Metabolomic Data Analysis with MetaboAnalyst 3.0

User ID: guest4588077676407564518

October 12, 2015

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. Table 1 summarizes the result of the data processing steps.

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.

Samples are in columns and features in rows. The uploaded file is in comma separated values (.csv) format. The uploaded data file contains 8 (samples) by 467 (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 should 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)

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 K-Nearest Neighbours, 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.

[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

72 variables were removed for threshold 50 percent. Missing variables were imputated using KNN

1.1.4 Data Filtering

The purpose of the data filtering is to identify and remove variables that are unlikely to be of use when modeling the data. No phenotype information are used in the filtering process, so the result can be used with any downstream analysis. This step can usually improves the results. Data filter is strongly recommended for datasets with large number of variables (> 250) datasets contain much noise (i.e.chemometrics data). Filtering can usually improve your results[2].

For data with number of variables < 250, this step will reduce 5% of variables; For variable number between 250 and 500, 10% of variables will be removed; For variable number bwteen 500 and 1000, 25% of variables will be removed; And 40% of variabled will be removed for data with over 1000 varaibles.

Reduce 10% features (39) based on Interquantile Range

Table 1: Summary of data processing results

	Features (positive)	Missing/Zero	Features (processed)
207	380	87	383
208	345	122	383
209	356	111	383
243	375	92	383
213	437	30	383
215	382	85	383
217	442	25	383
218	443	24	383

[2] Hackstadt AJ, Hess AM.*Filtering for increased power for microarray data analysis*, BMC Bioinformatics. 2009; 10: 11.

1.2 Data Normalization

The data is stored as a table with one sample per row and one variable (bin/peak/metabolite) per column. The normalization procedures implemented below are grouped into four categories. Sample specific normalization allows users to manually adjust concentrations based on biological inputs (i.e. volume, mass); row-wise normalization allows general-purpose adjustment for differences among samples; data transformation and scaling are two different approaches to make features more comparable. You can use one or combine both to achieve better results.

The normalization consists of the following options:

1. Sample specific normalization (i.e. normalize by dry weight, volume)
2. Row-wise procedures:
   - Normalization by the sum
   - Normalization by the sample median
   - Normalization by a reference sample (probabilistic quotient normalization)[3]
   - Normalization by a reference feature (i.e. creatinine, internal control)
3. Data transformation :
   - Generalized log transformation (glog 2)
   - Cube root transformation
4. Data scaling:
   - 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)

Figure 1 shows the effects before and after normalization.

[3] 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

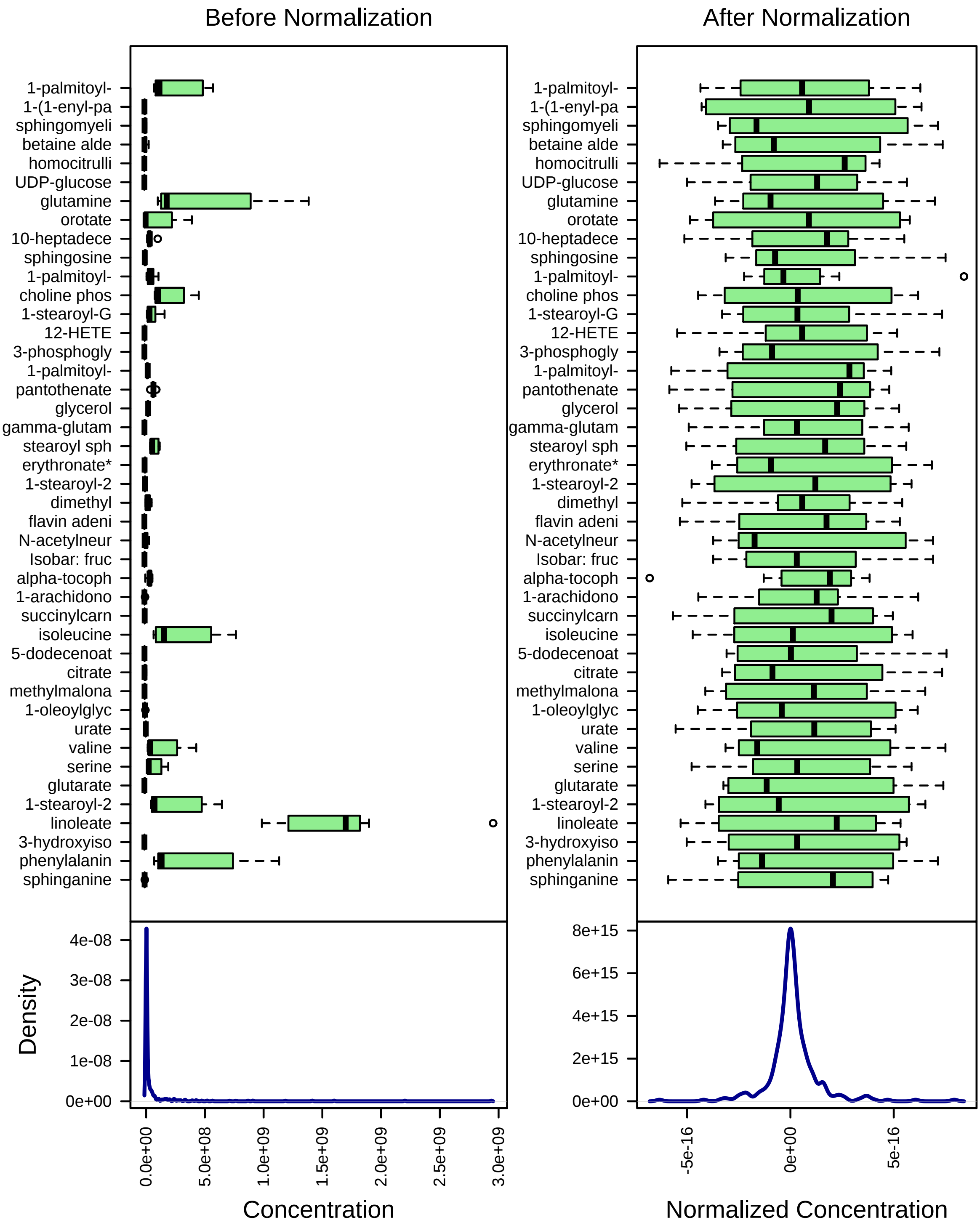

Figure 1: Box plots and kernel density plots before and after normalization. The boxplots show at most 50 features due to space limit. The density plots are based on all samples. Selected methods : Row-wise normalization: Normalization to sample median; Data transformation: Log Normalization; Data scaling: Range Scaling.

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
   - One-way ANOVA and post-hoc analysis
   - Correlation analysis
2. Multivariate analysis 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)

`Please note: some advanced methods are available only for two-group sample analyais.`

2.1 Univariate Analysis

Univariate analysis methods are the most common methods used for exploratory data analysis. For two-group data, MetaboAnalyst provides 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. For multi-group analysis, MetaboAnalyst provides two types of analysis - one-way analysis of variance (ANOVA) with associated post-hoc analyses, and correlation analysis to identify signficant compounds that follow a given pattern. The univariate analyses provide a preliminary overview about features that are potentially significant in discriminating the conditions 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)

Figure 2 shows the important features identified by fold change analysis. Figure 3 shows the important features identified by t-tests. Table 2 shows the details of these features; Figure 4 shows the important features identified by volcano plot.

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.

Figure 2: Important features selected by fold-change analysis with threshold . 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

[1] ”No significant features were found using the given threshold for fold change analysis”

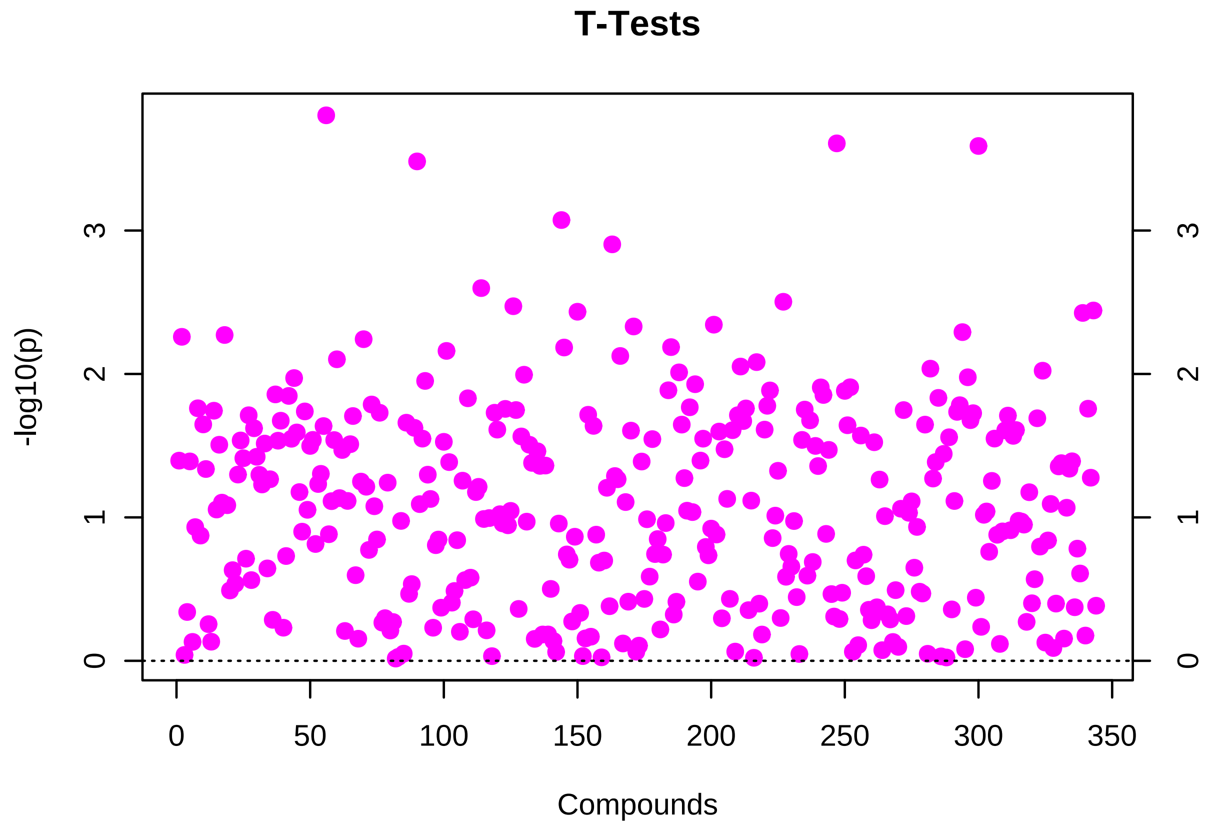

Figure 3: Important features selected by t-tests with threshold 1. The red circles represent features above the threshold. Note the p values are transformed by -log10 so that the more significant features (with smaller p values) will be plotted higher on the graph.

Table 2: Top 50 features identified by t-tests

	Compounds	p.value	-log10(p)	FDR
1	1-stearoyl-2-arachidonoyl-GPI	0.00015732	3.8032	0.02835
2	N-acetylglycine	0.00024677	3.6077	0.02835
3	S-adenosylhomocysteine	0.00025738	3.5894	0.02835
4	3-hydroxybutyrylcarnitine	0.00032965	3.4819	0.02835
5	cysteine sulfinic acid	0.00084518	3.0731	0.058148
6	ergothioneine	0.0012486	2.9036	0.071587
7	alpha-ketoglutarate	0.0025203	2.5986	0.099434
8	lysine	0.0031385	2.5033	0.099434
9	beta-alanine	0.0033706	2.4723	0.099434
10	xanthine	0.0036097	2.4425	0.099434
11	dehydroascorbate	0.003683	2.4338	0.099434
12	uridine	0.0037543	2.4255	0.099434
13	guanine	0.004534	2.3435	0.099434
14	gamma-carboxyglutamate	0.0046656	2.3311	0.099434
15	pyridoxal	0.0051092	2.2917	0.099434
16	1-(1-enyl-stearoyl)-2-oleoyl-GPE	0.0053503	2.2716	0.099434
17	1,2-dioleoyl-GPC	0.0055092	2.2589	0.099434
18	12-HETE	0.005723	2.2424	0.099434
19	glutarate	0.0064994	2.1871	0.099434
20	cytidine	0.0065409	2.1844	0.099434
21	5-oxoproline	0.0069086	2.1606	0.099434
22	flavin adenine dinucleotide	0.0074931	2.1253	0.099434
23	1-stearoyl-2-oleoyl-GPC	0.0079015	2.1023	0.099434
24	isocitrate	0.0082698	2.0825	0.099434
25	hypoxanthine	0.0088778	2.0517	0.099434
26	phenylalanine	0.0091841	2.037	0.099434
27	threonine	0.0094896	2.0228	0.099434
28	glutathione, reduced (GSH)	0.0097477	2.0111	0.099434
29	betaine aldehyde	0.010126	1.9946	0.099434
30	pyridoxamine phosphate	0.010546	1.9769	0.099434
31	1-palmitoyl-2-linoleoyl-GPC	0.010683	1.9713	0.099434
32	3-phosphoglycerate	0.011184	1.9514	0.099434
33	glycerophosphoinositol*	0.01179	1.9285	0.099434
34	N-acetylserine	0.012378	1.9074	0.099434
35	myristoylcarnitine	0.012404	1.9064	0.099434
36	glutamine	0.012996	1.8862	0.099434
37	laurylcarnitine	0.01303	1.885	0.099434
38	N-acetylmethionine	0.013121	1.882	0.099434
39	1-palmitoleoyl-2-linoleoyl-GPC	0.013886	1.8574	0.099434
40	N-acetylalanine	0.014012	1.8535	0.099434
41	1-palmitoyl-2-arachidonoyl-GPE	0.014229	1.8468	0.099434
42	phosphoenolpyruvate	0.014683	1.8332	0.099434
43	adenosine 5'-monophosphate	0.014779	1.8304	0.099434
44	16-hydroxypalmitate	0.01633	1.787	0.099434
45	putrescine	0.016513	1.7822	0.099434
46	lactate	0.016668	1.7781	0.099434
47	glycerophosphoethanolamine	0.017072	1.7677	0.099434
48	1-(1-enyl-palmitoyl)-2-arachidonoyl-GPC	0.017351	1.7607	0.099434
49	inosine	0.017395	1.7596	0.099434
50	uridine 5'-monophosphate	0.017468	1.7578	0.099434

Figure 4: Important features selected by volcano plot with fold change threshold (x) and t-tests threshold (y) . 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.

```
[1] "No significant features were found using the given threshold for volcano plot"
```

2.2 Principal Component Analysis (PCA)

PCA is an unsupervised method aiming to find the directions that best explain the variance in a data set (X) without referring to class labels (Y). The data are summarized into much fewer variables called *scores* which are weighted average of the original variables. The weighting profiles are called *loadings*. The PCA analysis is performed using the `prcomp` package. The calculation is based on singular value decomposition.

The Rscript `chemometrics.R` is required. Figure 5 is pairwise score plots providing an overview of the various seperation patterns among the most significant PCs; Figure 6 is the scree plot showing the variances explained by the selected PCs; Figure 7 shows the 2-D scores plot between selected PCs; Figure 8 shows the 3-D scores plot between selected PCs; Figure 9 shows the loadings plot between the selected PCs; Figure 10 shows the biplot between the selected PCs.

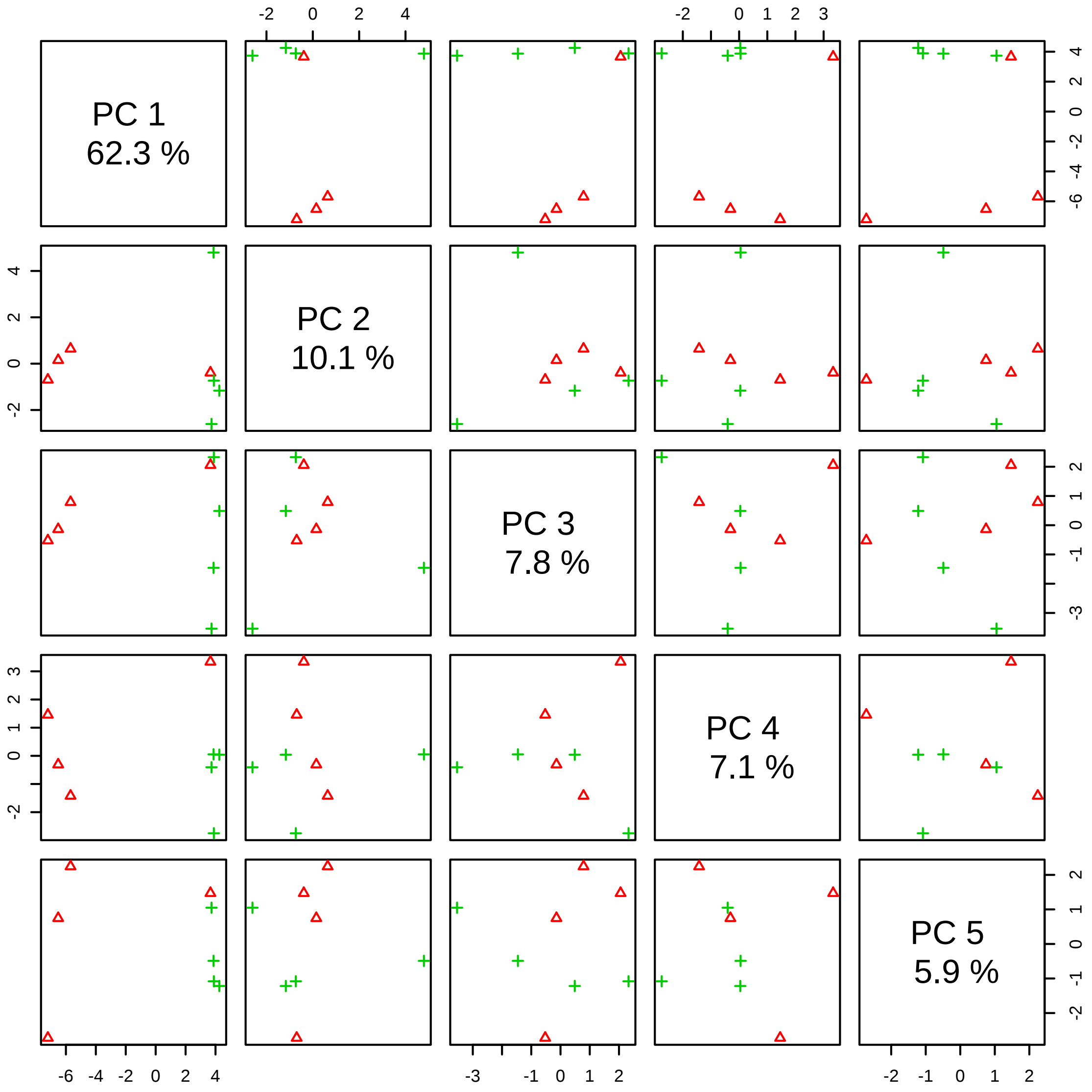

Figure 5: Pairwise score plots between the selected PCs. The explained variance of each PC is shown in the corresponding diagonal cell.

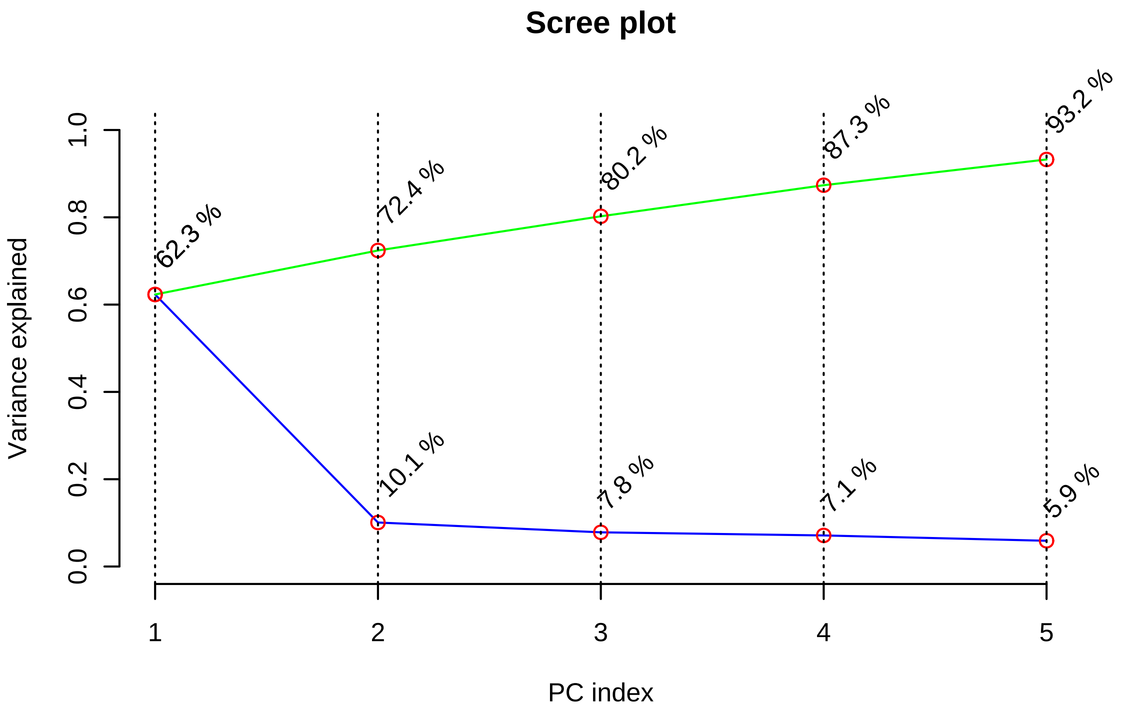

Figure 6: Scree plot shows the variance explained by PCs. The green line on top shows the accumulated variance explained; the blue line underneath shows the variance explained by individual PC.

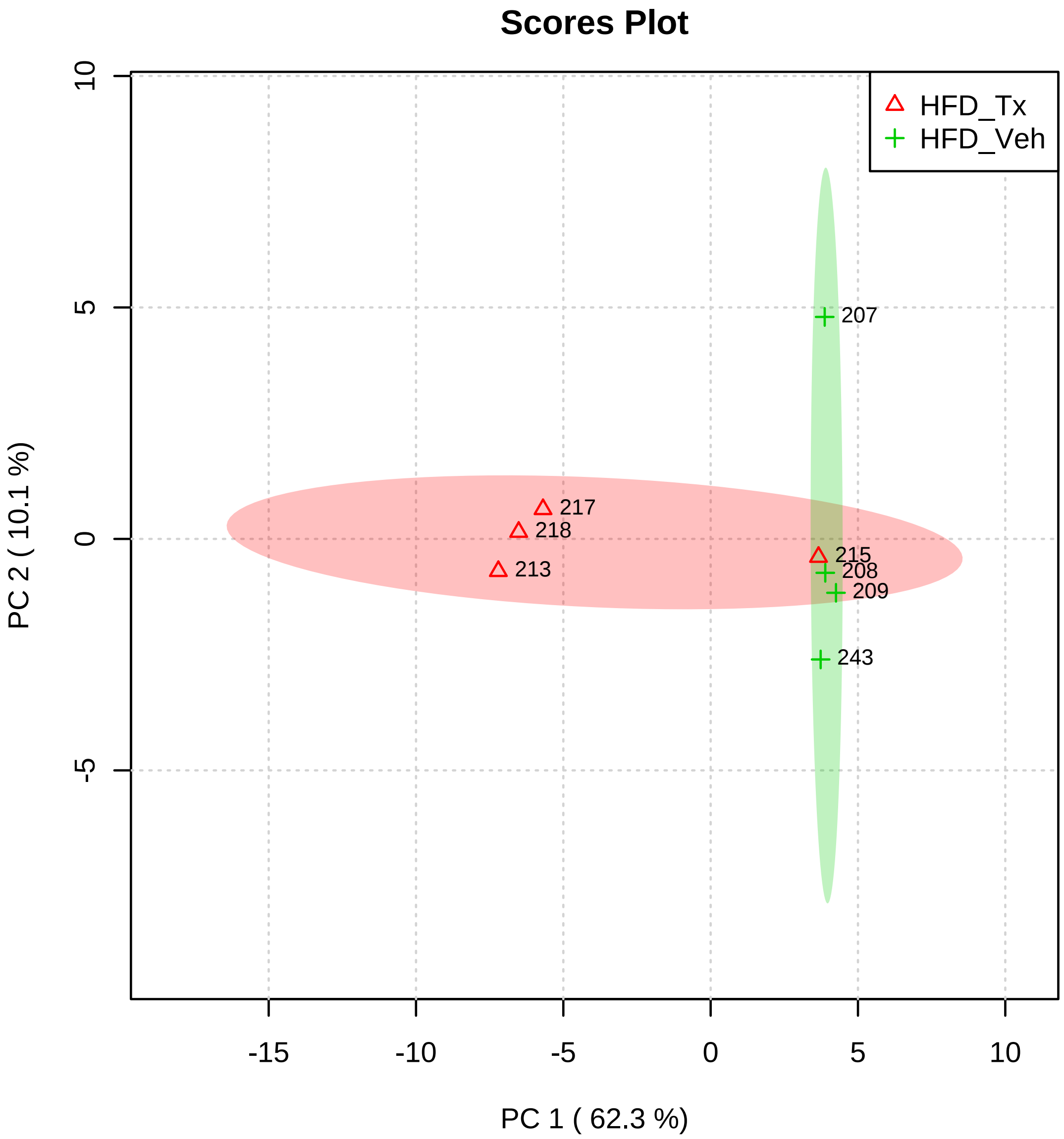

Figure 7: Scores plot between the selected PCs. The explained variances are shown in brackets.

Figure 8: 3D score plot between the selected PCs. The explained variances are shown in brackets.

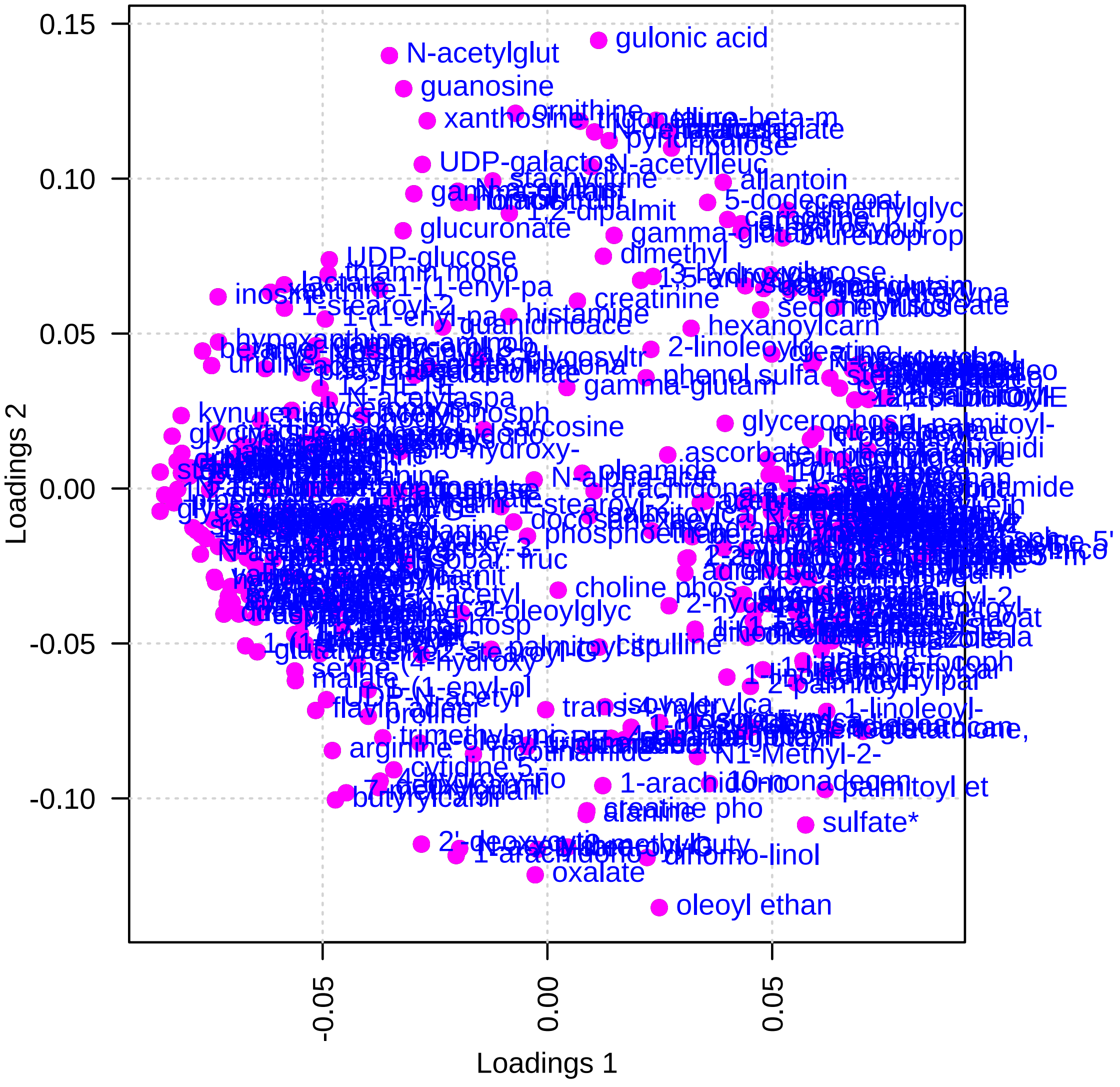

Figure 9: Loadings plot for the selected PCs.

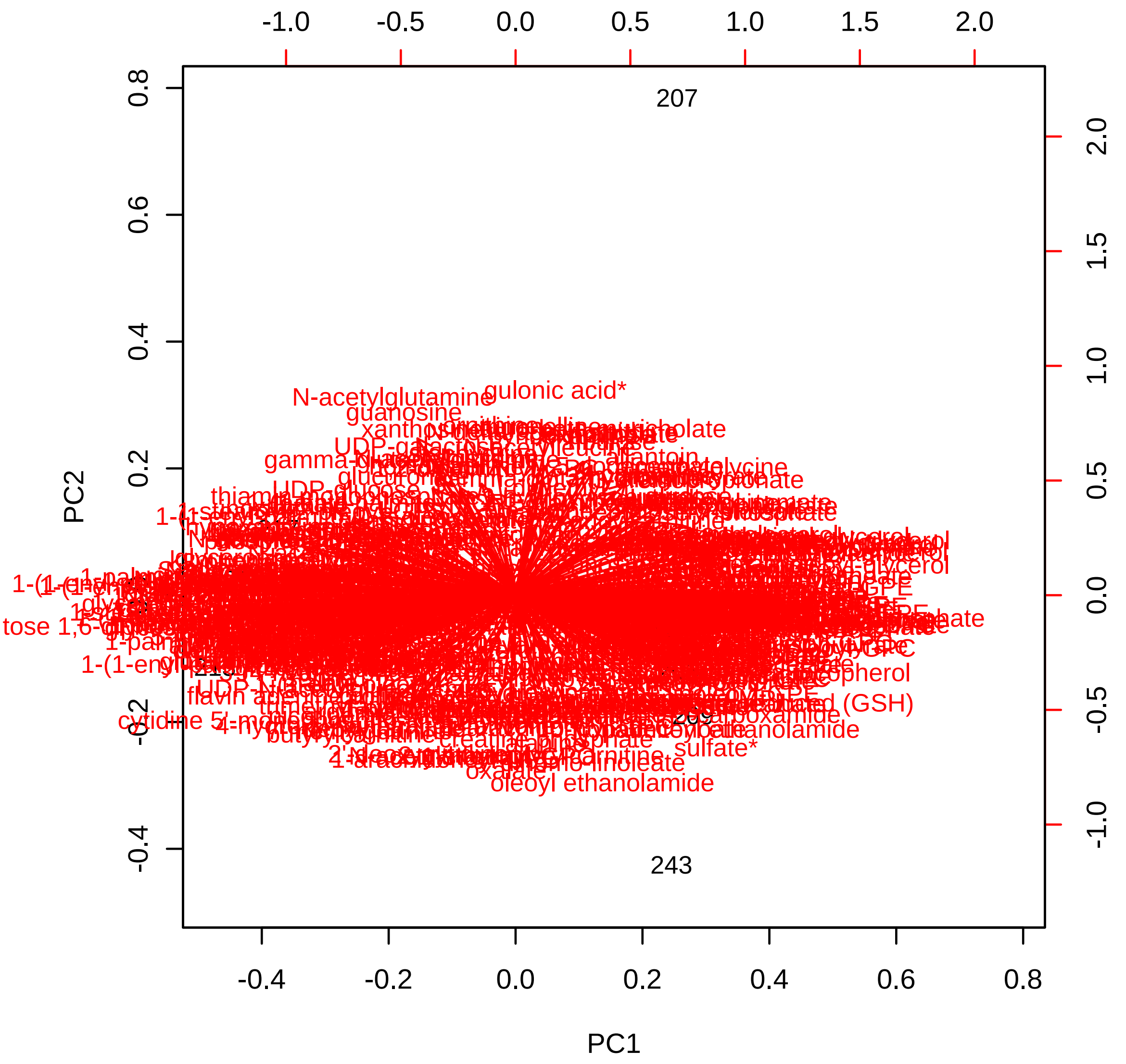

Figure 10: PCA biplot between the selected PCs. Note, you may want to test different centering and scaling normalization methods for the biplot to be displayed properly.

2.3 Random Forest (RF)

Random Forest is a supervised learning algorithm suitable for high dimensional data analysis. It uses an ensemble of classification trees, each of which is grown by random feature selection from a bootstrap sample at each branch. Class prediction is based on the majority vote of the ensemble. RF also provides other useful information such as OOB (out-of-bag) error, variable importance measure, and outlier measures. During tree construction, about one-third of the instances are left out of the bootstrap sample. This OOB data is then used as test sample to obtain an unbiased estimate of the classification error (OOB error). Variable importance is evaluated by measuring the increase of the OOB error when it is permuted. The outlier measures are based on the proximities during tree construction.

RF analysis is performed using the `randomForest` package[4]. Table 3 shows the confusion matrix of random forest. Figure 11 shows the cumulative error rates of random forest analysis for given parameters. Figure 12 shows the important features ranked by random forest. Figure 13 shows the outlier measures of all samples for the given parameters. The OOB error is 0.125

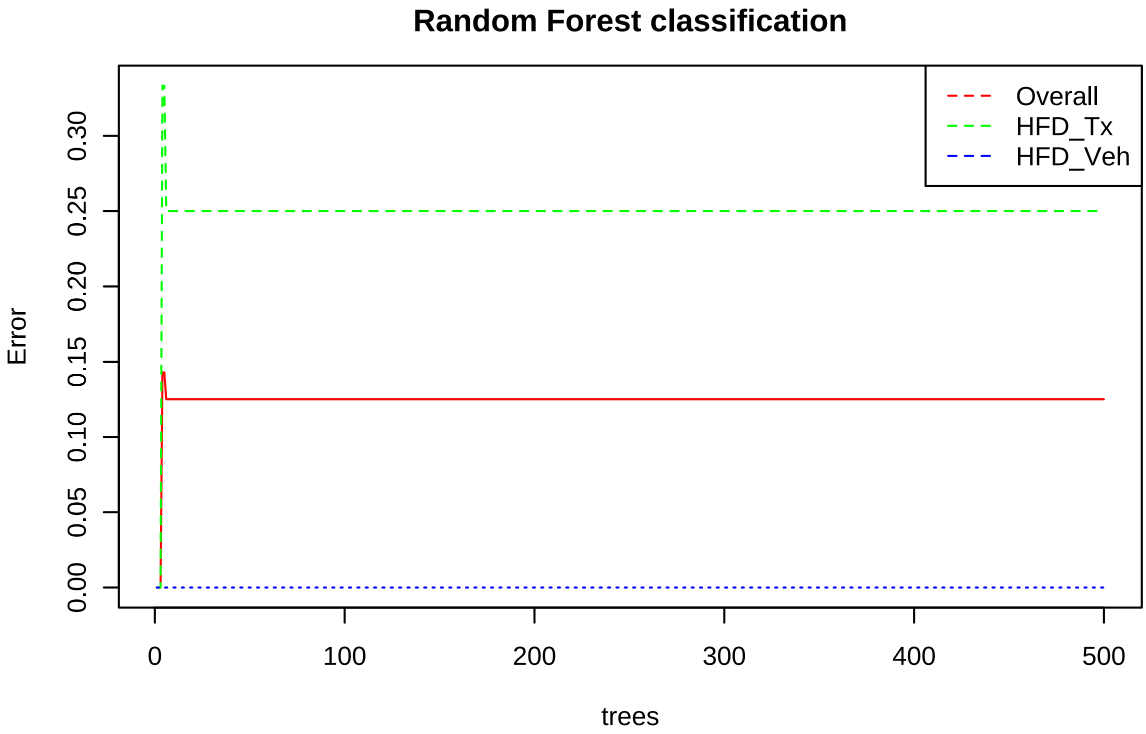

Figure 11: Cumulative error rates by Random Forest classification. The overall error rate is shown as the black line; the red and green lines represent the error rates for each class.

	HFD_Tx	HFD_Veh	class.error
HFD_Tx	3.00	1.00	0.25
HFD_Veh	0.00	4.00	0.00

Table 3: Random Forest Classification Performance

[4] Andy Liaw and Matthew Wiener. *Classification and Regression by randomForest*, 2002, R News

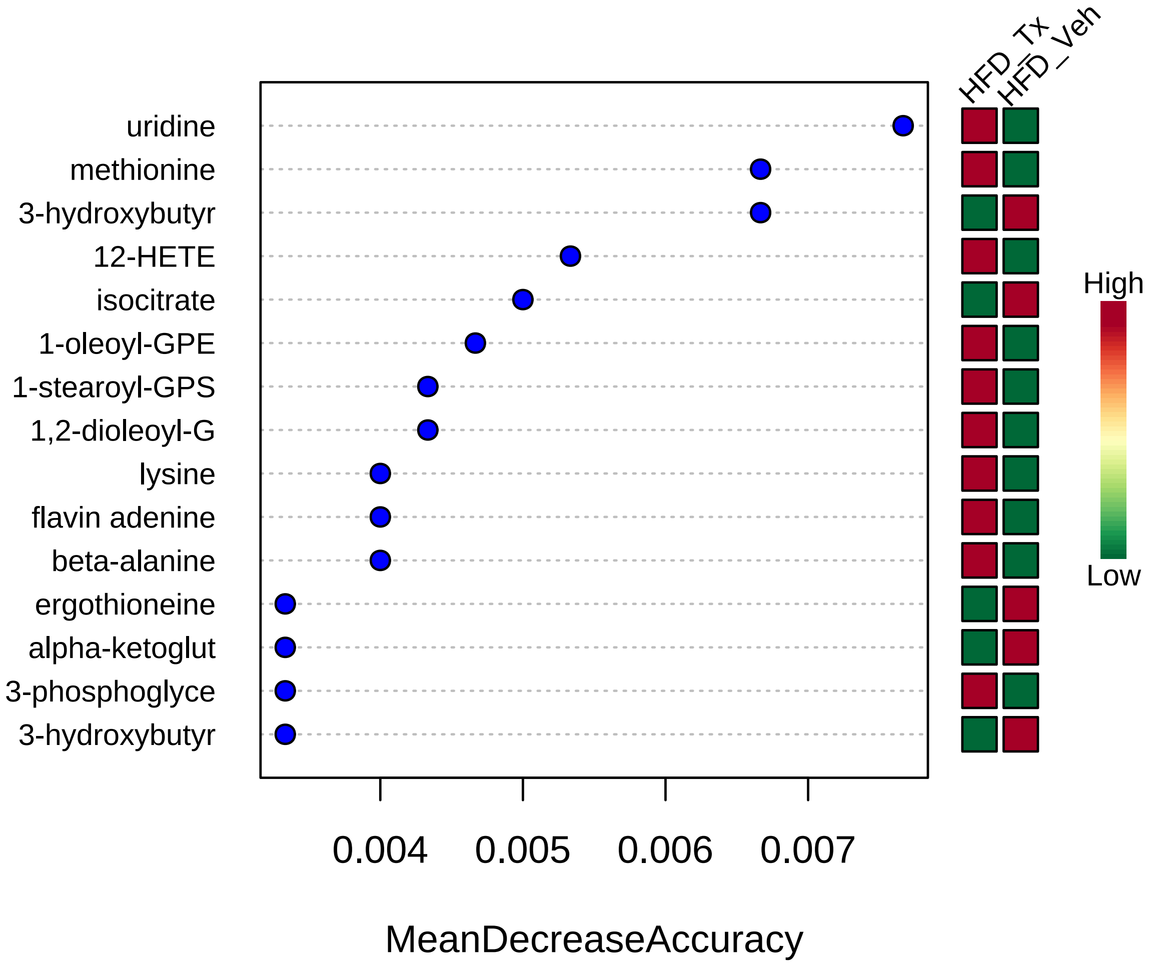

Figure 12: Significant features identified by Random Forest. The features are ranked by the mean decrease in classification accuracy when they are permuted.

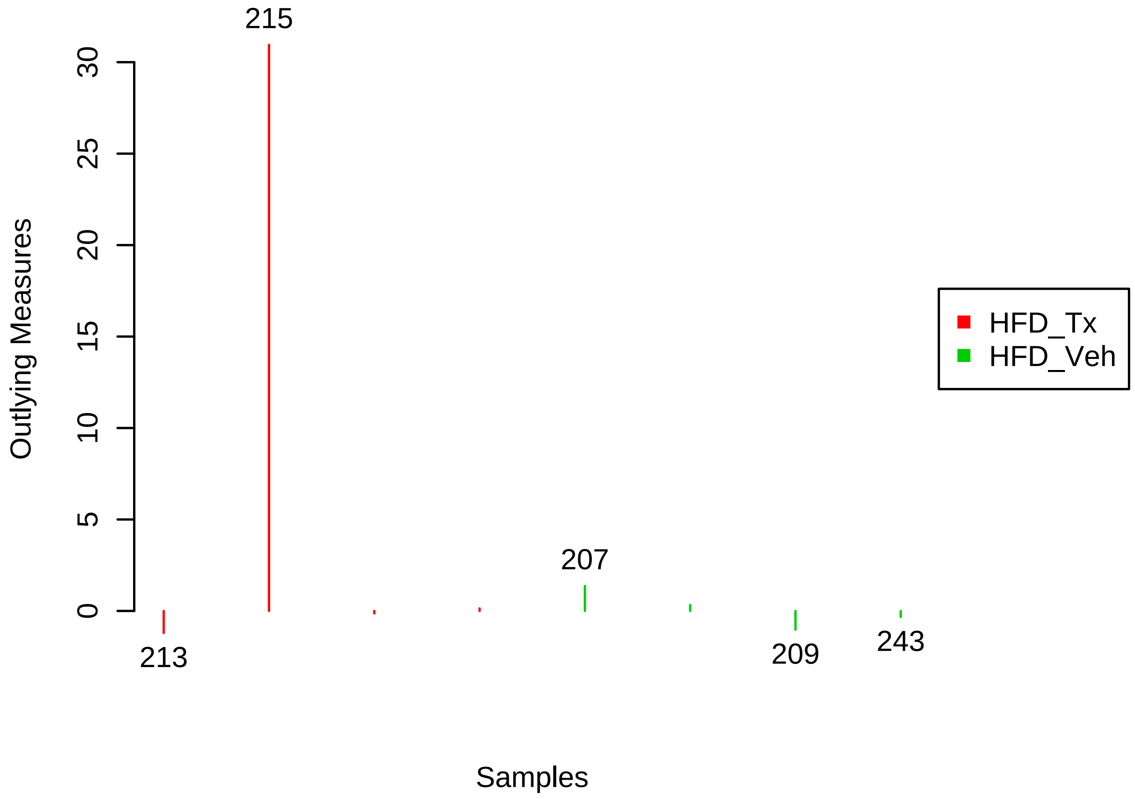

Figure 13: Potential outliers identified by Random Forest. Only the top five are labeled.

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 metabolite set enrichment analysis and metabolic pathway analysis.

The report was generated on Mon Oct 12 08:02:13 2015 with R version 3.0.3 (2014-03-06). Thank you for using MetaboAnalyst! For suggestions and feedback please contact Jeff Xia (*jeff.xia@mcgill.ca*).