Metabolomic Data Analysis with MetaboAnalyst

User ID: guest4965371694680211620

April 14, 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 rows and features in columns The uploaded data file contains 57 (samples) by 54 (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: 0.5

[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

Table 1: Summary of data processing results

	Features (positive)	Missing/Zero	Features (processed)
PIF178	40	14	54
PIF087	40	14	54
PIF090	38	16	54
NETL5	45	9	54
PIF115	42	12	54
PIF110	41	13	54
NETL19	47	7	54
PIF108	34	20	54
PIF171	35	19	54
PIF154	39	15	54
PIF105	40	14	54
NETL8	40	14	54
PIF146	39	15	54
PIF119	38	16	54
PIF099	32	22	54
PIF160	40	14	54
PIF113	45	9	54
PIF137	39	15	54
NETL20	40	14	54
PIF100	32	22	54
NETCR12	45	9	54
PIF094	41	13	54
PIF132	47	7	54
NETL10	38	16	54
PIF163	36	18	54
NETCR10	42	12	54
NETCR3	26	28	54
NETCR9	42	12	54
NETCR13	42	12	54
CACH192	35	19	54
NETL11	42	12	54
PIF004	31	23	54
NETCR19	45	9	54
NETCR4	36	18	54
NETL23	47	7	54
NETCR14	43	11	54
NETCR21	43	11	54
NETL2	43	11	54
CACH191	35	19	54
PIF164	35	19	54
NETL13	35	19	54
CACH188	30	24	54
CACH195	35	19	54
NETL12	42	12	54
NETCR7	38	16	54
NETCR15	45	9	54
PIF102	39	15	54
NETL1	36	18	54
NETCR5	38	16	54
PIF111	40	14	54
PIF153	39	15	54
PIF143	37	17	54
CACH190	28	26	54
NETL7	38	16	54
PIF112	35	19	54
PIF162	32	22	54
NETL4	42	12	54

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

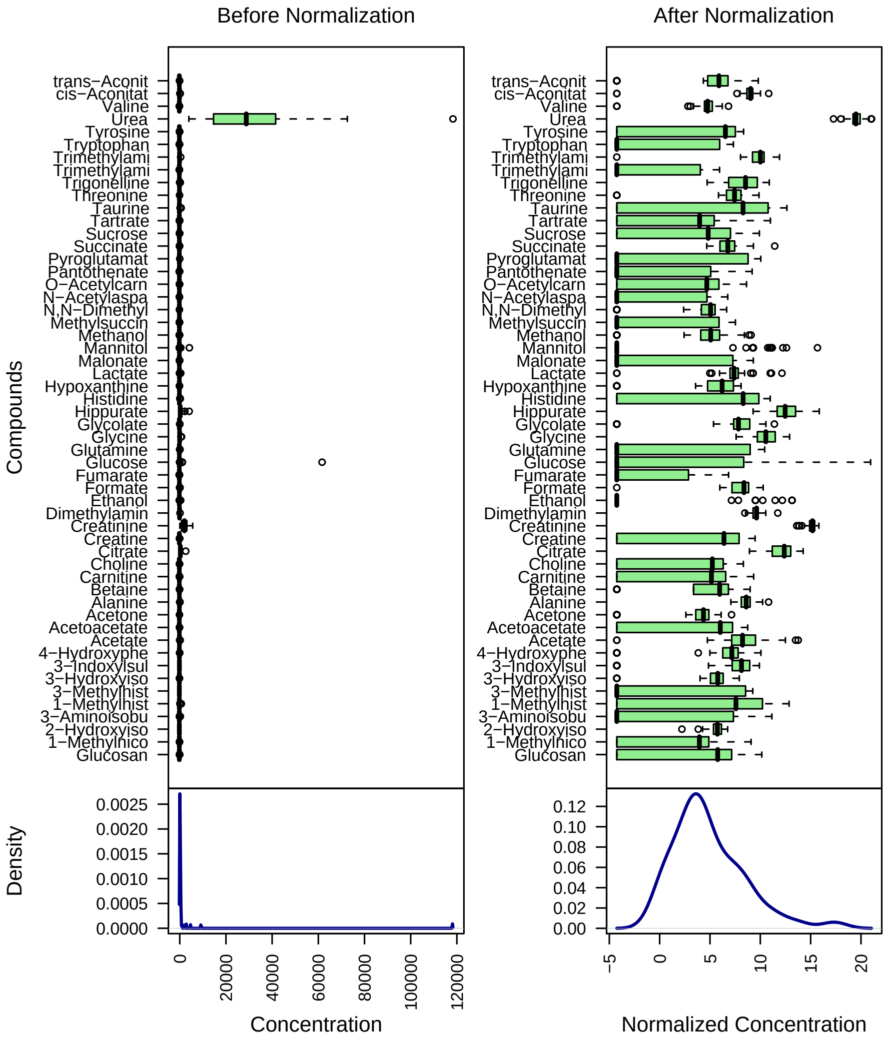

Figure 1: Box plots and kernel density plots before and after normalization. The boxplots show at most 80 features due to space limit. The density plots are based on all samples. Row-wise normalization: Probabilistic Quotient Normalization Column-wise normalization: Log Normalization.

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	Fold Change	log2(FC)
1	Glucose	0.0056	-7.4705
2	Fumarate	0.3103	-1.6882
3	Mannitol	0.3111	-1.6847
4	3-Methylhistidine	2.5432	1.3466
5	Malonate	0.4604	-1.1191
6	Acetoacetate	2.1161	1.0814
7	Pyroglutamate	0.4854	-1.0427
8	Lactate	0.4862	-1.0403
9	1-Methylnicotinamide	2.007	1.0051

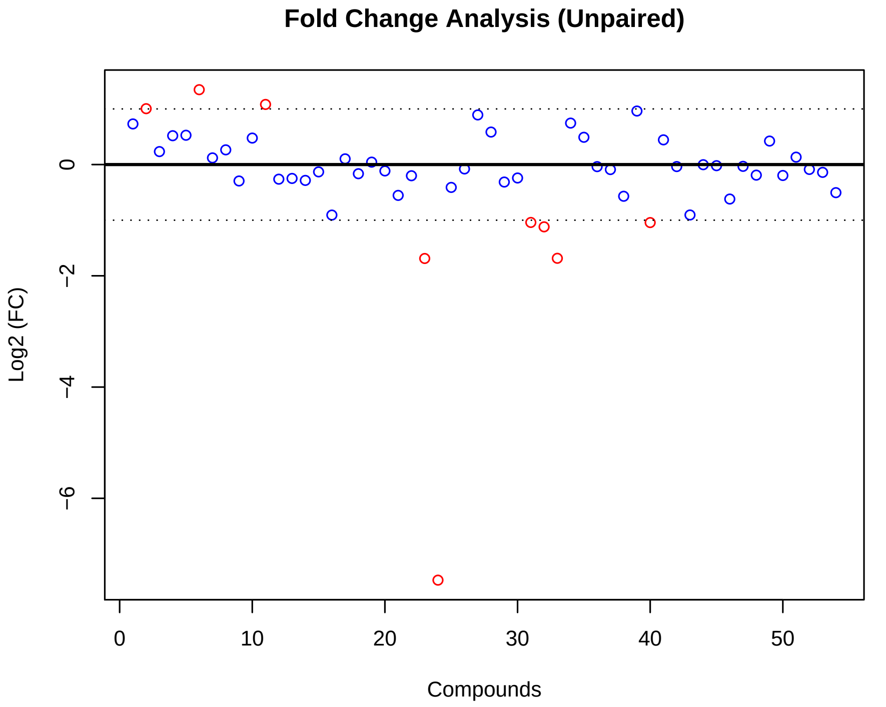

Figure 2: Important features selected by fold-change analysis with threshold 2. The red circles represent features above the threshold. Note the values are on log scale, so that both up-regulated and downregulated features can be plotted in a symmetrical way

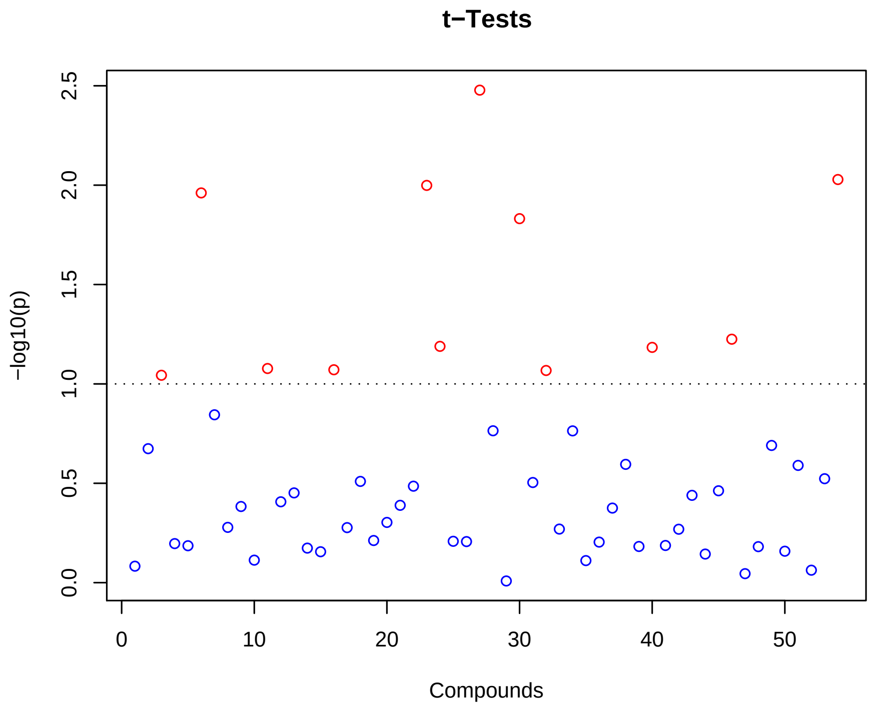

Figure 3: Important features selected by t-tests with threshold 0.1. The red circles represent features above the threshold. Note the p values are transformed by -log10 so that the more significant features (with smaller p values) will be plotted higher on the graph.

Table 3: Important features identified by t-tests

	Compounds	p.value	-log10(p)
1	Glycolate	0.00333	2.47814
2	trans-Aconitate	0.00937	2.02846
3	Fumarate	0.01003	1.99887
4	3-Methylhistidine	0.01094	1.96101
5	Hypoxanthine	0.01474	1.83148
6	Trigonelline	0.05956	1.22504
7	Glucose	0.06471	1.18901
8	Pyroglutamate	0.06548	1.18386
9	Acetoacetate	0.08363	1.07761
10	Choline	0.08486	1.0713
11	Malonate	0.08559	1.06759
12	2-Hydroxyisobutyrate	0.09045	1.04362

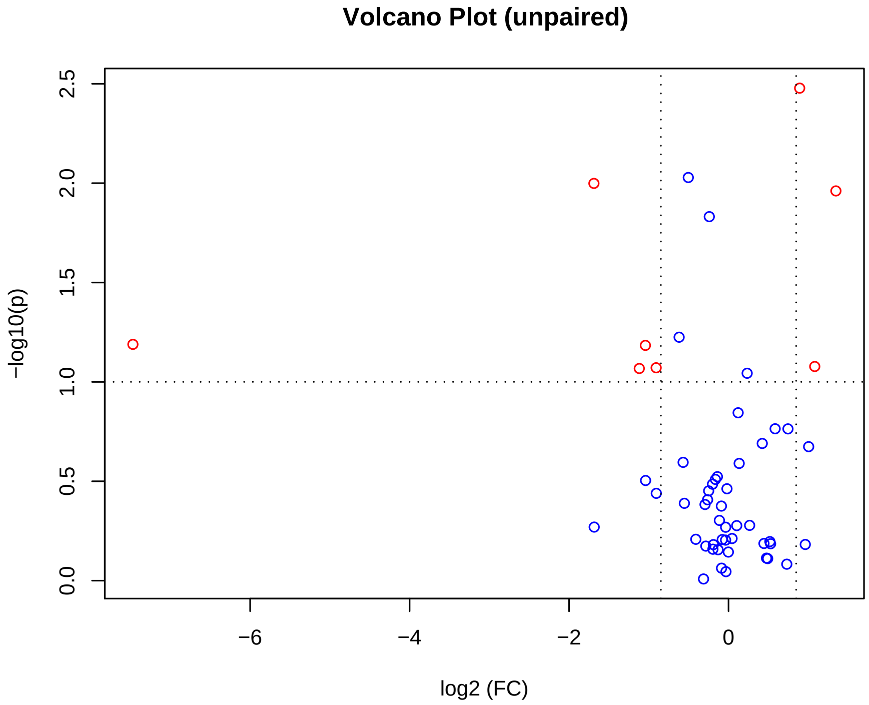

Figure 4: Important features selected by volcano plot with fold change threshold (x) 1.8 and t-tests threshold (y) 0.1. The red circles represent features above the threshold. Note both fold changes and p values are log transformed. The further its position away from the (0,0), the more significant the feature is.

Table 4: Important features identified by volcano plot

	Compounds	FC	log2(FC)	p.value	-log10(p)
1	Glycolate	1.856	0.893	0.003	2.478
2	Fumarate	0.31	-1.688	0.01	1.999
3	3-Methylhistidine	2.543	1.347	0.011	1.961
4	Glucose	0.006	-7.471	0.065	1.189
5	Pyroglutamate	0.485	-1.043	0.065	1.184
6	Acetoacetate	2.116	1.081	0.084	1.078
7	Choline	0.533	-0.908	0.085	1.071
8	Malonate	0.46	-1.119	0.086	1.068

2.2 Partial Least Squares - Discriminant Analysis (PLS-DA)

PLS is a supervised method that uses multivariate regression techniques to extract via linear combination of original variables (X) the information that can predict the class membership (Y). The PLS regression is performed using the `plsr` function provided by R `pls` package[3]. The classification and cross-validation are performed using the corresponding wrapper function offered by the `caret` package[4].

To assess the significance of class discrimination, a permutation test was performed. In each permutation, a PLS-DA model was built between the data (X) and the permuted class labels (Y) using the optimal number of components determined by cross validation for the model based on the original class assignment. The ratio of the between sum of the squares and the within sum of squares (B/W-ratio) for the class assignment prediction of each model was calculated. If the B/W ratio of the original class assignment is a part of the distribution based on the permuted class assignments The contrast between the two class assignment cannot be considered significant from a statistical point of view.

There are two variable importance measures in PLS-DA. The first, Variable Importance in Projection (VIP) is a weighted sum of squares of the PLS loadings taking into account the amount of explained Y-variation in each dimension. The other importance measure is based on the weighted sum of PLS-regression The weights are a function of the reduction of the sums of squares across the number of PLS components. coefficients[5].

The R script `chemometrics.R` is required. Figure 5 shows the overview of score plots; Figure 6 shows the 2-D score plot between selected components; Figure 7 shows the 3-D score plot between selected components; Figure 8 shows the loading plot between the selected components; Figure 9 shows the classification performance with different number of components. Figure 10 shows the important features identified by PLS-DA. Figure 11 shows the permutation test results for model validation.

[3] Ron Wehrens and Bjorn-Helge Mevik.*pls: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR)*, 2007, R package version 2.1-0

[4] Max Kuhn. Contributions from Jed Wing and Steve Weston and Andre Williams.*caret: Classification and Regression Training*, 2008, R package version 3.45

[5] Bijlsma et al.*Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation*, Anal Chem. 2006, 78 567 - 574

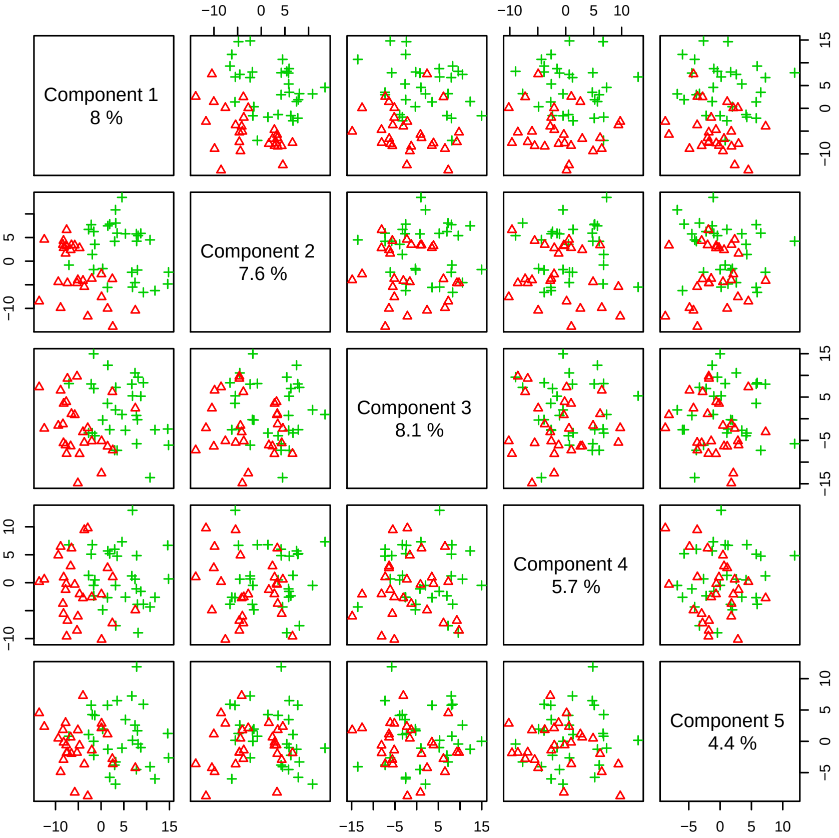

Figure 5: Pairwise score plots between the selected components. The explained variance of each component is shown in the corresponding diagonal cell.

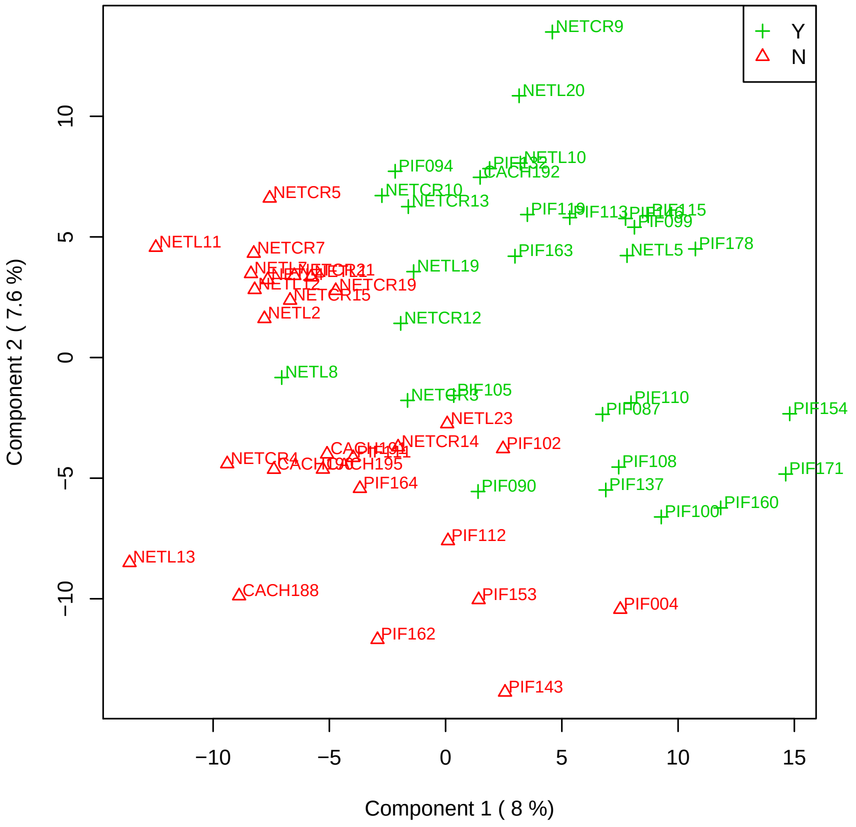

Figure 6: Score plot between the selected PCs. The explained variances are shown in brackets.

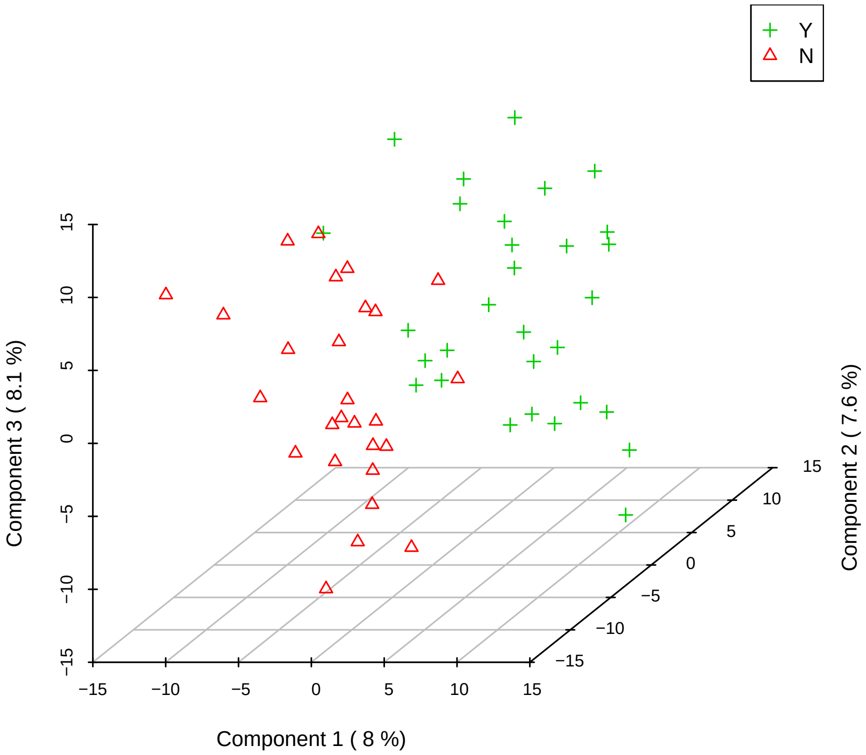

Figure 7: 3D score plot between the selected PCs. The explained variances are shown in brackets.

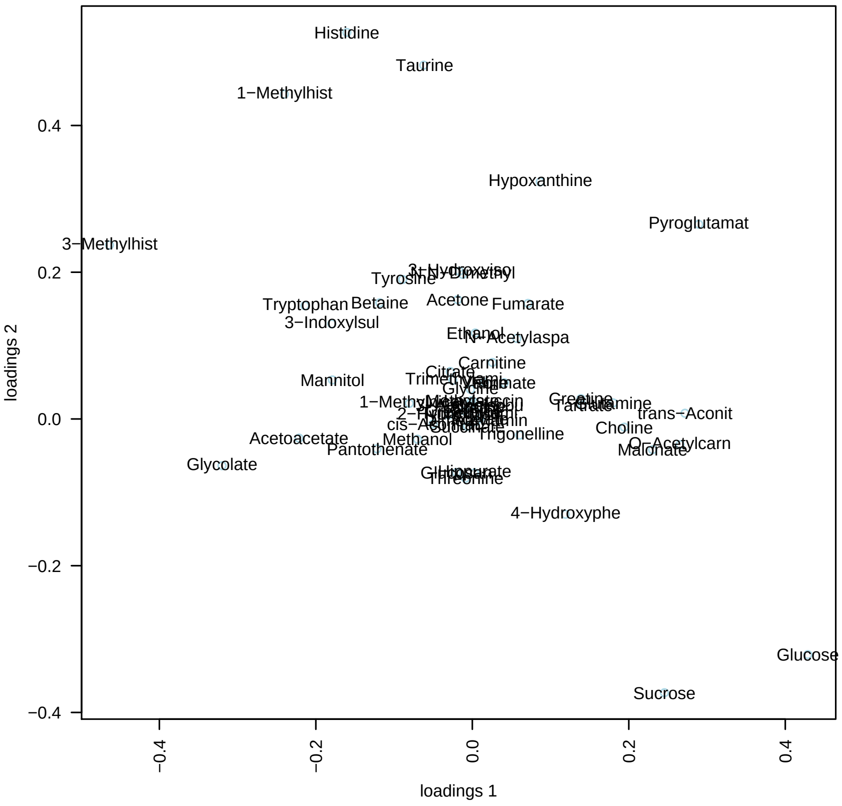

Figure 8: Loading plot between the selected PCs.

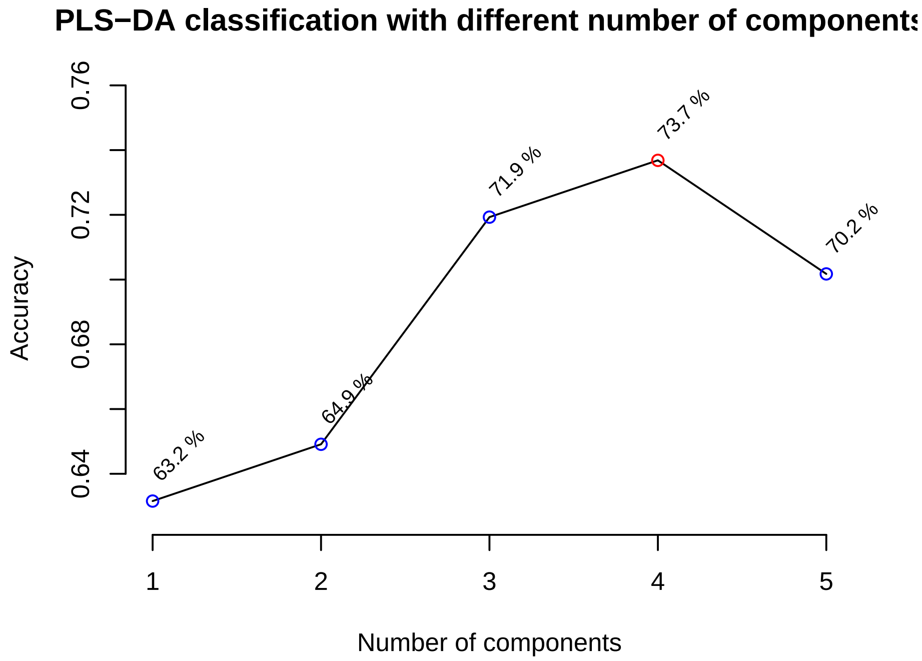

Figure 9: PLS-DA classification using different number of components. The red circle indicates the best classifier.

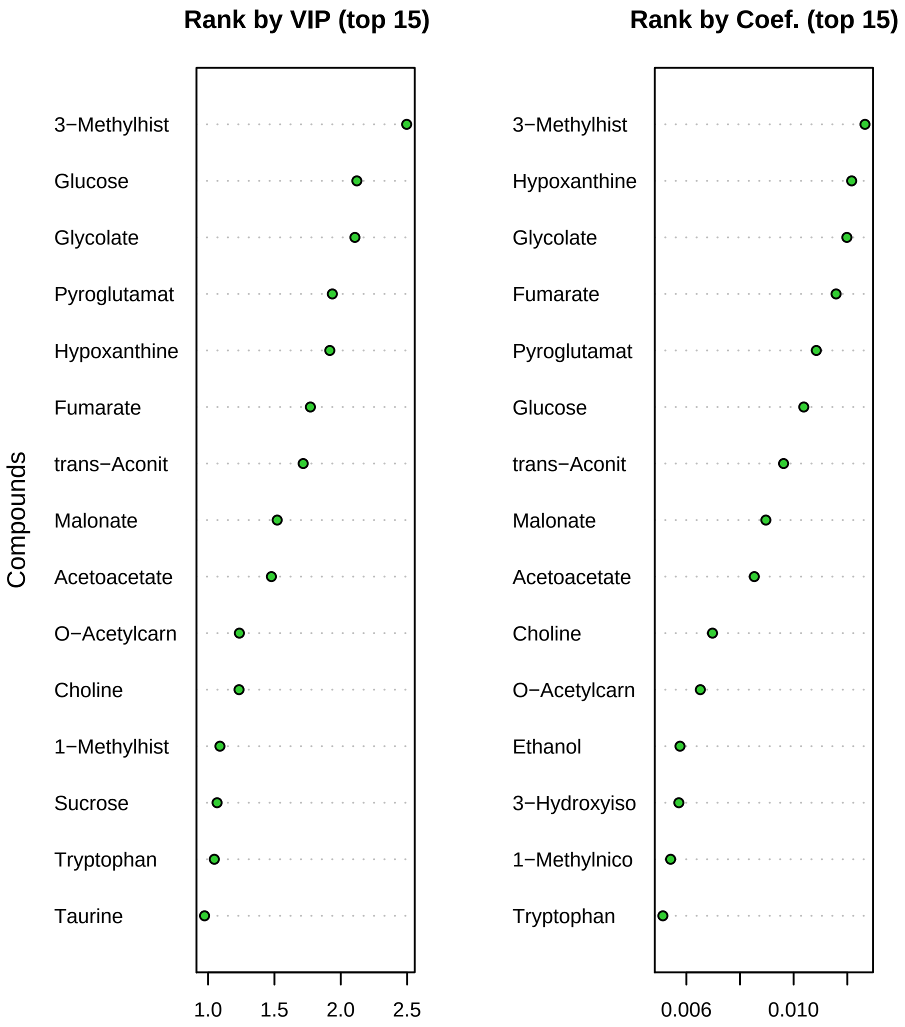

Figure 10: Important features identified by PLS-DA. The left panel shows the features ranked by VIP score. The right panel shows the features ranked based on their regression coefficients.

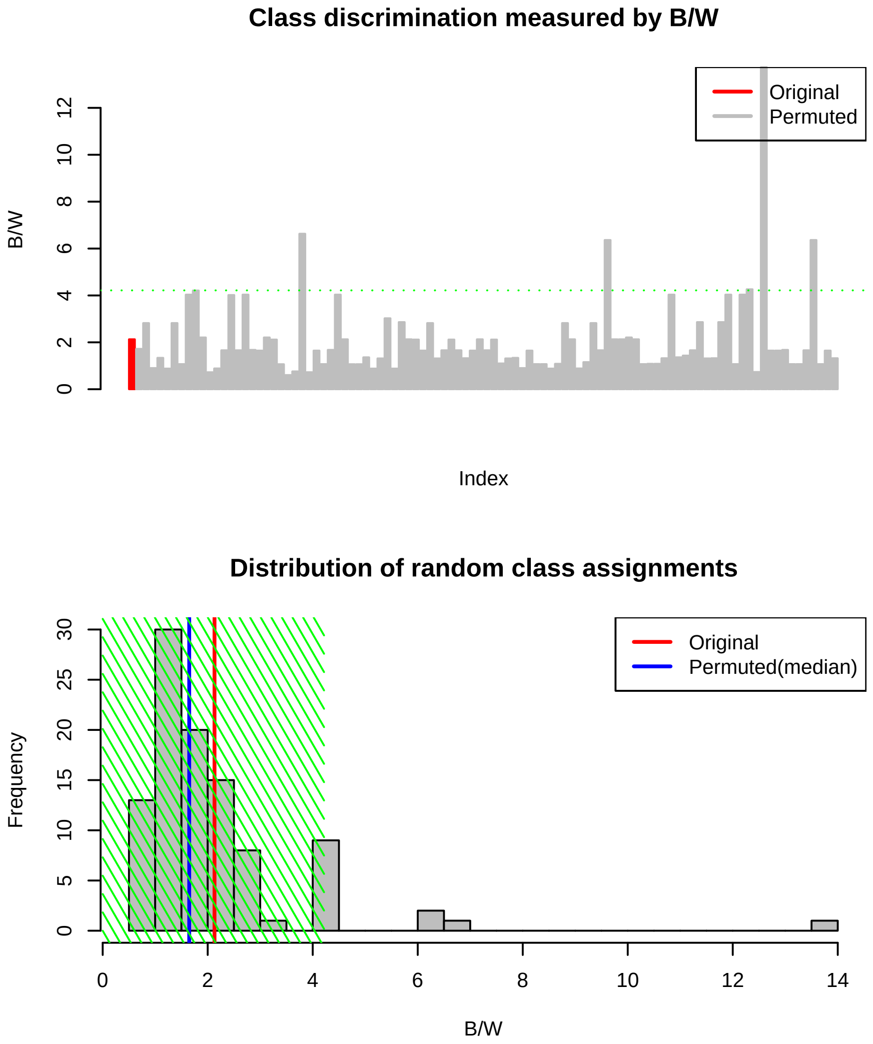

Figure 11: PLS-DA model validation by permutation tests. The top panel shows B/W ratio calculated for both original and permuted PLS-DA models. The bottom panel shows the distribution of random class assignments based on the frequencies of permuted B/W ratios. The green line (top) and green area (bottom) mark the 95% confidence regions of B/W for the permuted data.

2.3 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[6]. Users need to specify the `Delta` value to control FDR in order to proceed.

The R script `sigfeatures.R` is required. Figure 12 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	Glycolate	-3.114	0.986	0.002	0.027	0.119
2	trans-Aconitate	2.729	0.919	0.005	0.027	5.69
3	Fumarate	2.667	0.803	0.007	0.027	4.414
4	3-Methylhistidine	-2.639	1.35	0.008	0.027	0.085
5	Hypoxanthine	2.569	0.93	0.01	0.028	5.235
6	Trigonelline	1.925	0.361	0.056	0.102	1.618
7	Glucose	1.887	1.541	0.061	0.102	7.505
8	Pyroglutamate	1.879	1.418	0.063	0.102	6.343
9	Choline	1.756	1.026	0.081	0.102	3.487
10	Malonate	1.751	1.246	0.082	0.102	4.536
11	3-Hydroxyisovalerate	1.507	0.618	0.13	0.144	1.908

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

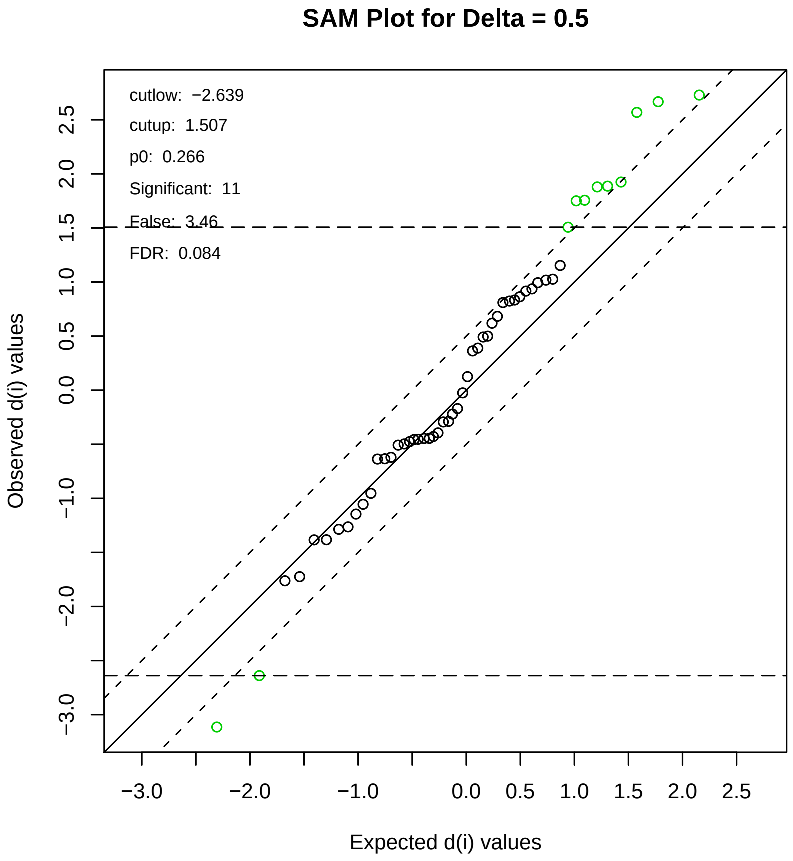

Figure 12: Significant features identified by SAM. The green circles represent features that exceed the specified threshold.

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 Tue Apr 14 19:04:35 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*).