Metabolomic Data Analysis with MetaboAnalyst 3.0

User ID: guest1281228020898600557

September 24, 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 481 (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

73 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.

No data filtering was applied

Table 1: Summary of data processing results

	Features (positive)	Missing/Zero	Features (processed)
207	393	88	396
208	358	123	396
209	369	112	396
243	388	93	396
213	451	30	396
215	395	86	396
217	456	25	396
218	457	24	396

[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

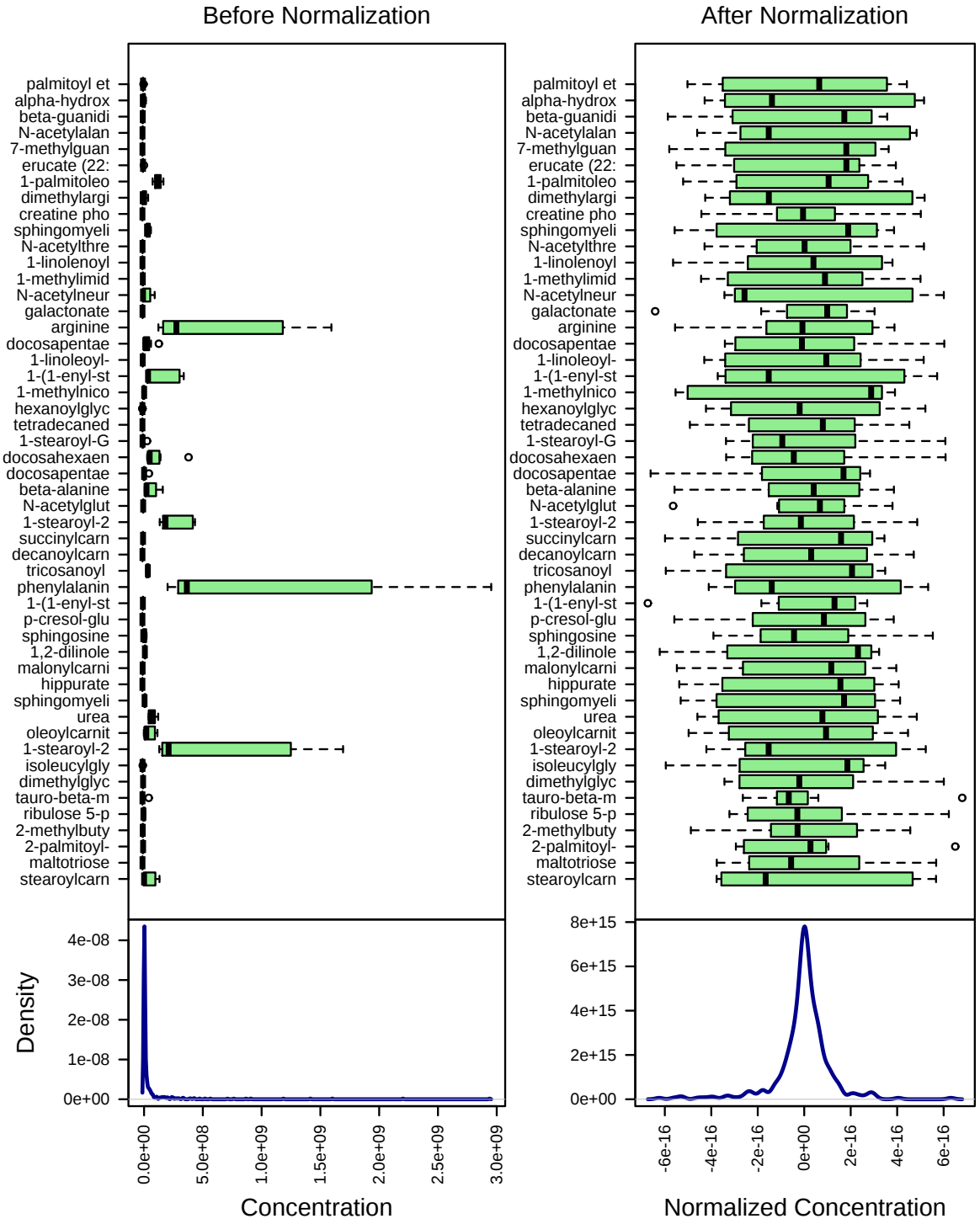

Figure 1: Box plots and kernel density plots before and after normalization. The boxplots show at most 50 features due to space limit. The density plots are based on all samples. Selected methods : Row-wise normalization: Normalization to sample median; Data transformation: Log Normalization; Data scaling: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

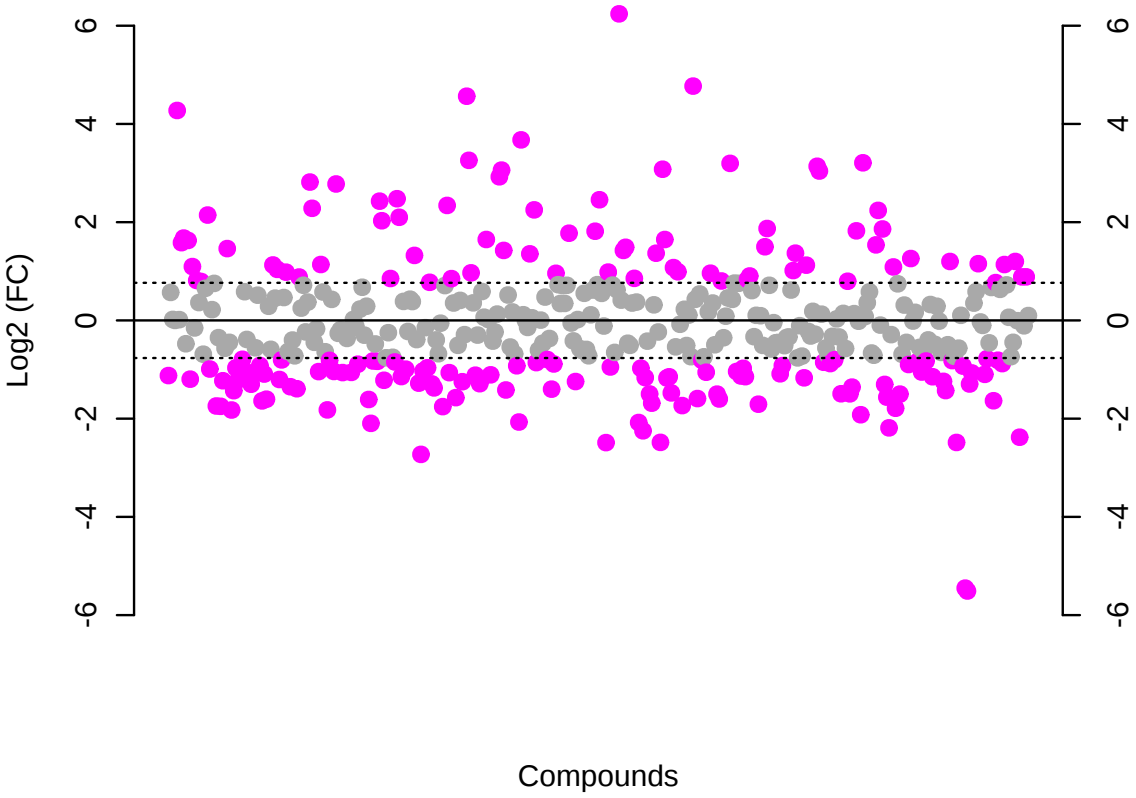

Figure 2: Important features selected by fold-change analysis with threshold 1.7. The red circles represent features above the threshold. Note the values are on log scale, so that both up-regulated and downregulated features can be plotted in a symmetrical way

Table 2: Top 50 features identified by fold change analysis

	Compounds	Fold Change	log2(FC)
1	glycerophosphoinositol*	75.655	6.2414
2	taurocholate	0.021962	-5.5088
3	tauro-beta-muricholate	0.022868	-5.4506
4	kynurenine	27.297	4.7706
5	betaine	23.663	4.5646
6	1,2-distearoyl-GPC (18:0/18:0)	19.352	4.2744
7	dehydroascorbate	12.781	3.6759
8	betaine aldehyde	9.5746	3.2592
9	pipecolate	9.2465	3.2089
10	methionine sulfoxide	9.1715	3.1972
11	orotate	8.8053	3.1384
12	hypotaurine	8.4446	3.078
13	cysteine sulfinic acid	8.3413	3.0603
14	orotidine	8.2332	3.0415
15	cysteine	7.5909	2.9243
16	1-stearoyl-GPI (18:0)	7.0467	2.8169
17	2-aminoadipate	6.8578	2.7777
18	adenosine 5'-monophosphate (AMP)	0.15097	-2.7277
19	glutathione, reduced (GSH)	0.17826	-2.4879
20	sucrose	0.17866	-2.4847
21	homostachydrine*	0.17886	-2.4831
22	5-methylthioadenosine (MTA)	5.5699	2.4777
23	glutarate (pentanedioate)	5.4896	2.4567
24	3-methylhistidine	5.3809	2.4279
25	uridine 5'-monophosphate (UMP)	0.19238	-2.378
26	argininosuccinate	5.0667	2.341
27	1-stearoyl-GPS (18:0)*	4.8649	2.2824
28	dimethylarginine (SDMA + ADMA)	4.7592	2.2507
29	heme	0.21062	-2.2473
30	pyridoxal	4.7242	2.2401
31	retinol (Vitamin A)	0.21954	-2.1874
32	1-(1-enyl-stearoyl)-GPE (P-18:0)*	4.4229	2.145
33	5-oxoproline	4.2862	2.0997
34	3-hydroxybutyrylcarnitine (1)	0.23394	-2.0958
35	guanosine 5'- monophosphate (5'-GMP)	0.23673	-2.0787
36	decanoylcarnitine	0.23836	-2.0688
37	3-phosphoglycerate	4.0821	2.0293
38	phosphopantetheine	0.2642	-1.9203
39	N-acetylneuraminate	3.6581	1.8711
40	pyridoxamine phosphate	3.6335	1.8614
41	1-methylnicotinamide	0.28243	-1.824
42	phosphoenolpyruvate (PEP)	3.5399	1.8237
43	16-hydroxypalmitate	0.28291	-1.8216
44	glutamate, gamma-methyl ester	3.5209	1.816
45	salicylate	0.28949	-1.7884
46	gamma-carboxyglutamate	3.4231	1.7753
47	arachidonoyl ethanolamide	0.29629	-1.7549
48	1-linoleoyl-GPC (18:2)	0.29749	-1.7491
49	1-linolenoyl-GPC (18:3)*	0.2989	-1.7423
50	isocitrate	0.30059	-1.7341

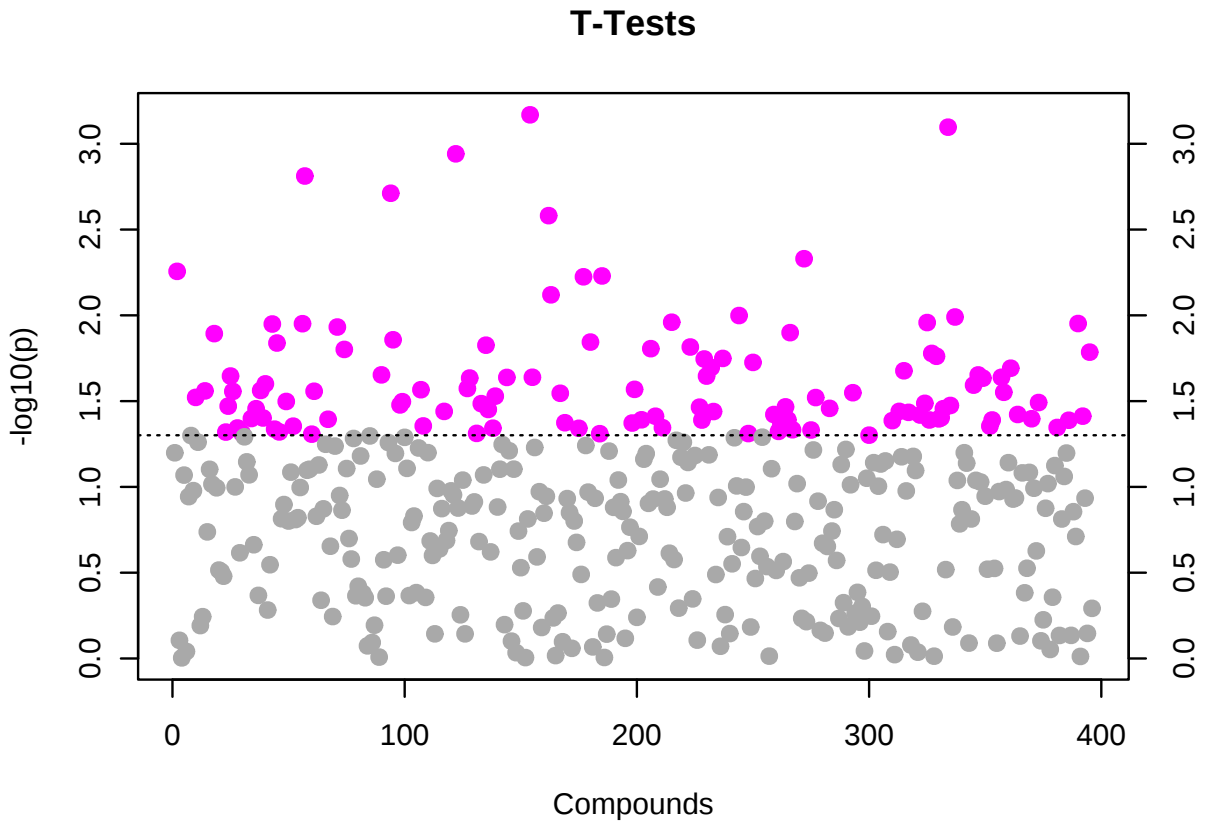

Figure 3: Important features selected by t-tests with threshold 0.05. The red circles represent features above the threshold. Note the p values are transformed by -log10 so that the more significant features (with smaller p values) will be plotted higher on the graph.

Table 3: Top 50 features identified by t-tests

	Compounds	p.value	-log10(p)	FDR
1	cysteine sulfinic acid	0.00067813	3.1687	0.15109
2	S-adenosylhomocysteine (SAH)	0.0008003	3.0967	0.15109
3	alpha-ketoglutarate	0.0011446	2.9413	0.15109
4	1-stearoyl-2-arachidonoyl-GPI (18:0/20:4)	0.0015415	2.8121	0.15261
5	3-hydroxybutyrylcarnitine (1)	0.0019426	2.7116	0.15386
6	decanoylcarnitine	0.0026239	2.5811	0.1625
7	N-acetylglycine	0.004677	2.33	0.1625
8	1,2-dioleoyl-GPC (18:1/18:1)*	0.0055416	2.2564	0.1625
9	gamma-carboxyglutamate	0.005894	2.2296	0.1625
10	ergothioneine	0.0059619	2.2246	0.1625
11	dehydroascorbate	0.0075935	2.1196	0.1625
12	laurylcarnitine	0.010027	1.9988	0.1625
13	sebacate (decanedioate)	0.010224	1.9904	0.1625
14	guanine	0.010966	1.9599	0.1625
15	pseudouridine	0.011017	1.9579	0.1625
16	uridine	0.011163	1.9522	0.1625
17	1-stearoyl-2-arachidonoyl-GPE (18:0/20:4)	0.011189	1.9512	0.1625
18	1-palmitoyl-2-arachidonoyl-GPE (16:0/20:4)*	0.011234	1.9495	0.1625
19	12-HETE	0.011698	1.9319	0.1625
20	myristoylcarnitine	0.012614	1.8991	0.1625
21	1-(1-enyl-stearoyl)-2-oleoyl-GPE (P-18:0/18:1)	0.012781	1.8934	0.1625
22	3-hydroxybutyrylcarnitine (2)	0.013882	1.8575	0.1625
23	flavin adenine dinucleotide (FAD)	0.014315	1.8442	0.1625
24	1-palmitoyl-2-linoleoyl-GPC (16:0/18:2)	0.014499	1.8387	0.1625
25	beta-alanine	0.014928	1.826	0.1625
26	hippurate	0.015295	1.8154	0.1625
27	glycerophosphoethanolamine	0.015639	1.8058	0.1625
28	16-hydroxypalmitate	0.015801	1.8013	0.1625
29	xanthine	0.016389	1.7855	0.1625
30	pyridoxal	0.016668	1.7781	0.1625
31	pyridoxamine phosphate	0.017322	1.7614	0.1625
32	isocitrate	0.01783	1.7489	0.1625
33	hypoxanthine	0.017962	1.7457	0.1625
34	lysine	0.018808	1.7257	0.1625
35	indolelactate	0.02013	1.6962	0.1625
36	succinate	0.020331	1.6918	0.1625
37	phenyllactate (PLA)	0.021045	1.6769	0.1625
38	3-(4-hydroxyphenyl)lactate	0.022231	1.653	0.1625
39	sphingomyelin (d18:1/20:1, d18:2/20:0)*	0.02225	1.6527	0.1625
40	1-linoleoyl-GPC (18:2)	0.022581	1.6463	0.1625
41	imidazole lactate	0.022598	1.6459	0.1625
42	cytidine	0.022956	1.6391	0.1625
43	stearidonate (18:4n3)	0.02296	1.639	0.1625
44	cholesterol	0.023008	1.6381	0.1625
45	arginine	0.0232	1.6345	0.1625
46	sphingomyelin (d18:1/22:1, d18:2/22:0, d16:1/24:1)*	0.023244	1.6337	0.1625
47	1-palmitoleoyl-3-oleoyl-glycerol (16:1/18:1)*	0.025085	1.6006	0.1625
48	sphingomyelin (d18:1/18:1, d18:2/18:0)	0.025518	1.5932	0.1625
49	arachidonoyl ethanolamide	0.026679	1.5738	0.1625
50	glutarate (pentanedioate)	0.027002	1.5686	0.1625

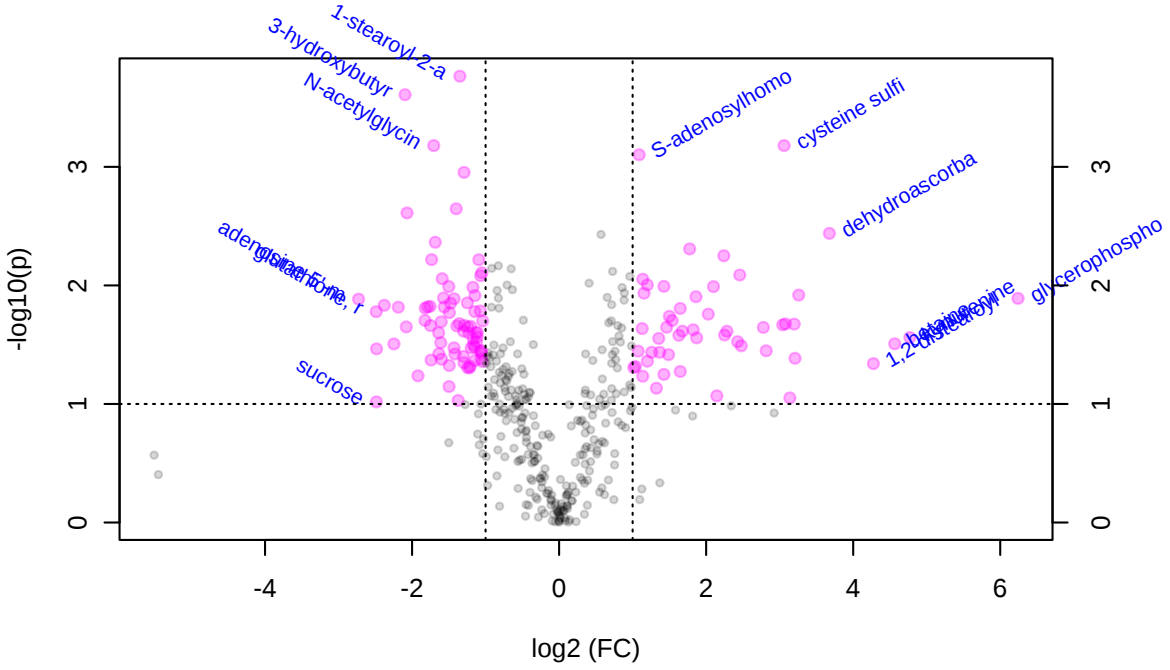

Figure 4: Important features selected by volcano plot with fold change threshold (x) 2 and t-tests threshold (y) 0.1. The red circles represent features above the threshold. Note both fold changes and p values are log transformed. The further its position away from the (0,0), the more significant the feature is.

Table 4: Top 50 features identified by volcano plot

	Compounds	FC	log2(FC)	p.value	-log10(p)
1	1-stearoyl-2-arachidonoyl-GPI (18:0/20:4)	0.39231	-1.3499	0.0001719	3.7647
2	3-hydroxybutyrylcarnitine (1)	0.23394	-2.0958	0.00024614	3.6088
3	cysteine sulfinic acid	8.3413	3.0603	0.00066122	3.1797
4	N-acetylglycine	0.30657	-1.7057	0.00066146	3.1795
5	S-adenosylhomocysteine (SAH)	2.1266	1.0886	0.00079069	3.102
6	alpha-ketoglutarate	0.40805	-1.2932	0.0011128	2.9536
7	ergothioneine	0.37901	-1.3997	0.0022575	2.6464
8	decanoylcarnitine	0.23836	-2.0688	0.0024459	2.6116
9	dehydroascorbate	12.781	3.6759	0.0036398	2.4389
10	hippurate	0.311	-1.685	0.0043228	2.3642
11	gamma-carboxyglutamate	3.4231	1.7753	0.0049217	2.3079
12	pyridoxal	4.7242	2.2401	0.0056192	2.2503
13	isocitrate	0.30059	-1.7341	0.0060591	2.2176
14	1-palmitoyl-2-linoleoyl-GPC (16:0/18:2)	0.4691	-1.092	0.0060712	2.2167
15	sphingomyelin (d18:1/20:1, d18:2/20:0)*	0.48251	-1.0514	0.0078239	2.1066
16	glutarate (pentanedioate)	5.4896	2.4567	0.0081712	2.0877
17	1-palmitoleoyl-2-linoleoyl-GPC (16:1/18:2)*	0.47717	-1.0674	0.0082831	2.0818
18	laurylcarnitine	0.33195	-1.5909	0.008775	2.0568
19	12-HETE	2.201	1.1382	0.0088622	2.0525
20	uridine	2.2978	1.2002	0.0098966	2.0045
21	cytidine	2.6853	1.4251	0.010185	1.992
22	sebacate (decanedioate)	0.35342	-1.5006	0.01022	1.9905
23	5-oxoproline	4.2862	2.0997	0.010234	1.9899
24	imidazole lactate	0.44355	-1.1728	0.010368	1.9843
25	threonine	2.2262	1.1546	0.011627	1.9345
26	betaine aldehyde	9.5746	3.2592	0.01206	1.9187
27	myristoylcarnitine	0.4521	-1.1453	0.012181	1.9143
28	pyridoxamine phosphate	3.6335	1.8614	0.012463	1.9044
29	azelate (nonanedioate)	0.33629	-1.5722	0.012762	1.8941
30	glycerophosphoinositol*	75.655	6.2414	0.012864	1.8906
31	stearoyl ethanolamide	0.37152	-1.4285	0.01296	1.8874
32	adenosine 5'-monophosphate (AMP)	0.15097	-2.7277	0.013052	1.8843
33	beta-guanidinopropanoate	0.42101	-1.2481	0.014034	1.8528
34	indolelactate	0.35947	-1.476	0.014133	1.8498
35	uridine 5'-monophosphate (UMP)	0.19238	-2.378	0.014711	1.8324
36	arachidonoyl ethanolamide	0.29629	-1.7549	0.015013	1.8235
37	salicylate	0.28949	-1.7884	0.015176	1.8188
38	pyroglutamine*	0.33871	-1.5619	0.015198	1.8182
39	retinol (Vitamin A)	0.21954	-2.1874	0.015266	1.8163
40	16-hydroxypalmitate	0.28291	-1.8216	0.015529	1.8089
41	hypoxanthine	3.1322	1.6472	0.01564	1.8058
42	1-palmitoleoyl-3-oleoyl-glycerol (16:1/18:1)*	0.47602	-1.0709	0.01641	1.7849
43	sphingomyelin (d18:2/23:0, d18:1/23:1, d17:1/24:1)*	0.4516	-1.1469	0.016618	1.7794
44	glutathione, reduced (GSH)	0.17826	-2.4879	0.016657	1.7784
45	pantothenate	0.35522	-1.4932	0.01697	1.7703
46	3-phosphoglycerate	4.0821	2.0293	0.017468	1.7577
47	N-acetylmethionine	2.8278	1.4997	0.018213	1.7396
48	1-methylnicotinamide	0.28243	-1.824	0.019747	1.7045
49	putrescine	2.9036	1.5378	0.019776	1.7039
50	methylsuccinate	0.48644	-1.0397	0.019947	1.7001

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.

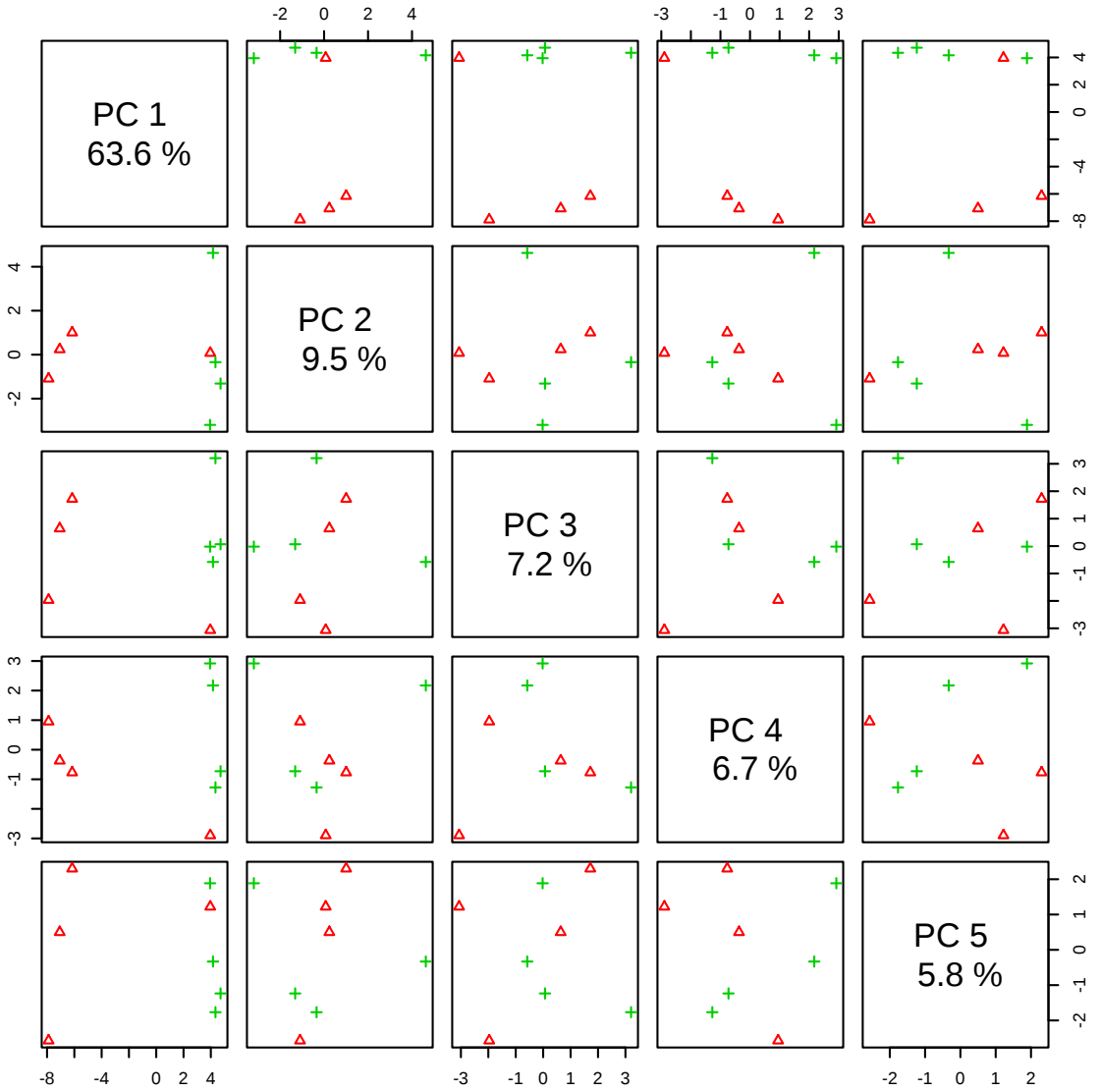

Figure 5: Pairwise score plots between the selected PCs. The explained variance of each PC is shown in the corresponding diagonal cell.

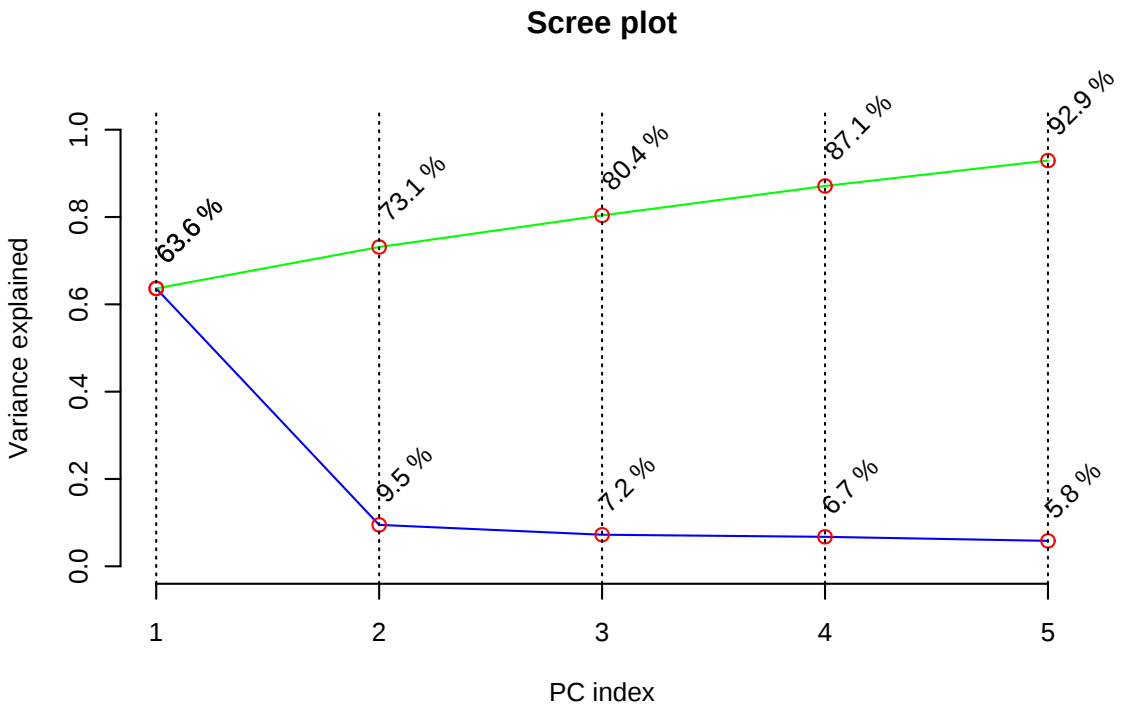

Figure 6: Scree plot shows the variance explained by PCs. The green line on top shows the accumulated variance explained; the blue line underneath shows the variance explained by individual PC.

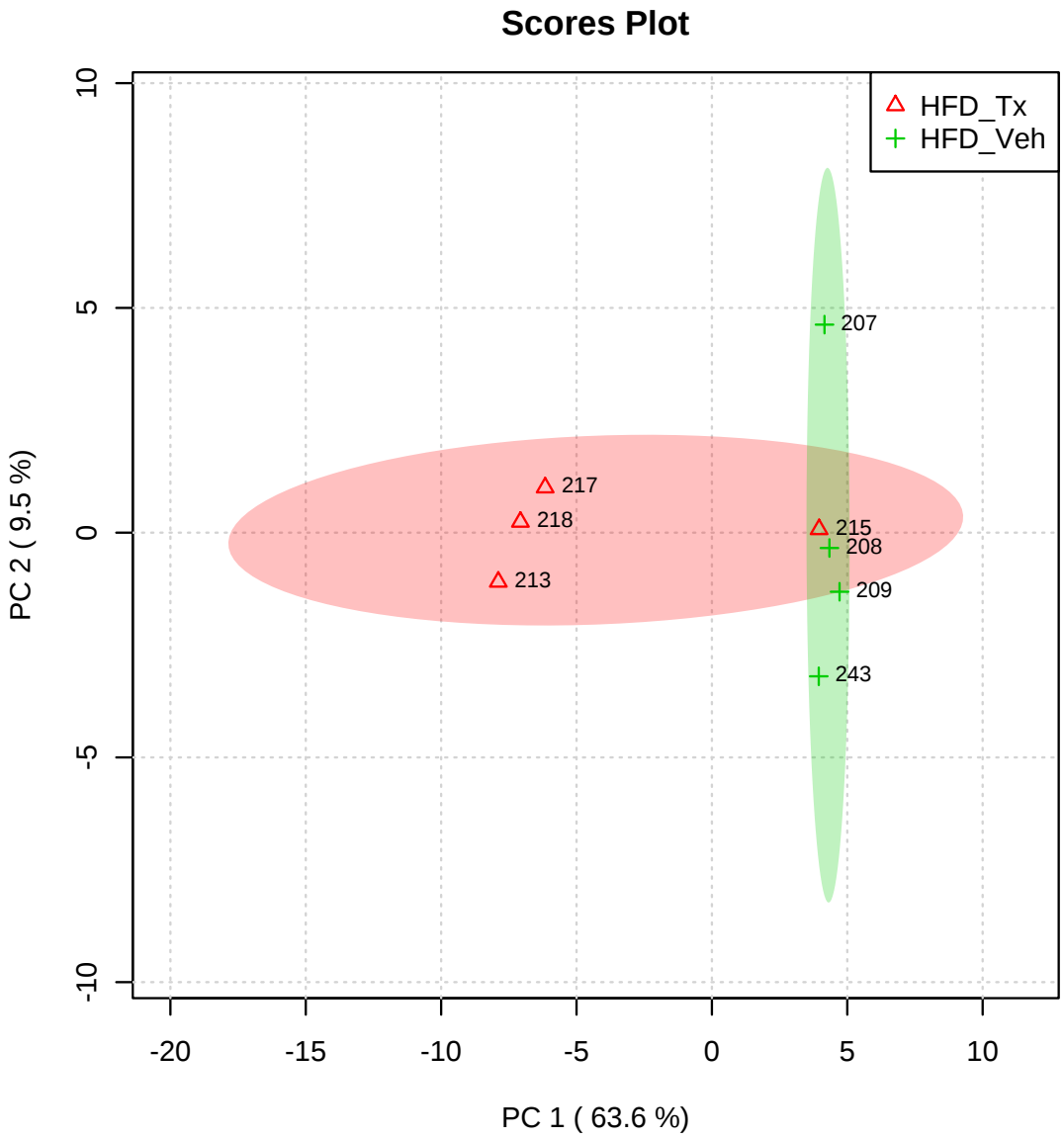

Figure 7: Scores plot between the selected PCs. The explained variances are shown in brackets.

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

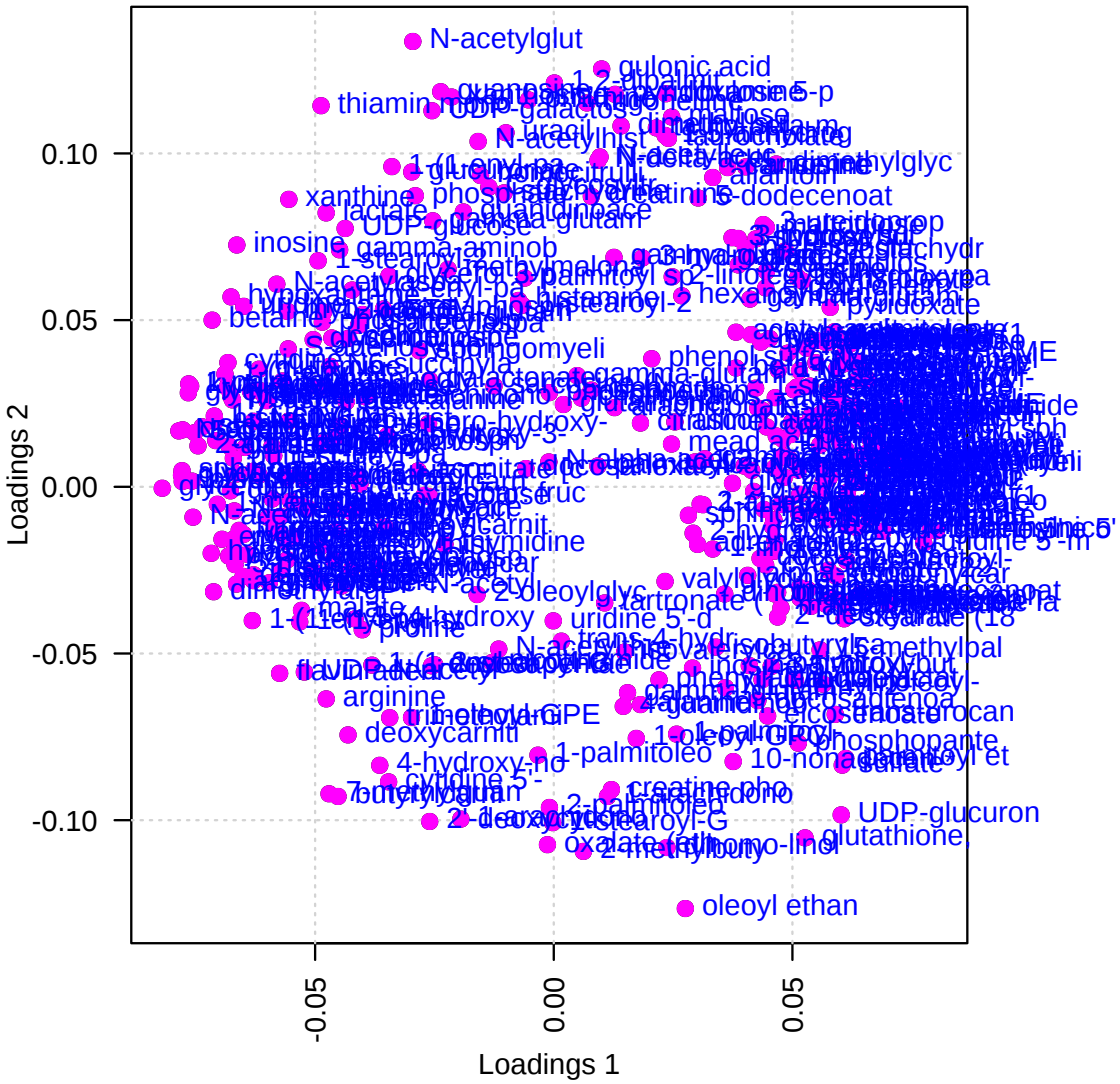

Figure 9: Loadings plot for the selected PCs.

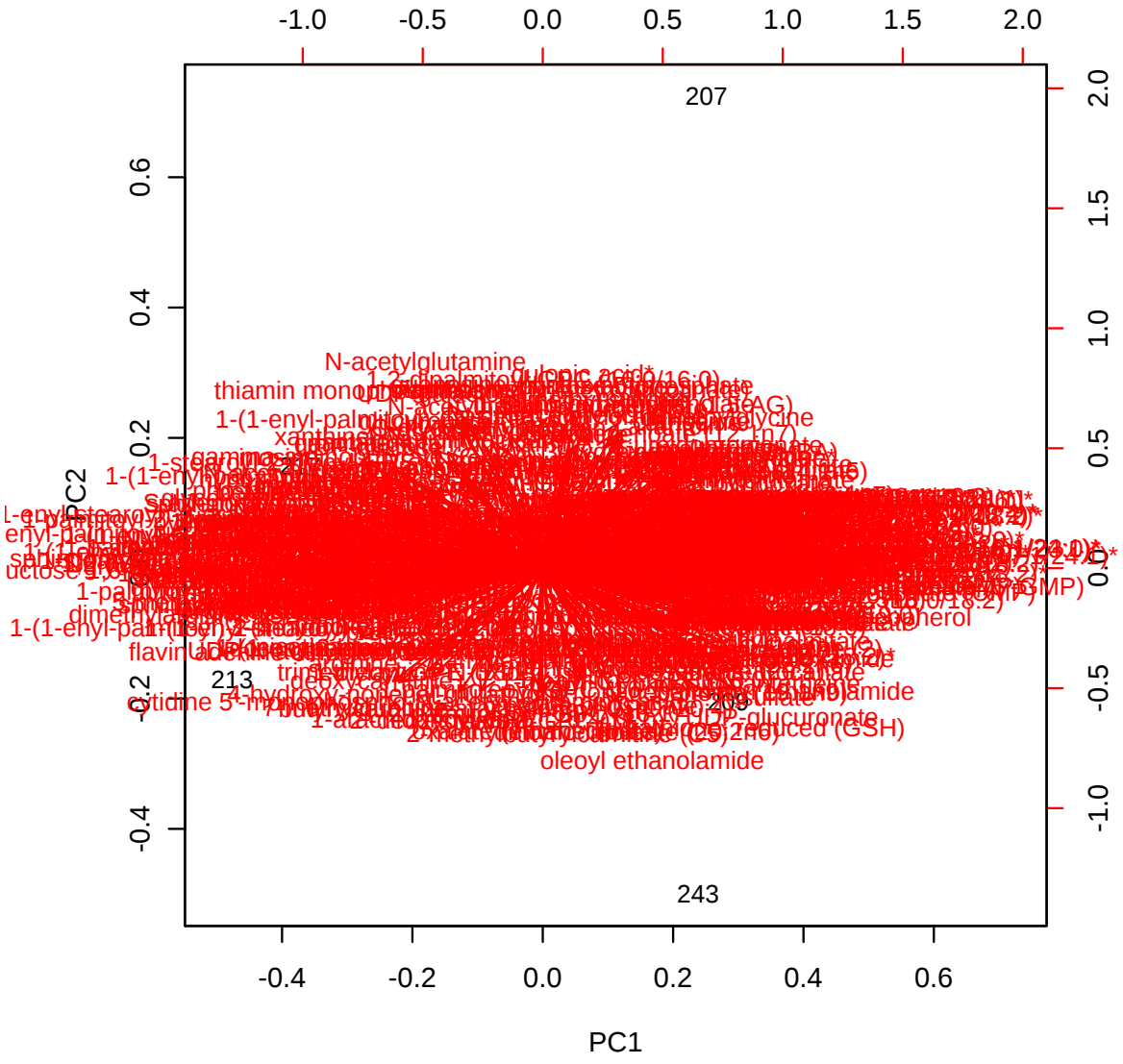

Figure 10: PCA biplot between the selected PCs. Note, you may want to test different centering and scaling normalization methods for the biplot to be displayed properly.

2.3 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[4]. The classification and cross-validation are performed using the corresponding wrapper function offered by the `caret` package[5].

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. MetaboAnalyst supports two types of test statistics for measuring the class discrimination. The first one is based on prediction accuracy during training. The second one is separation distance based on the ratio of the between group sum of the squares and the within group sum of squares (B/W-ratio). If the observed test statistic is part of the distribution based on the permuted class assignments, the class discrimination cannot be considered significant from a statistical point of view.[6].

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. Please note, VIP scores are calculated for each components. When more than componetnts are used to calculate the feature importance, the average of the VIP scores are used. 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. Please note, for multiple-group (more than two) analysis, the same number of predictors will be built for each group. Therefore, the coefficient of each feature will be different depending on which group you want to predict. The average of the feature coefficients are used to indicate the overall coefficient-based importance.

Figure 11 shows the overview of scores plots; Figure 12 shows the 2-D scores plot between selected components; Figure 13 shows the 3-D scores plot between selected components; Figure 14 shows the loading plot between the selected components;Figure 15 shows the classification performance with different number of components; Figure 16 shows the results of permutation test for model validation; Figure 17 shows important features identified by PLS-DA.

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

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

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

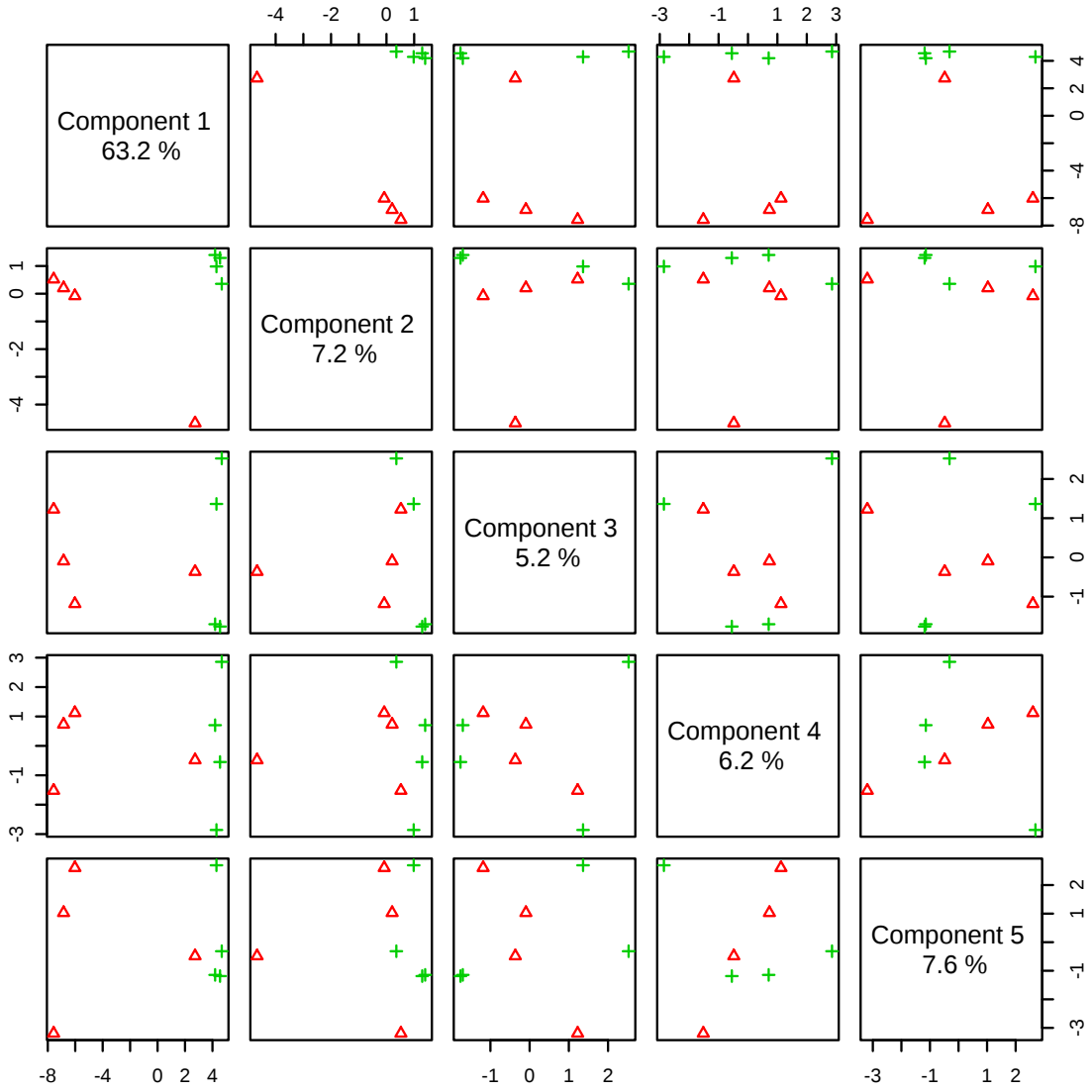

Figure 11: Pairwise scores plots between the selected components. The explained variance of each component is shown in the corresponding diagonal cell.

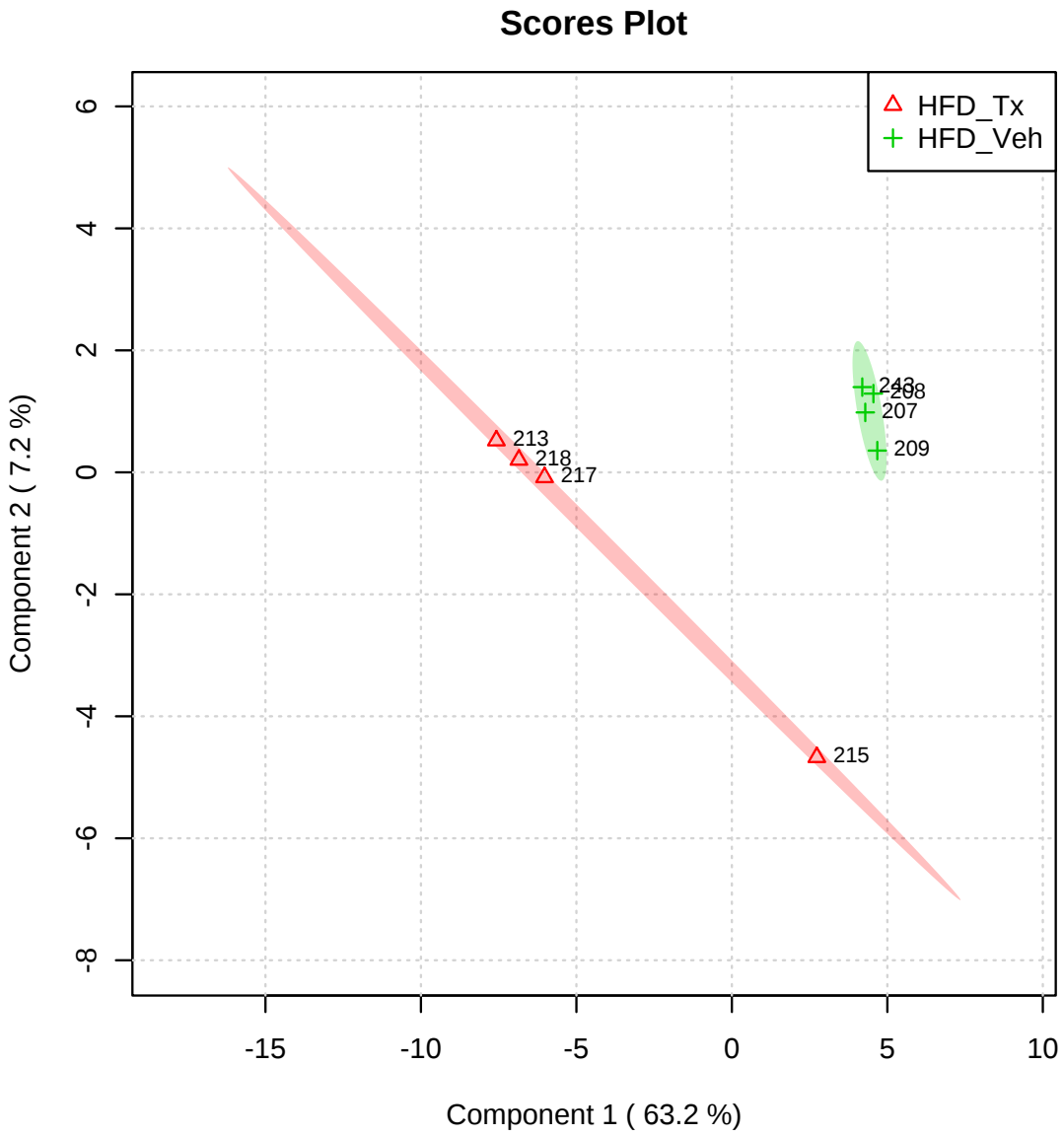

Figure 12: Scores plot between the selected PCs. The explained variances are shown in brackets.

Figure 13: 3D scores plot between the selected PCs. The explained variances are shown in brackets.

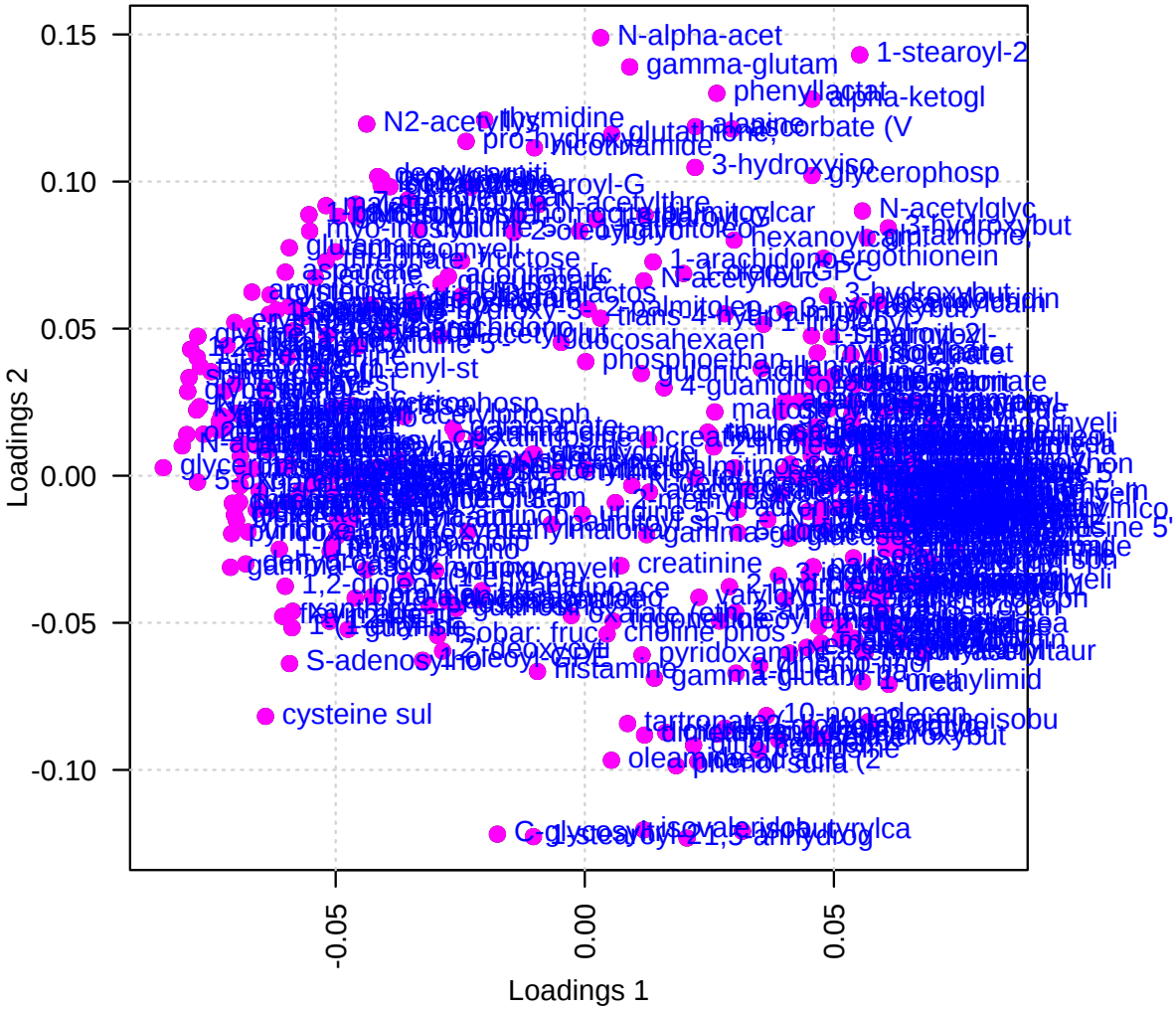

Figure 14: Loadings plot between the selected PCs.

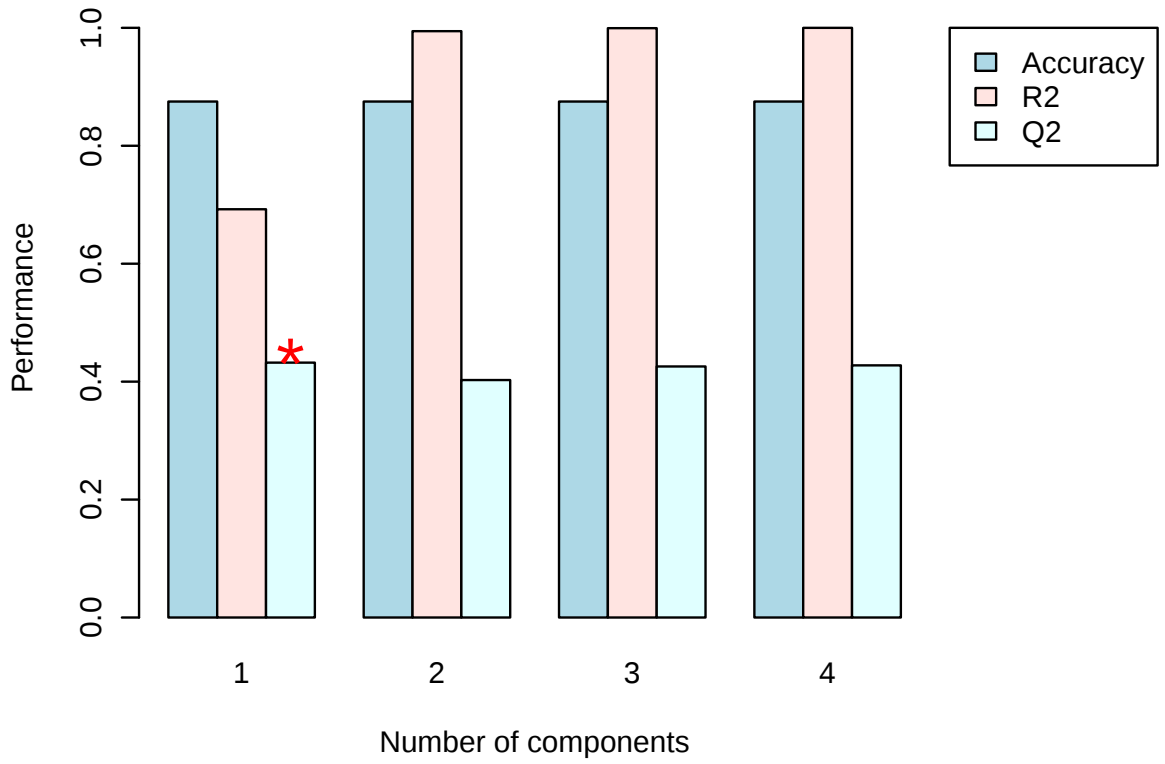

Figure 15: PLS-DA classification using different number of components. The red circle indicates the best classifier.

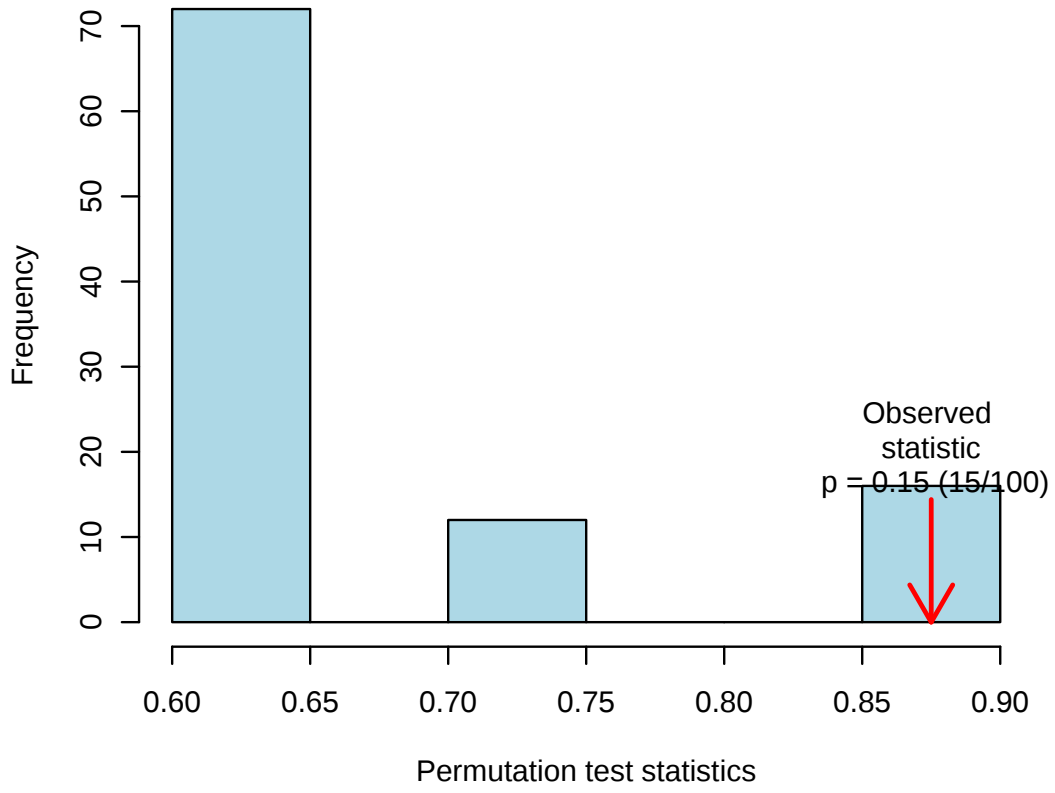

Figure 16: PLS-DA model validation by permutation tests based on prediction accuracy. The p value based on permutation is p = 0.15 (15/100).

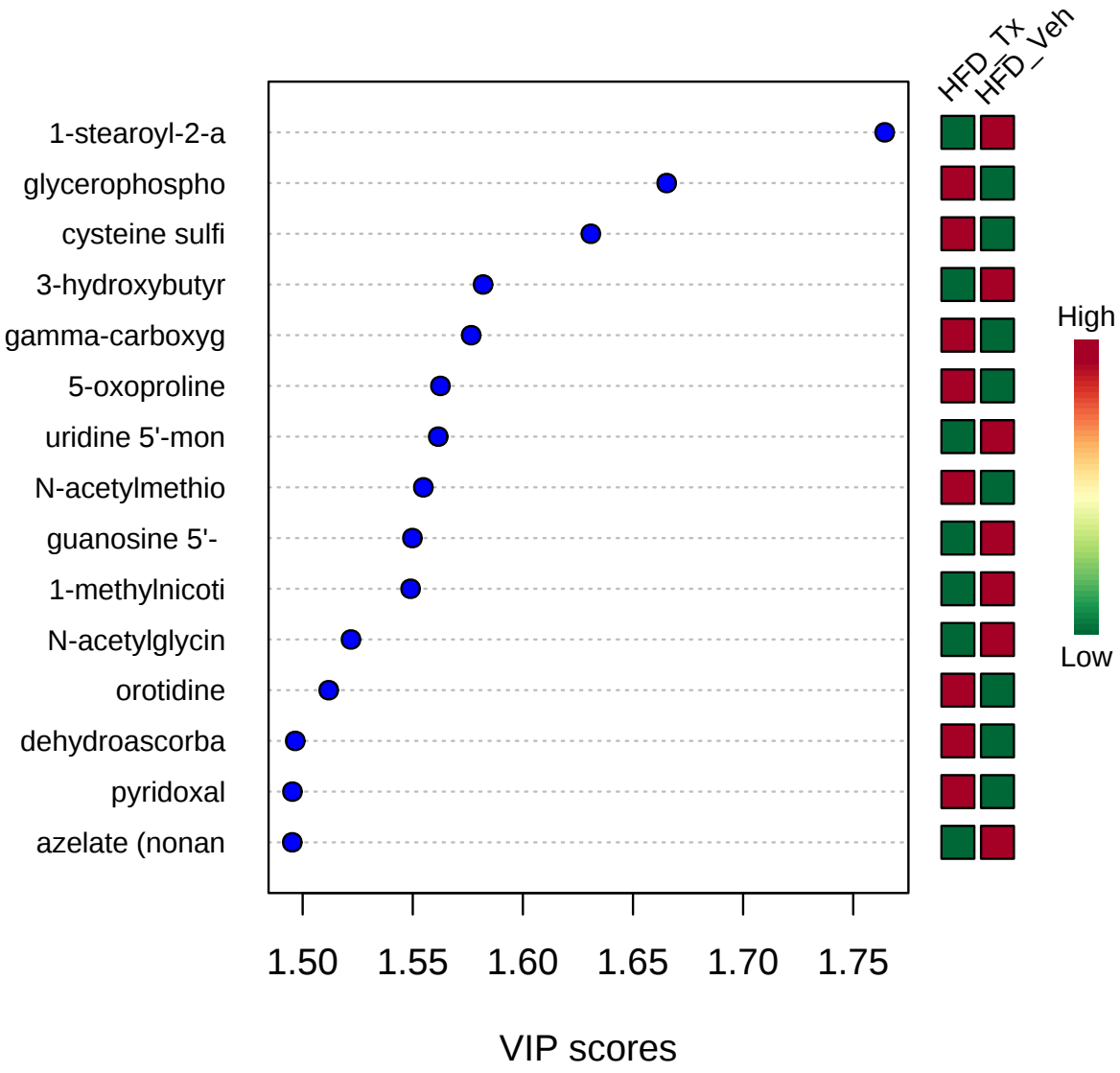

Figure 17: Important features identified by PLS-DA. The colored boxes on the right indicate the relative concentrations of the corresponding metabolite in each group under study.

2.4 Hierarchical Clustering

In (agglomerative) hierarchical cluster analysis, each sample begins as a separate cluster and the algorithm proceeds to combine them until all samples belong to one cluster. Two parameters need to be considered when performing hierarchical clustering. The first one is similarity measure - Euclidean distance, Pearson's correlation, Spearman's rank correlation. The other parameter is clustering algorithms, including average linkage (clustering uses the centroids of the observations), complete linkage (clustering uses the farthest pair of observations between the two groups), single linkage (clustering uses the closest pair of observations) and Ward's linkage (clustering to minimize the sum of squares of any two clusters). Heatmap is often presented as a visual aid in addition to the dendrogram.

Hierachical clustering is performed with the `hclust` function in package `stat`. Figure 18 shows the clustering result in the form of a dendrogram. Figure 19 shows the clustering result in the form of a heatmap.

Figure 18: Clustering result shown as heatmap (distance measure using `euclidean`, and clustering algorithm using `ward`).

2.5 K-means Clustering

K-means clustering is a nonhierarchical clustering technique. It begins by creating k random clusters (k is supplied by user). The program then calculates the mean of each cluster. If an observation is closer to the centroid of another cluster then the observation is made a member of that cluster. This process is repeated until none of the observations are reassigned to a different cluster.

K-means analysis is performed using the `kmeans` function in the package `stat`. Figure 20 shows clustering the results. Table 5 shows the members in each cluster from K-means analysis.

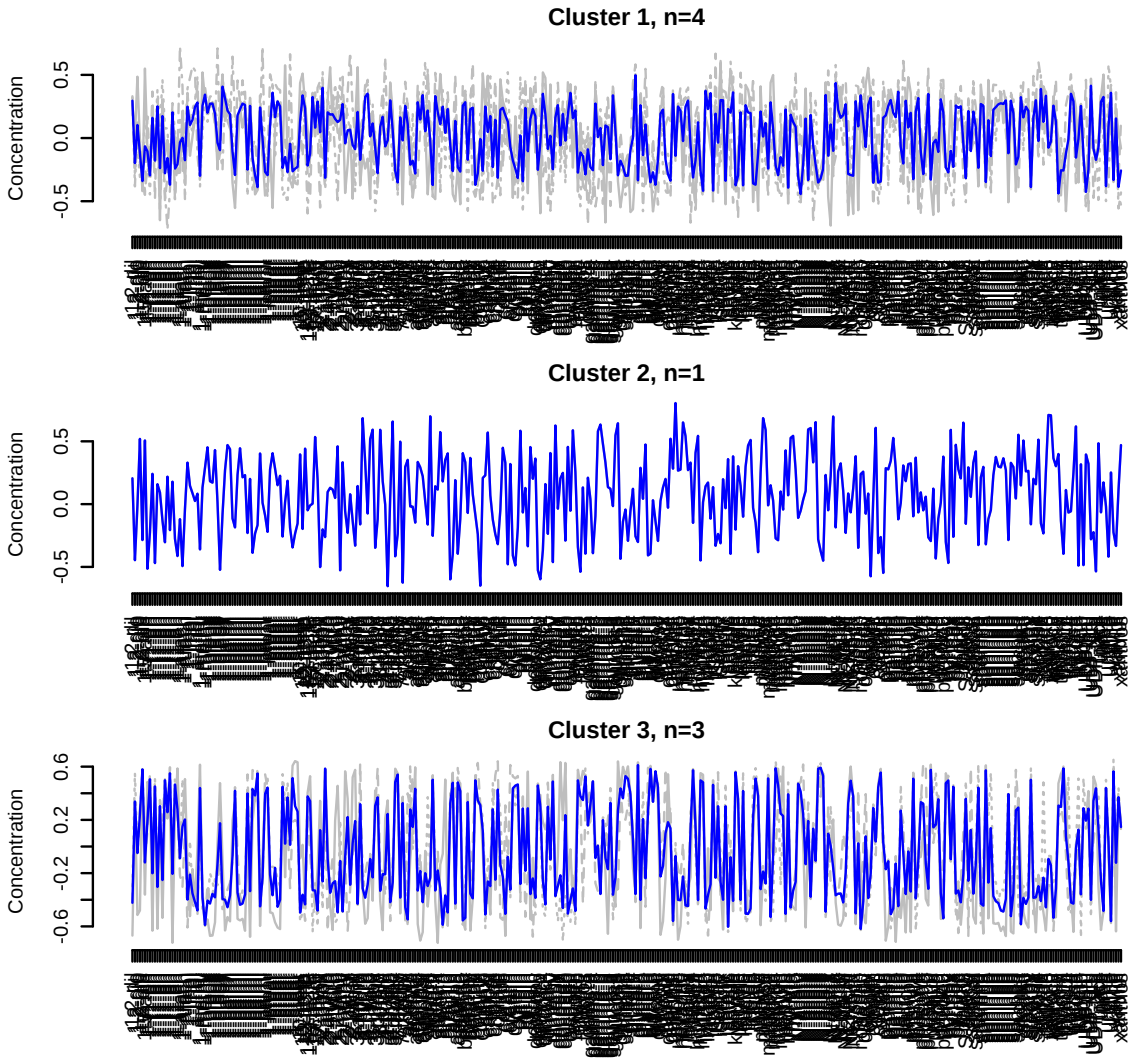

Figure 19: K-means cluster analysis. The x-axes are variable indices and y-axes are relative intensities. The blue lines represent median intensities of corresponding clusters

Table 5: Clustering result using K-means

	Samples in each cluster
Cluster(1)	215 208 209 243
Cluster(2)	207
Cluster(3)	213 217 218

2.6 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[7]. Table 6 shows the confusion matrix of random forest. Figure 21 shows the cumulative error rates of random forest analysis for given parameters. Figure 22 shows the important features ranked by random forest. Figure 23 shows the outlier measures of all samples for the given parameters. The OOB error is 0.125

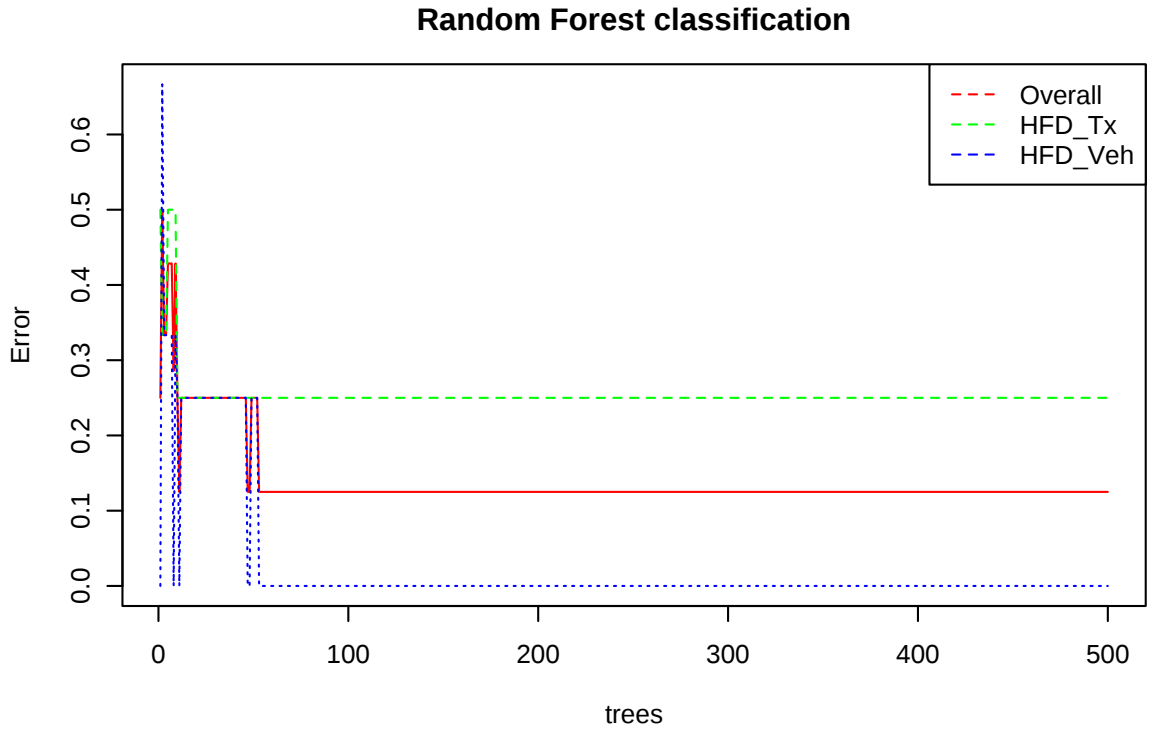

Figure 20: Cumulative error rates by Random Forest classification. The overall error rate is shown as the black line; the red and green lines represent the error rates for each class.

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

Table 6: Random Forest Classification Performance

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

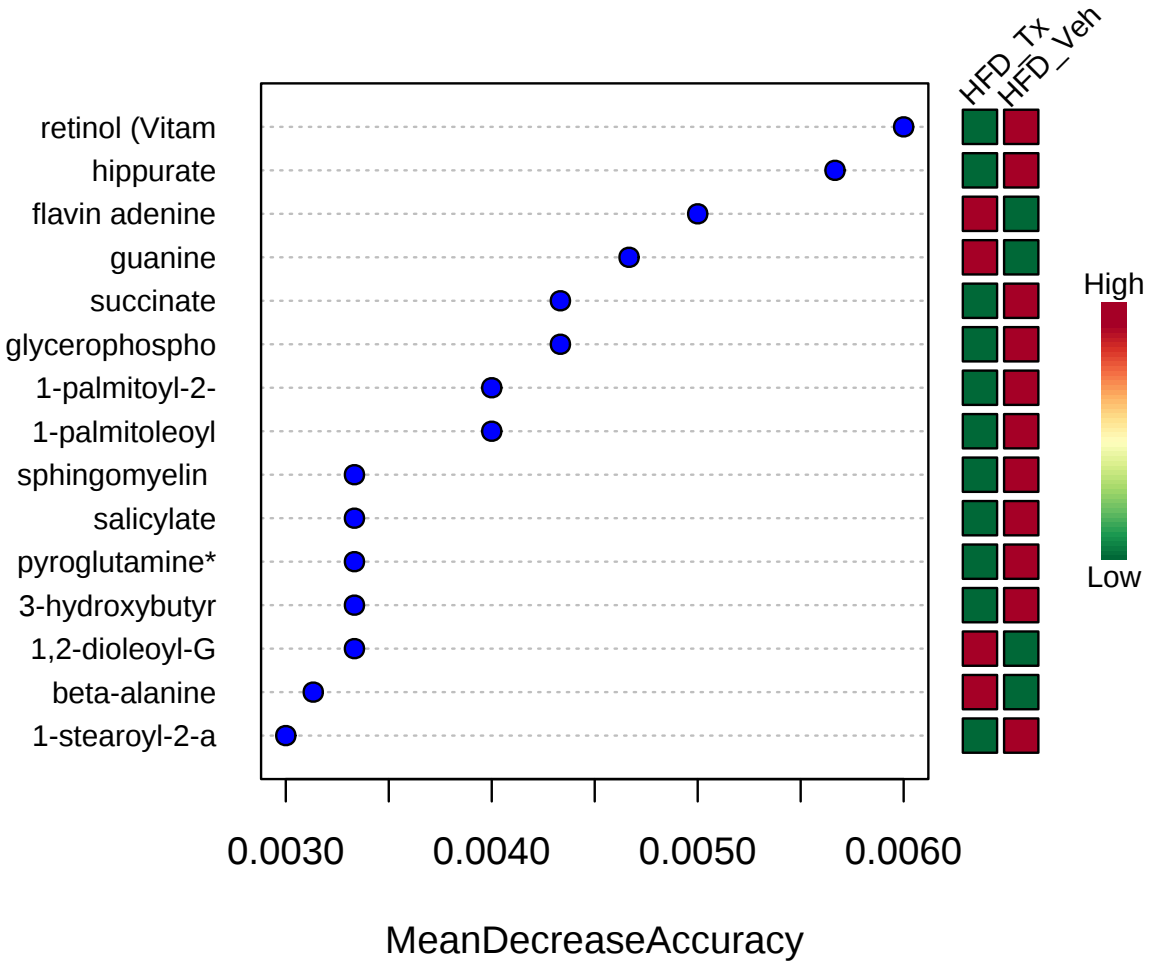

Figure 21: Significant features identified by Random Forest. The features are ranked by the mean decrease in classification accuracy when they are permuted.

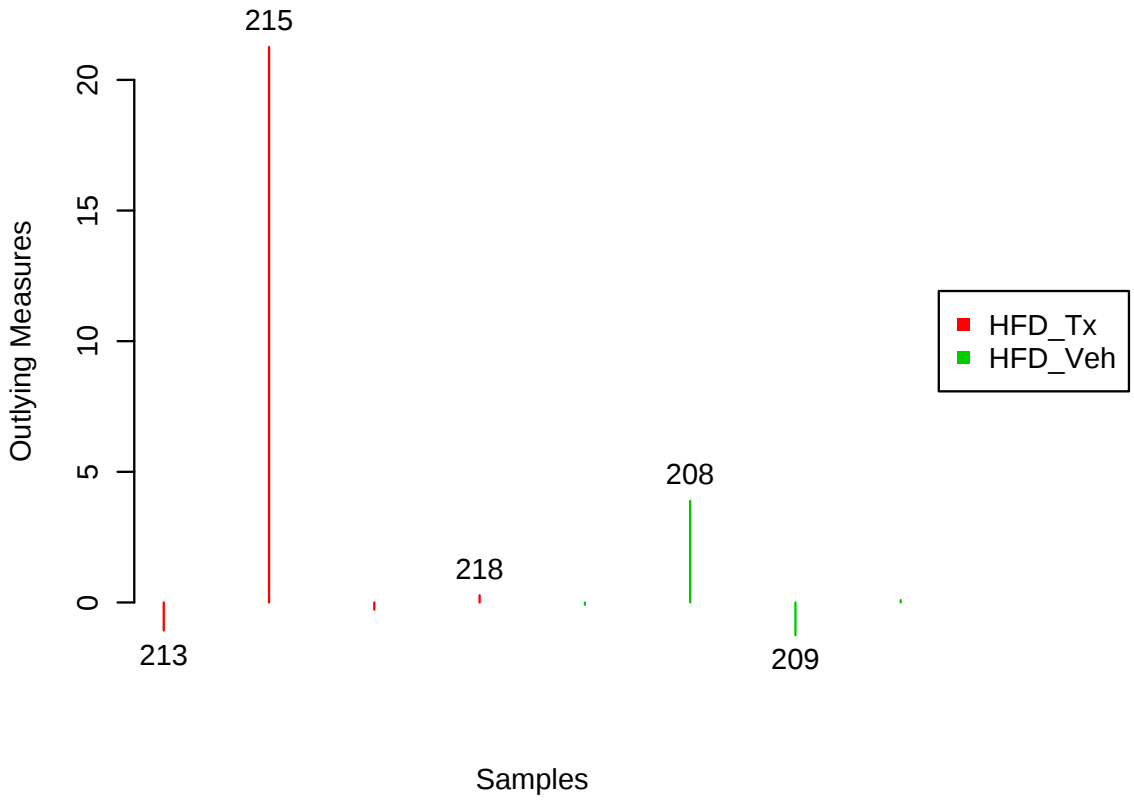

Figure 22: Potential outliers identified by Random Forest. Only the top five are labeled.

2.7 Support Vector Machine (SVM)

SVM aims to find a nonlinear decision function in the input space by mapping the data into a higher dimensional feature space and separating it there by means of a maximum margin hyperplane. The SVM-based recursive feature selection and classification is performed using the `R-SVM` script[8]. The process is performed recursively using decreasing series of feature subsets (`ladder`) so that different classification models can be calculated. Feature importance is evaluated based on its frequencies being selected in the best classifier identified by recursive classification and cross-validation. Please note, R-SVM is very computationally intensive. Only the top 50 features (ranked by their p values from t-tests) will be evaluated.

In total, 10 models (levels) were created using 396, 158, 79, 40, 24, 18, 14, 10, 8, 6 selected feature subsets. Figure 24 shows the SVM classification performance using recursive feature selection. Figure 25 shows the signicant features used by the best classifiers.

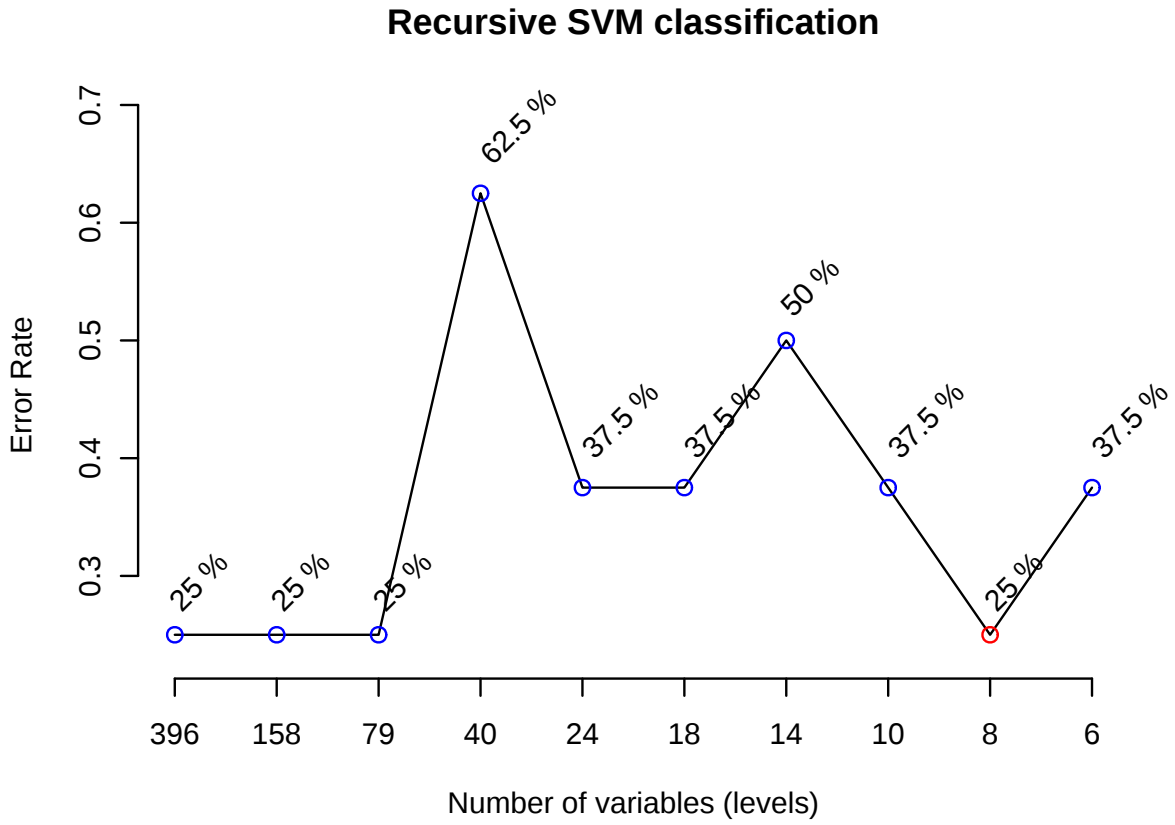

Figure 23: Recursive classification with SVM. The red circle indicates the best classifier.

[8] http://www.hsph.harvard.edu/bioinfocore/RSVMhome/R-SVM.html

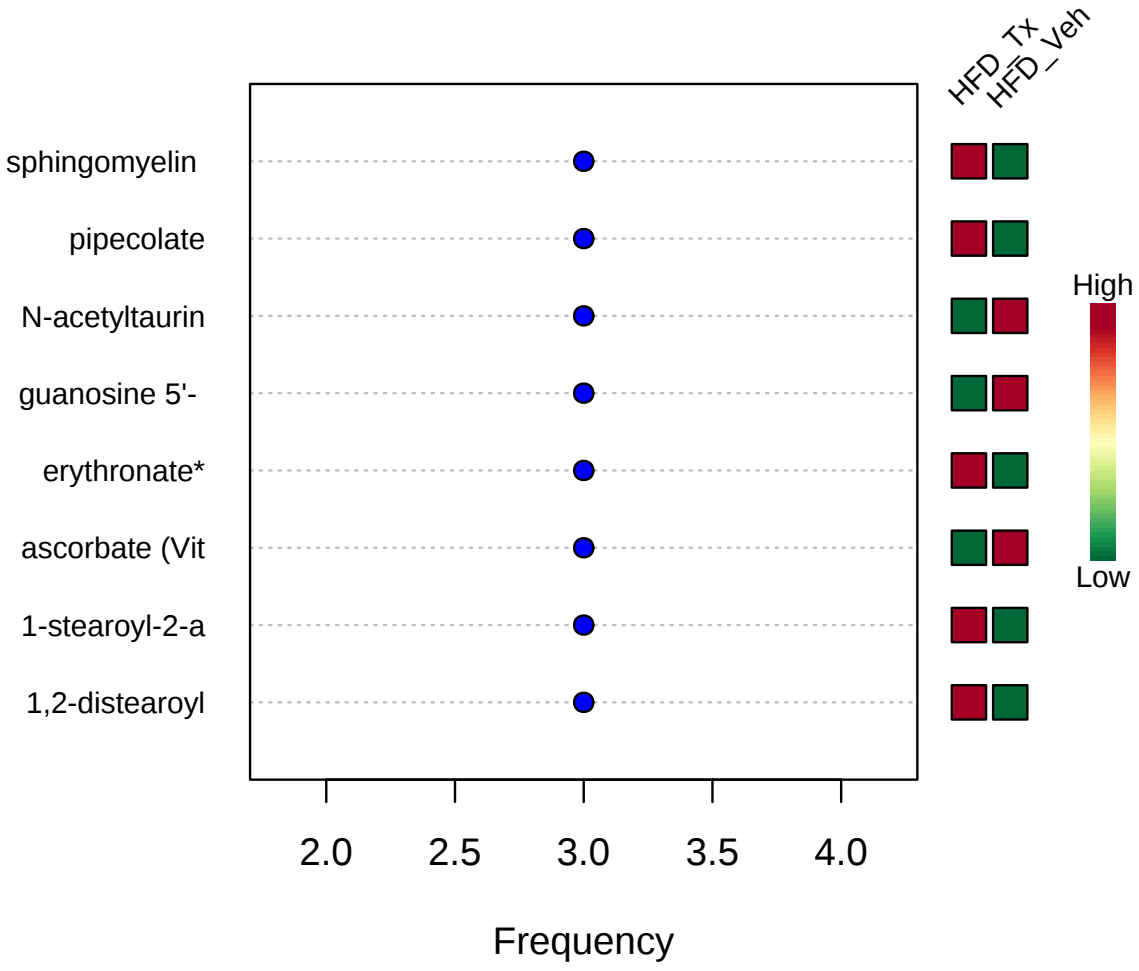

Figure 24: Significant features identified by R-SVM. Features are ranked by their frequencies of being selected in the classifer.

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 Thu Sep 24 17:27:49 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*).