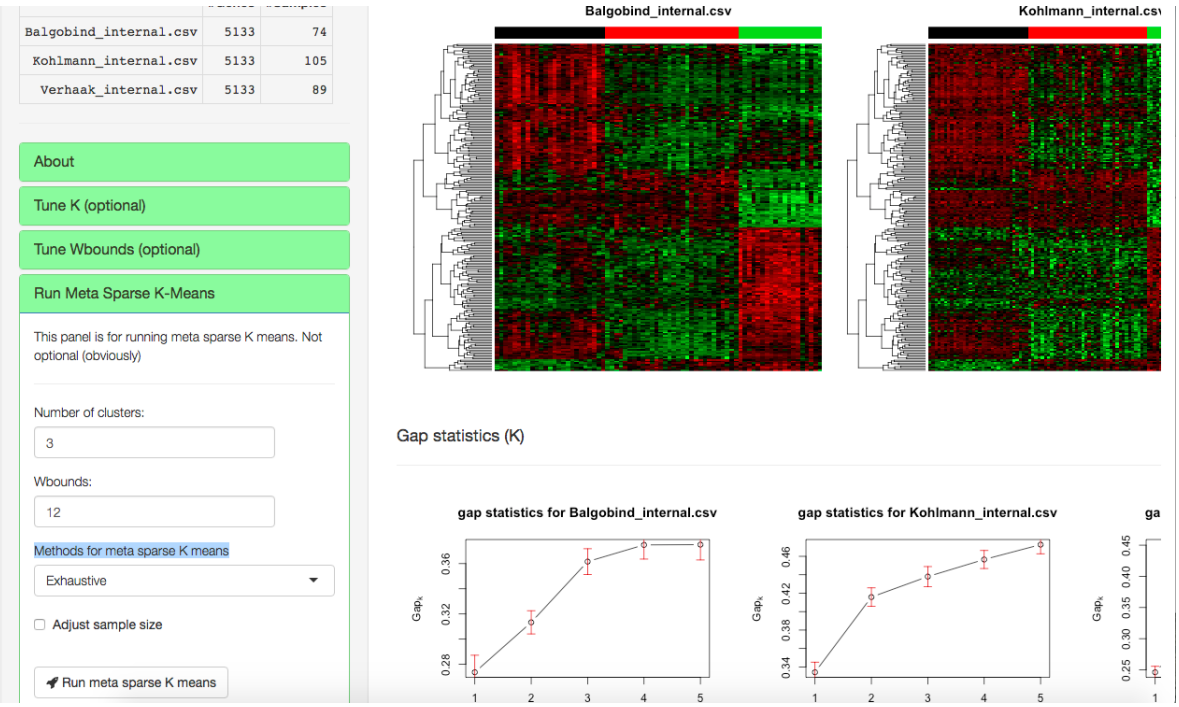


Figure 31: GUI Preprocessing page

There are three clustering matching methods: 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 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 31.

4.5 MetaPCA

4.6 MetaKTSP

4.7 MetaDCN

4.8 MetaLA

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.

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.

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.