Metabolomic Data Analysis with MetaboAnalystR

Name: guest7645290568106381084

February 7, 2018

1 Background

The combination of multiple independent metabolomics studies investigating the same condition in similar populations, which is often termed âĂIJhorizontal integrationâĂİ, or âĂIJmetabolomic meta-analysisâĂİ. The aim of metabolomic meta-analysis is to leverage the collective power of multiple studies to overcome potential noise, bias, and small effect sizes to improve the precision in identifying true patterns within data. Specifically, biomarker identification remains a large area of research in metabolomics, and their validation is challenging due to inconsistencies in identified biomarkers amongst similar experiments. Performing meta-analysis across similar studies will thereby increase the sample size and the power to identify robust and precise biomarkers of disease. The aim of the Meta-Analysis module for the integration of individual metabolomic studies to identify consistent and robust biomarkers of disease. This module supports three methods for performing meta-analysis: 1) Combining p-values, 2) Vote counting, and 3) Direct merging of data into a mega-dataset.

2 Meta-Analysis Overview

The Meta-Analysis module consists of six steps: 1) uploading the individual datasets; 2) data processing of each individual dataset, however it is suggested that the data-processing steps are consistent amongst the studies; 3) differential expression analysis of individual datasets; 4) data integrity check prior to meta-analysis ; 5) selection of the statistical method for meta-analysis, and 6) visualization of results as a Venn diagram to view all possible combinations of shared features between the datasets.

3 Data Input

The Meta-Analysis module accepts individual datasets which must be prepared by users prior to being uploaded. In general, the datasets must have been collected under comparable experimental conditions/share the same hypothesis or have the same mechanistic underpinnings. At the moment, the module only supports two-group comparisons (ex: control vs disease). Further, the module accepts either a compound concentration table, spectral binned data, or a peak intensity table. The format of the data must be specified, identifying whether the samples are in rows or columns, or may either be .csv or .txt files.

3.0.1 Sanity Check

Before data analysis, a sanity check is performed to make sure that all of the necessary information has been collected. The class labels must be present and must contain only two classes. If the samples are paired, the class label must be from -n/2 to -1 for one group, and 1 to n/2 for the second group (n is the sample number and must be an even number). Class labels with the same absolute value are assumed to be pairs. Compound concentration or peak intensity values must all be non-negative numbers. By

default, all missing values, zeros and negative values will be replaced by the half of the minimum positive value found within the data (see next section).

3.0.2 Normalization of Individual Data

Before differential expression analysis, datasets may be normalized using Log2 transformation. Additionally, users may choose to auto-scale their data. No normalization methods were applied. Autoscaling of data was performed.

3.0.3 Differential Expression Analysis of Individual Data

Before meta-analysis, differential expression analysis using linear models (Limma) may be performed for exploratory analysis. Here, users must specify the p-value (FDR) cut-off and the fold-change (FC) cutoff. No differential-expression analysis was performed.

3.0.4 Data Integrity Check

Before meta-analysis, one final data integrity check is performed to ensure meta-data are consistent between datasets and that there are at least more than 25 percent common features between the collective datasets.The following is information about your uploaded dataset for meta-analysis: Sample : 337 Common ID : 137 Condition: Adenocarcinoma vs. Control

3.1 Meta-Analysis Output

After the data has passed the final integrity check, users have the option to select one of three methods to perform meta-analysis: 1) Combining p-values, 2) vote counting, or 3) directly mering the datasets into a mega-dataset.

3.2 Combining P-Values

Calculating and combining p-values for the meta-analysis of microarray studies has been a long standing method and now we apply it to metabolomics studies. It includes two popular approaches, the Fisher's method and the Stouffer's method, which have similar levels of performance and are generally interpreted as larger scores reflecting greater differential abundance. The main difference between the two methods are weights (which are based on sample size), which are used in the Stouffer's method but not used in the Fisher's method. It should be noted that larger sample sizes do not warrant larger weights, as study quality can be variable. Further, users should use the Stouffer's method only when all studies are of similar quality.

3.3 Vote Counting

Vote counting is considered the most simple yet most intuitive method for meta-analysis. Here, significant features are selected based on a selected criteria (i.e. an adjusted p-value <0.05 and the same direction of FC) for each dataset. The votes are then calculated for each feature by counting the total of number of times a feature is significant across all included datasets. However, this method is statistically inefficient and should be considered the last resort in situations where other methods to perform meta-analysis cannot be applied.

3.4 Direct Merging

The final method of meta-analysis is the direct merging of individual data into a mega-dataset, which results in an analysis of that mega-dataset as if the individual data were derived from the same experiment. This method thereby ignores any inherent bias and heterogeneity between the different data. Because of this, there exists several confounders such as different experimental protocols, technical platforms, and raw data processing procedures that can mask true underlying differences. It is therefore highly suggested that this approach be used only when individual data are very similar (i.e. from the same lab, same platform, without batch effects).

P-value combination was the selected method to perform meta-analysis. The method of p-value combination used is: fisher The p-value significance threshold is: 0.05

Table 1: Predicted top-ranking features from meta-analysis

	V1	V2	V3	V4	CombinedTstat	CombinedPval
adenosine-5-phosphate	-1.75	-0.82	-0.89	-1.68	-187.97	0.00
pyrophosphate	-1.65	-0.69	-1.01	-1.62	-174.04	0.00
pyruvic acid	-1.73	-0.02	-1.15	-0.18	-123.22	0.00
maltotriose	-0.57	-0.78	-0.62	-0.34	-45.97	0.00
glutamine	0.25	0.58	0.92	0.23	42.53	0.00
lactamide	-0.16	-0.34	-0.99	-0.14	-37.84	0.00
citrulline	0.17	0.71	0.65	0.23	34.70	0.00
lactic acid	-0.05	-0.14	-1.04	0.01	-34.79	0.00
alpha ketoglutaric acid	-0.52	-0.28	-0.58	-0.40	-32.54	0.00
cystine	0.22	0.81	0.38	0.21	31.32	0.00
taurine	0.01	-0.21	-0.89	-0.27	-31.39	0.00
maltose	-0.36	-0.70	-0.38	-0.22	-30.37	0.00
fructose	0.55	0.29	0.60	0.12	29.70	0.00
asparagine	0.26	0.37	0.67	0.28	29.05	0.00
oxalic acid	-0.25	-0.55	-0.44	-0.32	-27.22	0.01
hippuric acid	-0.01	-0.32	-0.78	-0.11	-26.83	0.01
histidine	-0.10	0.67	0.40	0.24	24.77	0.01
lauric acid	0.03	-0.24	-0.61	-0.44	-24.72	0.01
tagatose	0.49	0.09	0.58	-0.04	24.06	0.02
inosine	0.37	0.40	0.49	0.08	24.17	0.02
glutamic acid	-0.00	-0.36	-0.61	-0.22	-23.31	0.02
arachidonic acid	0.57	-0.27	0.51	0.12	23.37	0.02
ethanolamine	-0.37	-0.36	-0.49	-0.07	-23.25	0.02
lysine	0.14	0.62	0.34	0.18	23.30	0.02
3-phosphoglycerate	-0.05	-0.82	0.02	-0.08	-22.98	0.02
5-hydroxynorvaline NIST	0.36	0.18	0.62	-0.33	22.17	0.02
caprylic acid	0.58	0.07	0.46	-0.31	21.99	0.03
tryptophan	0.01	0.66	0.13	0.25	20.95	0.04
nicotinic acid	-0.13	-0.41	-0.49	-0.08	-20.58	0.04
parabanic acid NIST	-0.27	-0.40	-0.25	-0.33	-20.63	0.04
xanthine	0.16	-0.35	-0.54	-0.23	-20.18	0.04
isothreonic acid	0.43	0.14	0.51	-0.34	19.99	0.04
N-acetylglutamate	-0.13	-0.34	-0.58	0.24	-19.90	0.04
glycerol-alpha-phosphate	0.50	-0.54	0.44	0.12	19.80	0.04

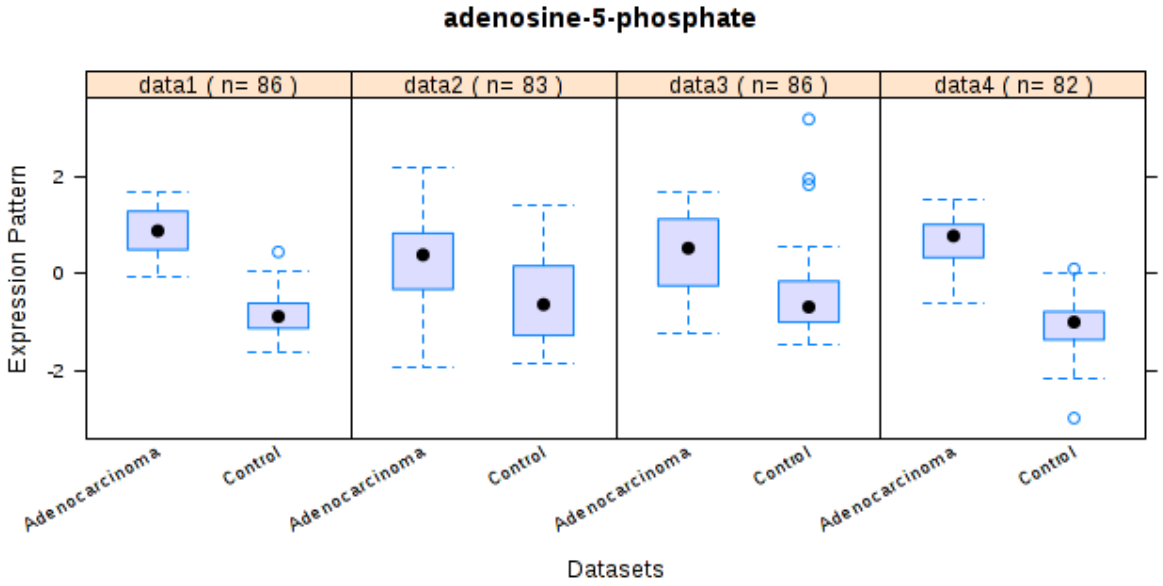

Figure 1: Box plot of the expression pattern of the selected feature between the two experimental groups across all studies. The expression pattern is on the y-axis, and the group labels are on the x-axis. The median expression for the feature is indicated with a black dot in the centre of the boxplot.
Selected feature: adenosine-5-phosphate

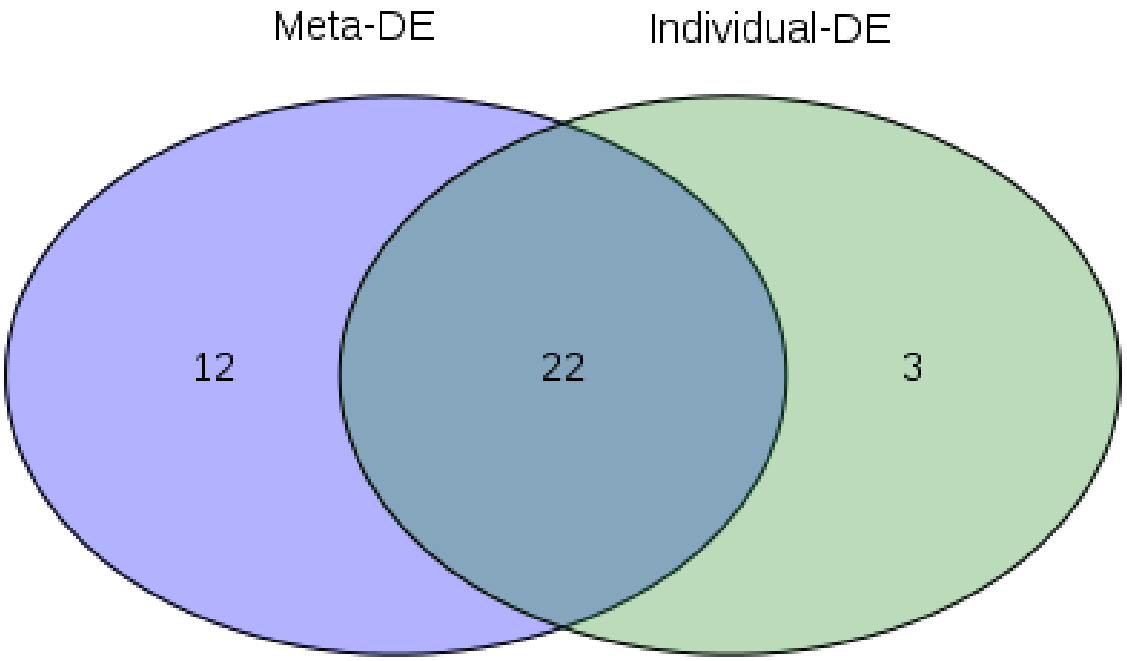

Figure 2: Venn diagram of the top differentially expressed features from the meta-analysis. On the left side are features that are DE only from the meta-analysis, in the center are DE features that were identified in both the meta-analysis and the individual studies, and on the right side are features that were DE in the individual analysis, but did not show up as DE during meta-analysis.

Table 2: Differentially expressed features by individual study and from meta-analysis

	Sum of DE Features	Names of DE Features
data1.csv	3	adenosine-5-phosphate, pyruvic acid, pyrophosphate
data2.csv	8	adenosine-5-phosphate, 3-phosphoglycerate, cystine, maltotriose, citrulline, maltose, pyrophosphate, histidine
data3.csv	21	pyruvic acid, lactic acid, pyrophosphate, lactamide, glutamine, adenosine-5-phosphate, taurine, hippuric acid, asparagine, salicylic acid, citrulline, maltotriose, aspartic acid, 5-hydroxynorvaline NIST, lauric
data4.csv	2	adenosine-5-phosphate, pyrophosphate
Meta-Analysis	12	oxalic acid, inosine, arachidonic acid, ethanolamine, lysine, caprylic acid, tryptophan, nicotinic acid, parabanic acid NIST, xanthine, isothreonic acid, glycerol-alpha-phosphate

4 Appendix: R Command History

```
 [1] "InitDataObjects(\"conc\", \"metadata\", FALSE)"
 [2] "mSet<-ReadIndData(mSet, \"data1.csv\", \"colu\");"
 [3] "mSet<-SanityCheckIndData(mSet, \"data1.csv\")"
 [4] "mSet<-PerformIndNormalization(mSet, \"data1.csv\", \"log\", 1);"
 [5] "mSet<-PerformLimmaDE(mSet, \"data1.csv\", 0.05, 0.0);"
 [6] "mSet<-ReadIndData(mSet, \"data2.csv\", \"colu\");"
 [7] "mSet<-SanityCheckIndData(mSet, \"data2.csv\")"
 [8] "mSet<-PerformIndNormalization(mSet, \"data2.csv\", \"log\", 1);"
 [9] "mSet<-PerformLimmaDE(mSet, \"data2.csv\", 0.05, 0.0);"
[10] "mSet<-ReadIndData(mSet, \"data3.csv\", \"colu\");"
[11] "mSet<-SanityCheckIndData(mSet, \"data3.csv\")"
[12] "mSet<-PerformIndNormalization(mSet, \"data3.csv\", \"log\", 1);"
[13] "mSet<-PerformLimmaDE(mSet, \"data3.csv\", 0.05, 0.0);"
[14] "mSet<-ReadIndData(mSet, \"data4.csv\", \"colu\");"
[15] "mSet<-SanityCheckIndData(mSet, \"data4.csv\")"
[16] "mSet<-PerformIndNormalization(mSet, \"data4.csv\", \"log\", 1);"
[17] "mSet<-PerformLimmaDE(mSet, \"data4.csv\", 0.05, 0.0);"
[18] "mSet<-CheckMetaDataConsistency(mSet, F);"
[19] "mSet<-PerformPvalCombination(mSet, \"fisher\", 0.05)"
[20] "mSet<-GetMetaResultMatrix(mSet, \"fc\")"
[21] "mSet<-PlotSelectedFeature(mSet, \"adenosine-5-phosphate\")"
[22] "mSet<-PrepareVennData(mSet);"
[23] "mSet<-GetSelectedDataNumber(mSet);"
[24] "mSet<-GetSelectedDataNames(mSet);"
[25] "mSet<-SaveTransformedData(mSet)"
```

The report was generated on Wed Feb 7 16:11:22 2018 with R version 3.4.1 (2017-06-30).