Metabolomic Data Analysis with MetaboAnalyst

User ID: guest6501

April 16, 2009

1 Data Processing and Normalization

1.1 Reading and Processing the Raw Data

MetaboAnalyst accepts a variety of data types generated in metabolomic studies, including compound concentration data, binned NMR/MS spectra data, NMR/MS peak list data, as well as MS spectra (NetCDF, mzXML, mzDATA). Users need to specify the data types when uploading their data in order for MetaboAnalyst to select the correct algorithm to process them.

The R scripts `datautils.R` and `processing.R` are required to read in and process the uploaded data.

1.1.1 Reading Concentration Data

The concentration data should be uploaded in comma separated values (.csv) format. Samples can be in rows or columns, with class labels immediately following the sample IDs.

The uploaded file is in comma separated values (.csv) format. Samples are in columns and features in rows. The uploaded data file contains 14 (samples) by 42 (compounds) data matrix.

1.1.2 Data Integrity Check

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

1.1.3 Missing value imputations

Too many zeroes or missing values will cause difficulties for downstream analysis. MetaboAnalyst offers several different methods for this purpose. The default method replaces all the missing and zero values with a small values (the half of the minimum positive values in the original data) assuming to be the detection limit. The assumption of this approach is that most missing values are caused by low abundance metabolites (i.e.below the detection limit). In addition, since zero values may cause problem for data normalization (i.e. log), they are also replaced with this small value. User can also specify other methods, such as replace by mean/median, or use Probabilistic PCA (PPCA), Bayesian PCA (BPCA) method, Singular Value Decomposition (SVD) method to impute the missing values [1]. Please choose the one that is the most appropriate for your data. Table 1 summarizes the result of the data processing steps.

Missing variables were replaced with a small value: 3.1

Table 1: Summary of data processing results

	Features (positive)	Missing/Zero	Features (processed)
87_day4	24	18	42
143_day4_2of2	29	13	42
143_day4_2of5	29	13	42
163_day4	30	12	42
225_day4	33	9	42
239_day4	27	15	42
241_day4	26	16	42
c87_day1	18	24	42
143_day1_1of4	28	14	42
143_day1_1of5	29	13	42
163_day1	30	12	42
225_day1	29	13	42
239_day1	28	14	42
241_day1	28	14	42

[1] Stacklies W, Redestig H, Scholz M, Walther D, Selbig J. *pcaMethods: a bioconductor package, providing PCA methods for incomplete data.*, Bioinformatics 2007 23(9):1164-1167

1.2 Data Normalization

The data is stored as a table with one sample per row and one variable (bin/ peak/metabolite) per column. There are two types of normalization. Row-wise normalization aims to bring each sample (row) comparable to each other (i.e. urine samples with different dilution effects). Column-wise normalization aims to make each variable (column) comparable to each other within the same sample. The procedure is useful when variables are of very different orders of magnitude.

The normalization consists of the following options:

1. Row-wise normalization:
   - Normalization by the sum
   - Normalization by a reference sample (probabilistic quotient normalization) [2]
   - Normalization by a reference feature (i.e. creatinine, internal control)
   - Sample specific normalization (i.e. normalize by dry weight, volume)
2. Column-wise normalization :
   - Log transformation (log 2)
   - Unit scaling (mean-centered and divided by standard deviation of each variable)
   - Pareto scaling (mean-centered and divided by the square root of standard deviation of each variable)
   - Range scaling (mean-centered and divided by the value range of each variable)

The R script `normalization.R` is required. Figure 1 shows the effects before and after normalization.

[2] Dieterle F, Ross A, Schlotterbeck G, Senn H. *Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics*, 2006, Anal Chem 78 (13);4281 - 4290

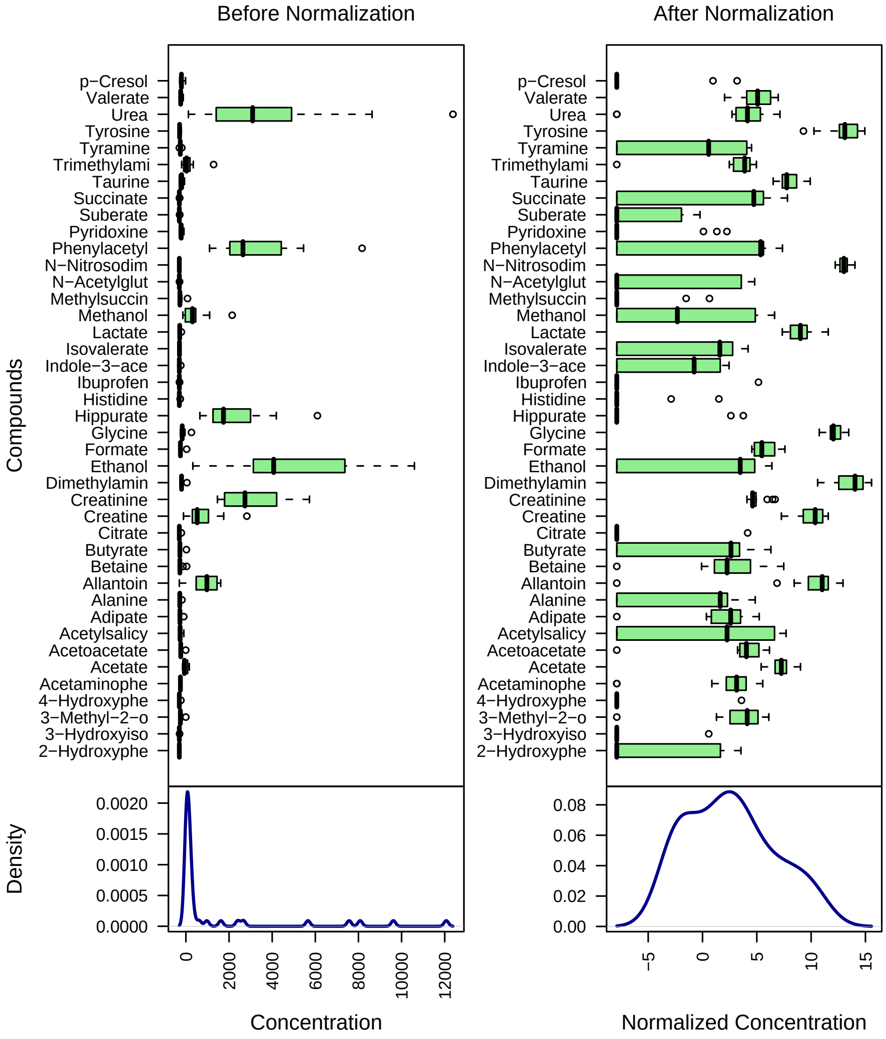

Figure 1: Box plots and kernel density plots before and after normalization. The boxplots show at most 80 features due to space limit. The density plots are based on all samples. Row-wise normalization: Normalization by a reference feature Column-wise normalization: Log Normalization.

2 Statistical and Machine Learning Data Analysis

MetaboAnalyst offers a variety of methods commonly used in metabolomic data analyses. They include:

1. Univariate analysis methods:
    - Fold Change Analysis
    - T-tests
    - Volcano Plot
2. Dimensional Reduction methods:
    - Principal Component Analysis (PCA)
    - Partial Least Squares - Discriminant Analysis (PLS-DA)
3. Robust Feature Selection Methods in microarray studies
    - Significance Analysis of Microarray (SAM)
    - Empirical Bayesian Analysis of Microarray (EBAM)
4. Clustering Analysis
    - Hierarchical Clustering
        - Dendrogram
        - Heatmap
    - Partitional Clustering
        - K-means Clustering
        - Self-Organizing Map (SOM)
5. Supervised Classification and Feature Selection methods
    - Random Forest
    - Support Vector Machine (SVM)

2.1 Univariate Analysis

There are three methods available for univariate analyses, including Fold Change (FC) analysis, t-tests, and volcano plot which is a combination of the first two methods. All three these methods support both unpaired and paired analyses. They provide a preliminary overview about features that are potentially significant in discriminating the two groups under study.

For paired fold change analysis, the algorithm first counts the total number of pairs with fold changes that are consistently above/below the specified FC threshold for each variable. A variable will be reported as significant if this number is above a given count threshold (default > 75% of pairs/variable)

The R script `univartests.R` is required. Figure 2 shows the important features identified by fold change analysis. Table 2 shows the details of these features; Figure 3 shows the important features identified by t-tests. Table 3 shows the details of these features; Figure 4 shows the important features identified by volcano plot. Table 4 shows the details of these features.

Please note, the purpose of fold change is to compare absolute value changes between two group means. Therefore, the data before column normlaization will be used instead. Also note, the result is plotted in log2 scale, so that same fold change (up/down-regulated) will have the same distance to the zero baseline.

Table 2: Important features identified by fold change analysis

	Compounds	Count (up)	Count (down)
1	Tyramine	6	0
2	N-Nitrosodimethylamine	5	0
3	Pyridoxine	5	0
4	Trimethylamine N-oxide	6	1
5	Acetylsalicylate	5	1
6	p-Cresol	5	1
7	Acetaminophen	5	2
8	Methanol	2	5

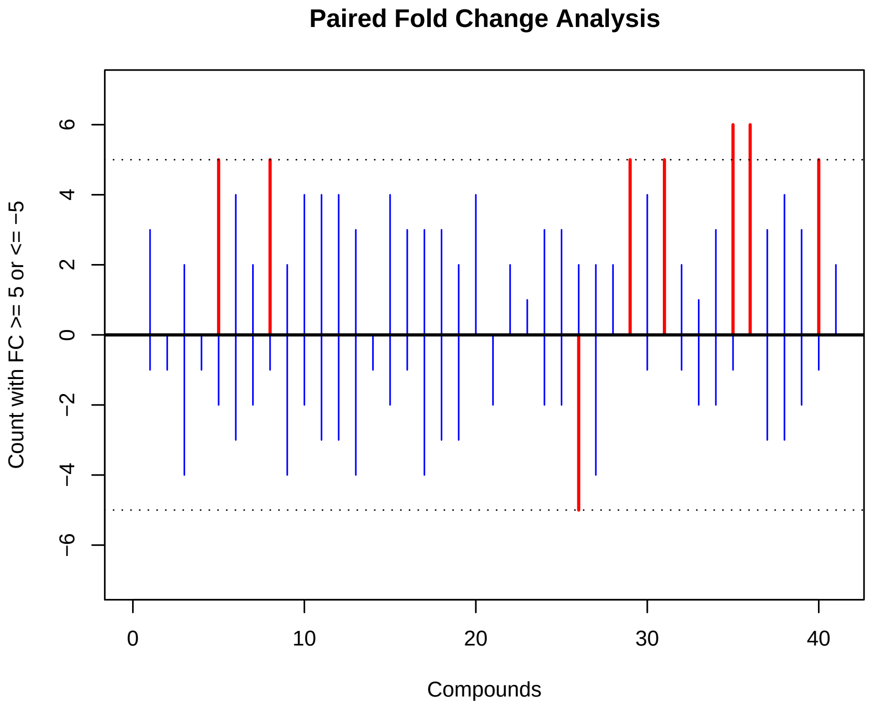

Figure 2: Important features selected by fold-change analysis with threshold 1.1. The red circles represent features above the threshold. Note the values are on log scale, so that both up-regulated and downregulated features can be plotted in a symmetrical way

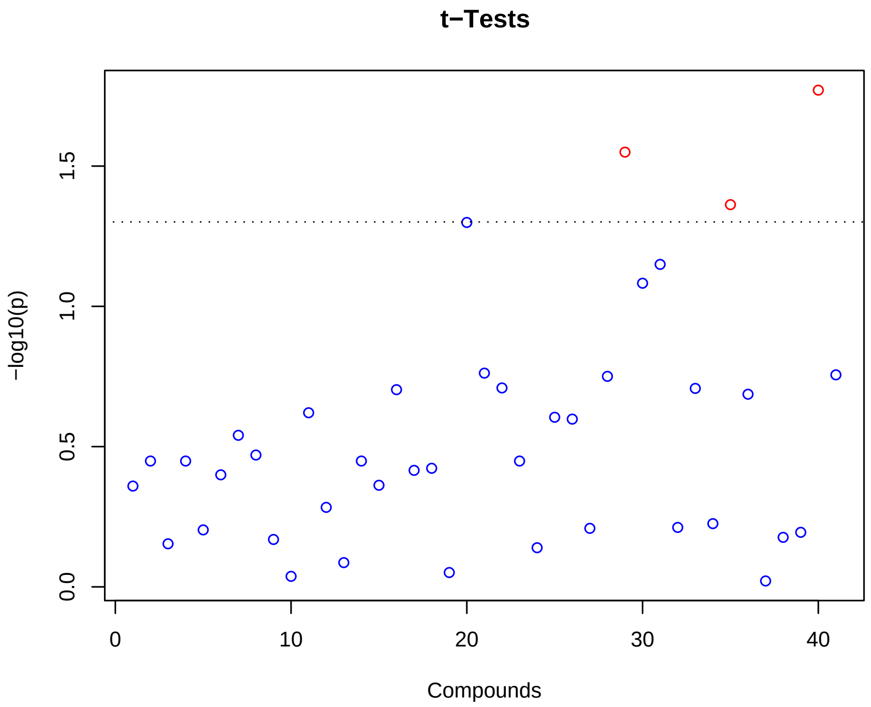

Figure 3: Important features selected by t-tests with threshold 0.05. The red circles represent features above the threshold. Note the p values are transformed by -log10 so that the more significant features (with smaller p values) will be plotted higher on the graph.

Table 3: Important features identified by t-tests

	Compounds	p.value	-log10(p)
1	p-Cresol	0.01695	1.77079
2	N-Nitrosodimethylamine	0.02819	1.54989
3	Trimethylamine N-oxide	0.04339	1.36257

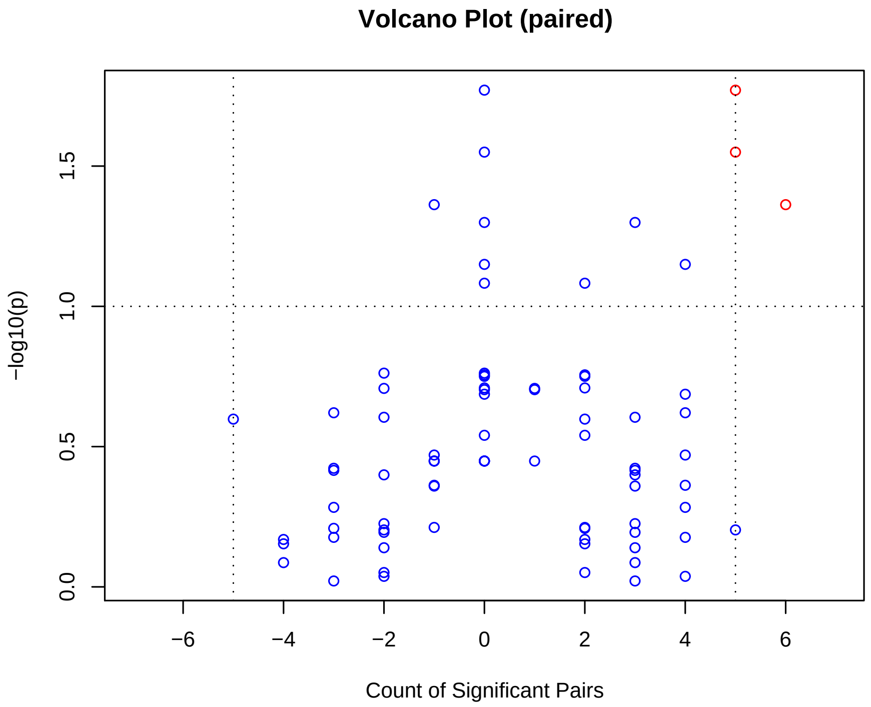

Figure 4: Important features selected by volcano plot with fold change threshold (x) 1.2 and t-tests threshold (y) 0.1. The red circles represent features above the threshold. Note both fold changes and p values are log transformed. The further its position away from the (0,0), the more significant the feature is.

Table 4: Important features identified by volcano plot

	Compounds	Counts (up)	Counts (down)	p.value	-log10(p)
1	p-Cresol	5	0	0.017	1.771
2	N-Nitrosodimethylamine	5	0	0.028	1.55
3	Trimethylamine N-oxide	6	1	0.043	1.363

2.2 Significance Analysis of Microarray (SAM)

SAM is a well-established statistical method for identification of differentially expressed genes in microarray data analysis. It is designed to address the false discovery rate (FDR) when running multiple tests on high-dimensional microarray data. SAM assigns a significance score to each variable based on its change relative to the standard deviation of repeated measurements. For a variable with scores greater than an adjustable threshold, its relative difference is compared to the distribution estimated by random permutations of the class labels. For each threshold, a certain proportion of the variables in the permutation set will be found to be significant by chance. The proportion is used to calculate the FDR. SAM is performed using the `siggenes` package[3]. Users need to specify the `Delta` value to control FDR in order to proceed.

The R script `sigfeatures.R` is required. Figure 5 shows the significant features identified by SAM. Table 5 shows the details of these features.

Table 5: Important features identified by SAM

	Compounds	d.value	stdev	rawp	q.value	R.fold
1	p-Cresol	-3.274	0.232	0.008	0.133	
2	N-Nitrosodimethylamine	-2.876	1.459	0.014	0.133	
3	Trimethylamine N-oxide	-2.552	0.197	0.027	0.137	
4	Hippurate	-2.444	0.113	0.03	0.137	
5	Pyridoxine	-2.192	1.81	0.045	0.166	
6	Phenylacetylglycine	-2.08	0.093	0.06	0.183	

[3] Holger Schwender. *siggenes: Multiple testing using SAM and Efron's empirical Bayes approaches*,2008, R package version 1.16.0

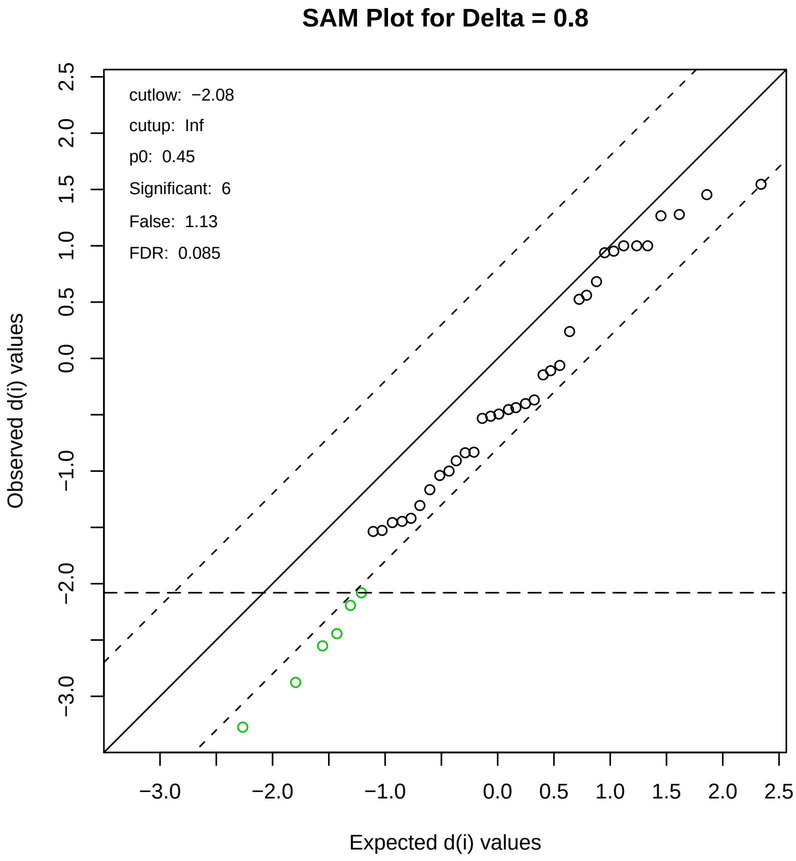

Figure 5: Significant features identified by SAM. The green circles represent features that exceed the specified threshold.

2.3 Empirical Bayesian Analysis of Microarray (EBAM)

EBAM is an empirical Bayesian method based on moderated t-statistics. EBAM uses a two-group mixture model for null and significant features. The prior and density parameters are estimated from the data. A feature is considered significant if its calculated posterior is larger than or equal to `delta` and no other features with a more extreme test score that is not called signicant. The default is `delta` = 0.9. The suggested fudge factor (`a0`) is chosen that leads to the largest number of significant features. EBAM is performed with `ebam` function in `siggenes` package[4].

The R script `sigfeatures.R` is required. Figure 6 shows the important features identified by EBAM. Table 6 shows the details of these features.

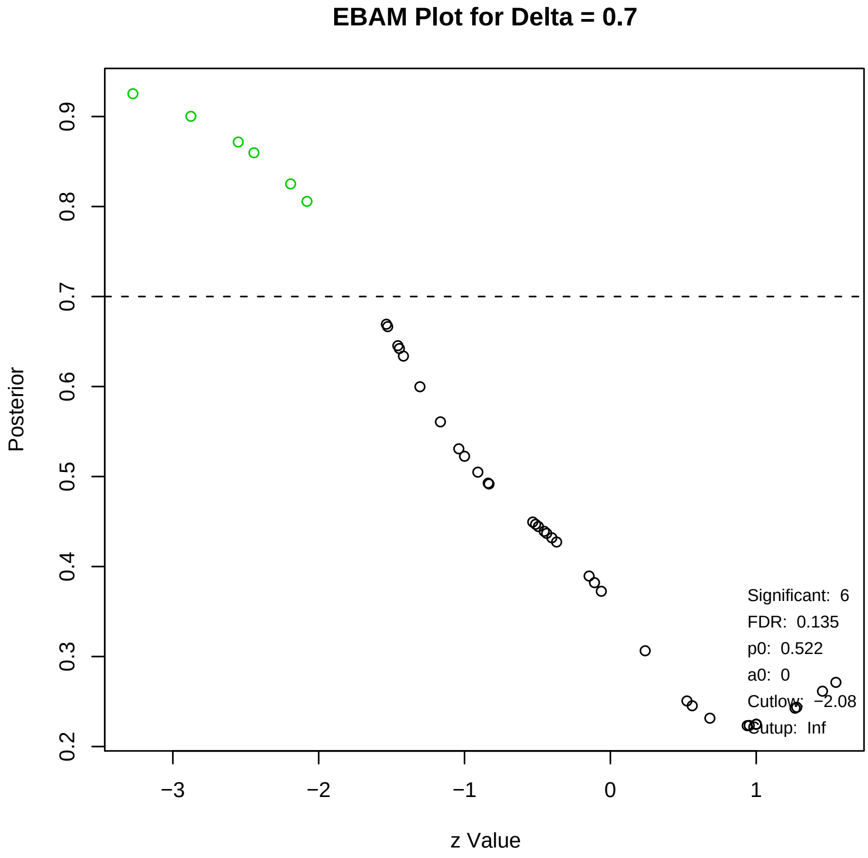

Figure 6: Significant features identified by EBAM. The green circles represent features that exceed the specified threshold.

[4] Holger Schwender. *siggenes: Multiple testing using SAM and Efron's empirical Bayes approaches*,2008,R package version 1.16.0

Table 6: Important features identified by EBAM

	Compounds	z.value	posterior	local.fdr
1	p-Cresol	-3.274	0.925	0.075
2	N-Nitrosodimethylamine	-2.876	0.9	0.1
3	Trimethylamine N-oxide	-2.552	0.872	0.128
4	Hippurate	-2.444	0.86	0.14
5	Pyridoxine	-2.192	0.825	0.175
6	Phenylacetylglycine	-2.08	0.806	0.194

3 Data Annotation

Please be advised that MetaboAnalyst also supports metabolomic data annotation. For NMR, MS, or GC-MS peak list data, users can perform peak identification by searching the corresponding libraries. For compound concentration data, users can perform pathway mapping. These tasks require a lot of manual efforts and are not performed by default.

The report was generated on Thu Apr 16 10:31:23 2009 with R version 2.8.1 (2008-12-22) on a i386-redhat-linux-gnu platform. Thank you for using MetaboAnalyst! For suggestions and feedback please contact Jeff Xia (*jianguox@ualberta.ca*).