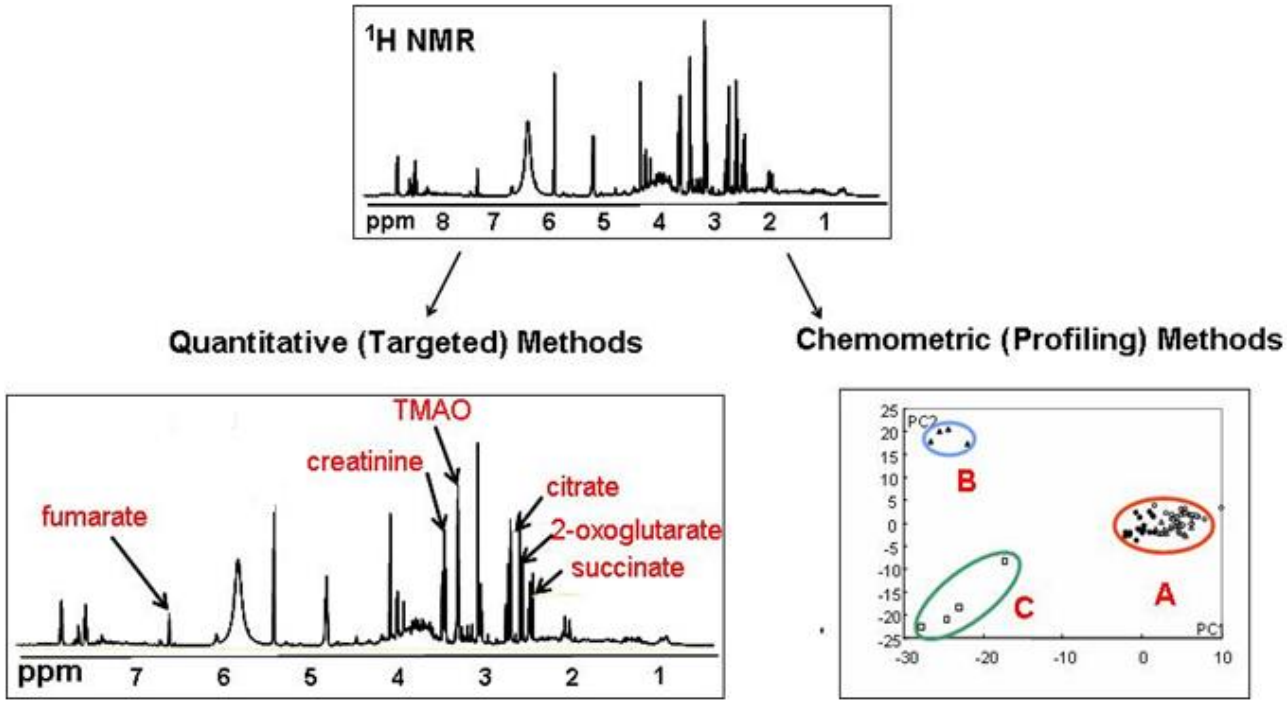
An Introduction to NMR-based Metabolomics and Statistical Data Analysis using MetaboAnalyst

Rohit Mahar (PhD)

Department of Biochemistry and molecular Biology

University of Florida

Routes of Metabolomics



General routes of metabolomics

Chemometrics or Multivariate Statistical Analysis: "Chemometrics is the chemical discipline that uses mathematics and statistics to design experimental procedures for maximum relevant chemical information by analyzing chemical data and obtain knowledge about chemical systems"

(1) Pattern recognition (2) Sample classification

NMR spectroscopy along with Chemometrics play a vital role in the field of Metabolomics.

Procedure of NMR-based Metabolomics

^{1}H-NMR analysis

- 2D-NMR Measurements*
  - Structure elucidation
    - Quantitative NMR for Target Metabolites
- Data bucketing
  - Multivariate data analysis
    - Signal sorting (using score and loading plots)

2D-NMR Measurements*: J-resolved, COSY, DQF-COSY, TOCSY, HSQC, HMBC, HSQC-TOCSY

Metabolites

Primary metabolites
- Amino acids
- Sugars
- Nucleic acids
- Lipids

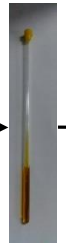
Secondary metabolites
- Cardiac glycosides
- Alkaloids
- Terpenoids
- Flavanoids
- Calotroposides

Tissue (e.g. liver)

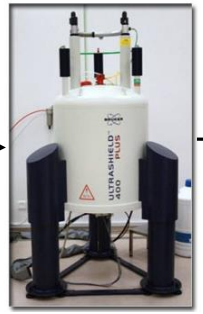
Extraction

Fractionation

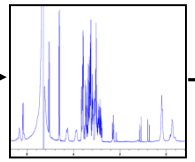
Sample preparation



NMR tube

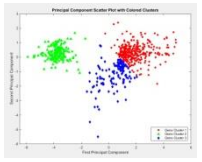


NMR magnet

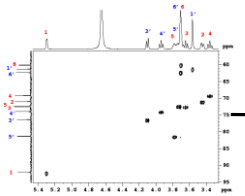


^{1}H NMR

Data Bucketing



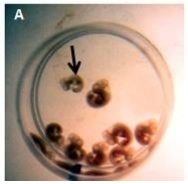
Multivariate analysis (PCA & PLS-DA)



2D NMR measurements

Structure elucidation

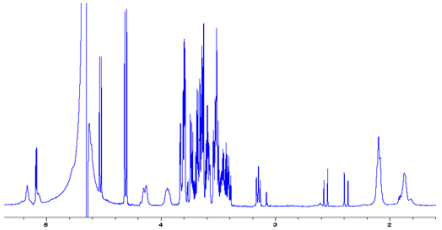
Overview of Solution state NMR-based metabolomics



Sample (rat embryo)

HR-MAS rotor

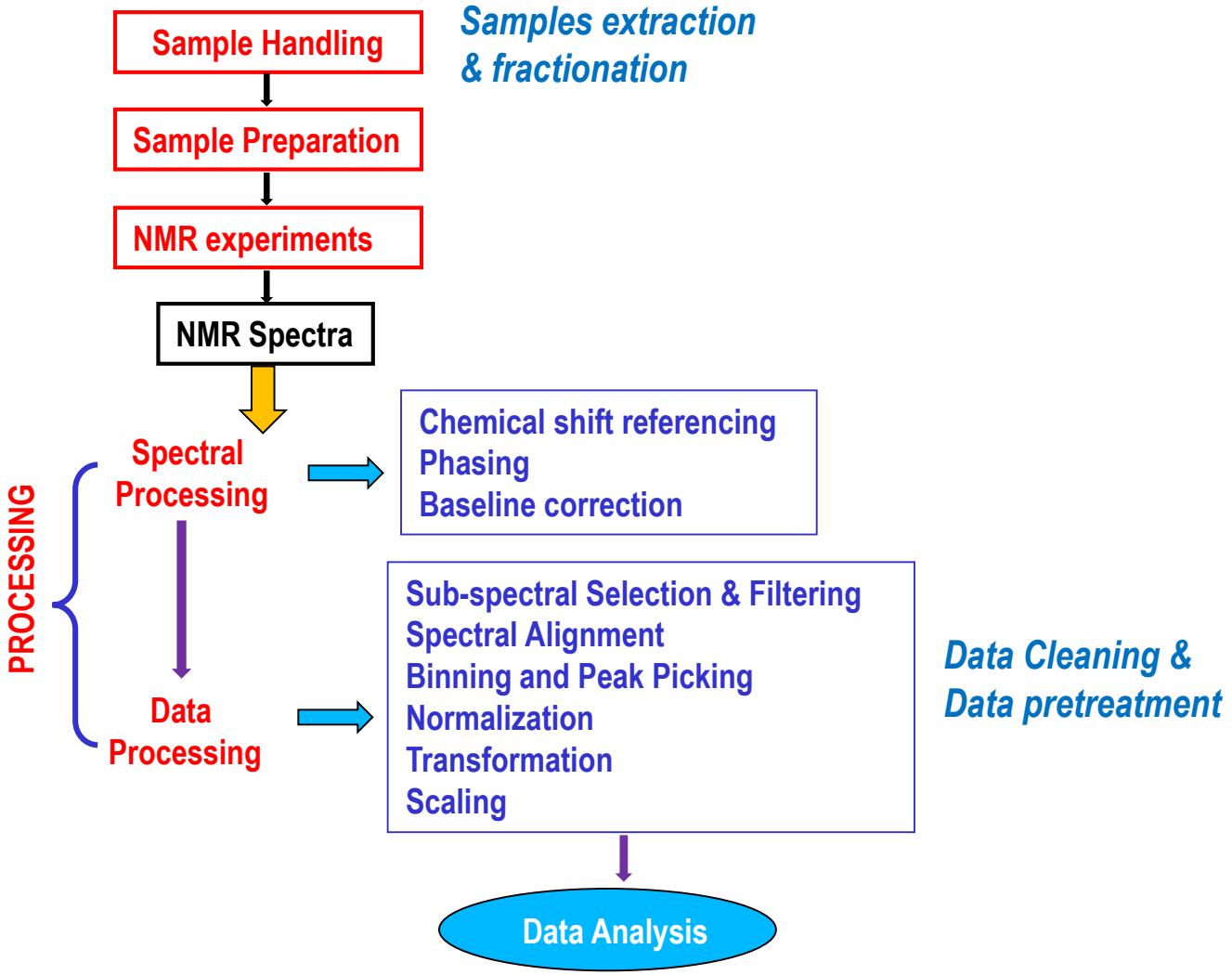
HR-MAS Probe



^{1}H NMR spectrum

Method to perform an experiment on HR-MAS NMR

- HR-MAS ^{1}H NMR spectroscopy provides better resolution of spectrum in case of semisolid samples.
- Solution state NMR needs solvent extraction of metabolites from tissues, which overcomes by HR-MAS.



Summary of spectral processing and post-processing steps on NMR-data in Metabolomics

Raja Roy and David S. Wishart; Metabolomics (2018) 14:31

Data Integrity Check:

- Checking the class labels - at least three replicates are required in each class.
- If the samples are paired, the pair labels must conform to the specified format.
- The data (except class labels) must not contain non-numeric values.
- The presence of missing values or features with constant values (i.e. all zeros).

Data Cleaning:

- This step is strongly recommended for untargeted metabolomics.
- Datasets (i.e. spectral binning data, peak lists) with large number of variables, many of them are from baseline noises.
- Data Filtering methods are used to remove bins that are null and do not displays any changes among spectra series.

Non-informative variables can be characterized in three groups:

1) Variables of very small values (close to baseline or detection limit) - these variables can be detected using mean or median.
2) Variables that are near-constant values throughout the experiment conditions (housekeeping or homeostasis) - these variables can be detected using standard deviation (SD); or the robust estimate such as interquantile range (IQR).
3) Variables that show low repeatability - this can be measured using the relative standard deviation (RSD = SD/mean).

Rules for Data Cleaning:

- Filtering Variable (bin) shows zeros among all rows (spectra) is discarded.
- In practice Standard Deviation, Median Absolute Deviation and Interquartile Range are calculated for all bins.
- In Standard Deviation, MAD and IQR a fixed fraction (default 10%) of the bins is discarded (e.g. if the matrix is composed by 100 bins it means that 10 bins are discarded, and the selection is based on the Filter method chosen).
- An amount of bins (with the lowest SD, or MAD or IQR values) are discarded with respect a percentage value of the total bins.

The following empirical rules are applied during data filtering:

- Less than 250 variables: 5% will be filtered.
- Between 250 - 500 variables: 10% will be filtered.
- Between 500 - 1000 variables: 25% will be filtered.
- Over 1000 variables: 40% will be filtered.

Normalization: Why normalization is essential?

To ensure that peak intensities are relatively similar from sample to sample.

Correction for sample variation due to sample dilution or sample concentration (technical or biological).

Methods of normalization

Sample normalization (row-wise)

To remove systematic variation between experimental conditions unrelated to the biological differences (i.e. dilutions, mass).

Feature normalization (column-wise)

To bring variances of all features close to equal.

Sample normalization:

Row-wise normalization aims to normalize each sample (row) so that it is comparable to the other.

it is an operation that is performed on the rows of the matrix

Sample-specific normalization (i.e. weight, volume)

Normalization by sum

every element on a row is divided by the sum of all elements of the same row

Normalization by median

every element on a row is reduced by the median value of all the bins that constitute the same row

Normalization by reference sample (PQN)

every element of a row is divided by the corresponding element of the row of the selected reference spectrum.

A most probable quotient between the signals of the corresponding spectrum and of a reference spectrum is calculated as normalization factor.

Normalization by a pooled sample from group

If you select a bundle of spectra (like all spectra belonging to the same class) normalization is performed on the calculated average spectrum.

Normalization by reference feature (selected bin)

Feature normalization (column-wise)

Data transformation and scaling are two different approaches to make individual features more comparable

Data transformation

Correction for deviation from normality and uneven variance (*heteroscedasticty*)

Log transformation: (generalized logarithm transformation or glog)

Cube root transformation: (take cube root of data values)

Data scaling

Mean centering: mean-centered only

Auto scaling: mean-centered and divided by the 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 range of each variable.

Robert A van den Berg, *BMC Genomics* 2006, **7**:142.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set.

PCA uses orthogonal transformation to convert a set of observations from correlated variables into a set of values of linearly uncorrelated variables (principal components).

PCA can be used to answer questions such as:

1) **What is the pattern of sample distribution?**

2) **Which variables describe the differences between samples?**

3) **Which variables contribute most to an observed difference?**

4) **Which variables contribute in the same way (i.e. are correlated)?**

Interpretation of Scores and Loadings

Relationship of Scores (samples) information to Loadings (variables) information.

PCA model is characterized by three complementary sets of attributes:

Scores

Scores describe the properties of the samples and are usually shown as a map of one PC plotted against another.

Loadings

Loadings describe the relationships between variables and may be plotted as a line (commonly used in spectral data interpretation) or a map (commonly used in process or sensory data analysis).

Explained (or Residual) Variances

These are error measures that tell how much information is taken into account by each PC.

Partial-least squares discriminant analysis (PLS-DA):

PLS-DA is a supervised method that uses multiple linear regression technique to find the direction of maximum covariance between a data set (X) and the class membership (Y).

- **PLS discriminant analysis is used to explain and predict the membership of observations to several classes using quantitative or qualitative explanatory variables or parameters.**
- **The quality of the mathematical model was described by the cross-validation parameters R^2 and Q^2.**
- **R^2 is the goodness of fit and Q^2 indicates the predictive ability and indicates the robustness of model.**
- **Typically, Q^2 is lower than R^2 for the PLS-DA.**

- **PLS-DA tends to overfit the data and therefore the model needs to be validated to see whether the separation is statistically significant or is due to random noise.**

- **In each permutation, a PLS-DA model is 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.**

Case Study

Statistical Data Analysis on Murine's Urine Samples

Two Groups of samples

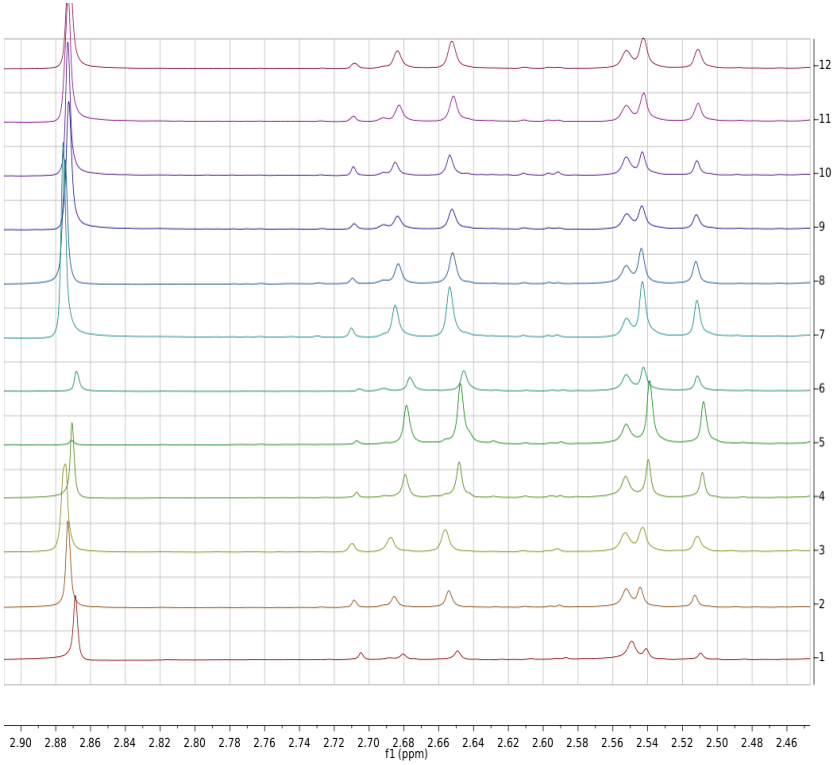
1. **KO_D3P**
2. **WT_D1A**

Spectral misalignment: **Peak positions are affected by:**

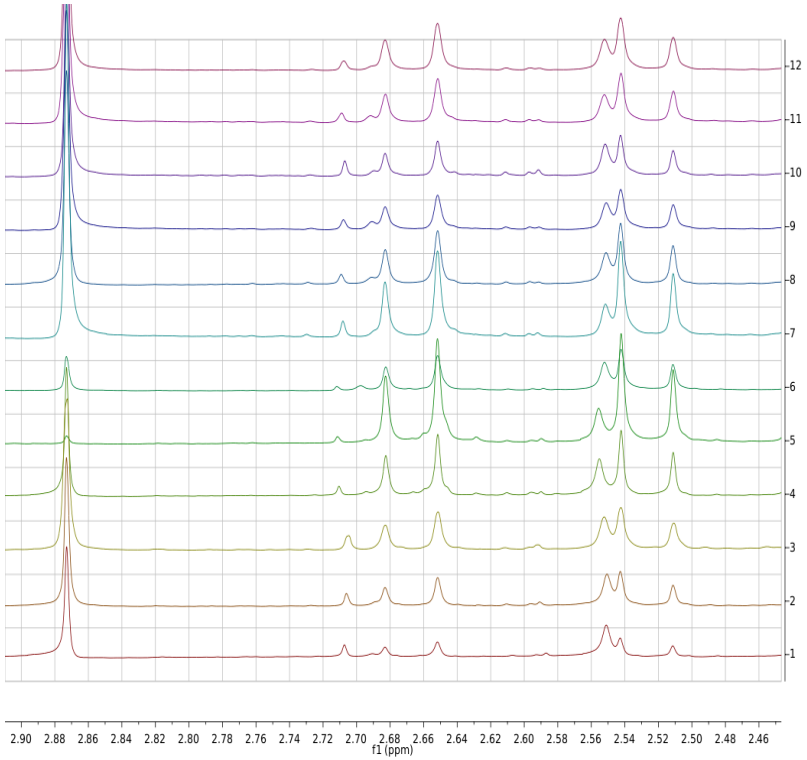
1. pH (Changes charge states), 2. Salt effects (Metal Chelation),

3. Line shapes (shimming), 4. Human and instrumental effects

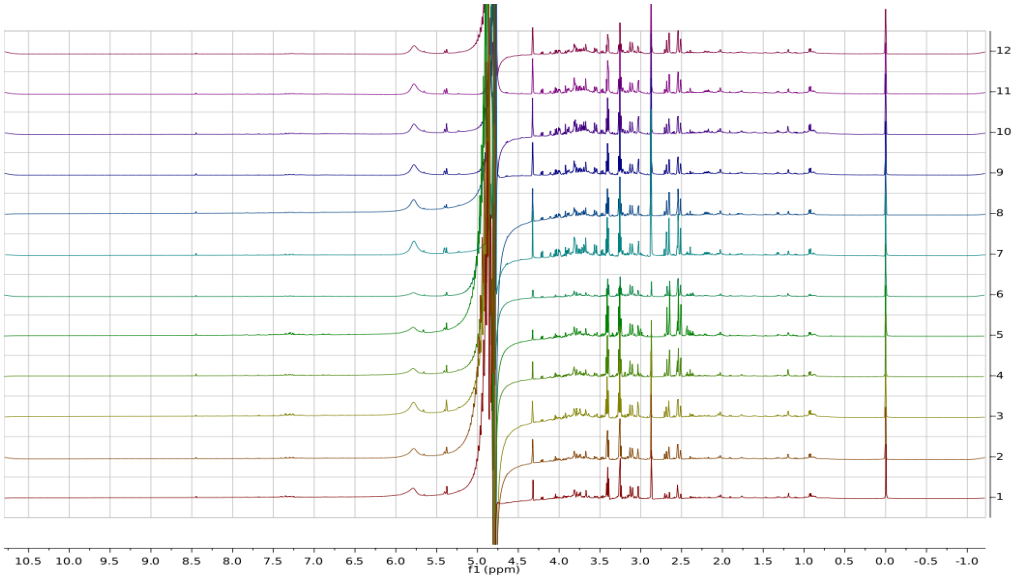
5. Line widths (shimming, diffusion , chemical exchange)



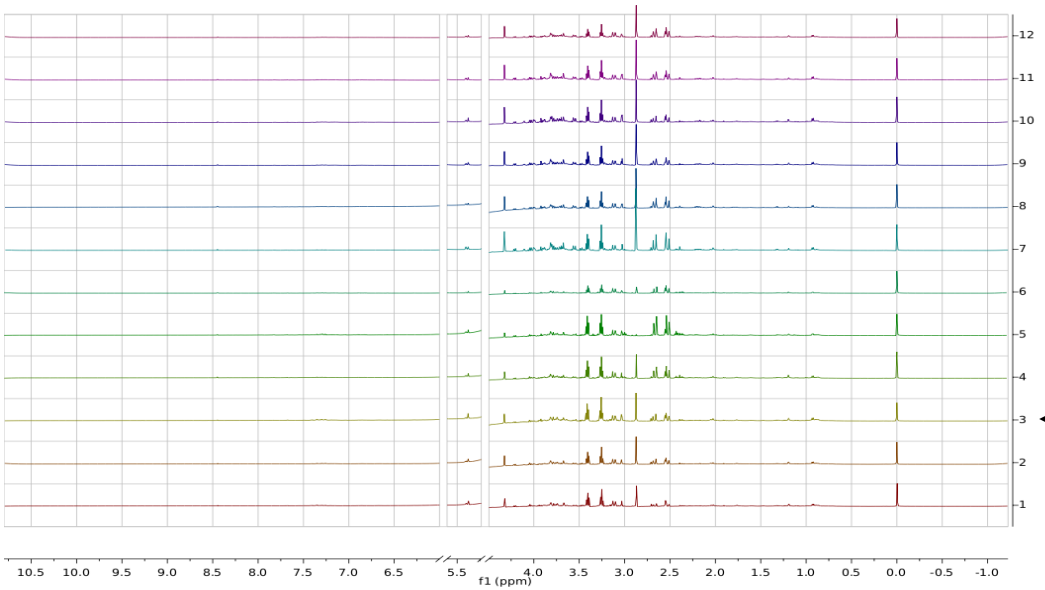
Before Alignment



After Alignment



NMR spectra



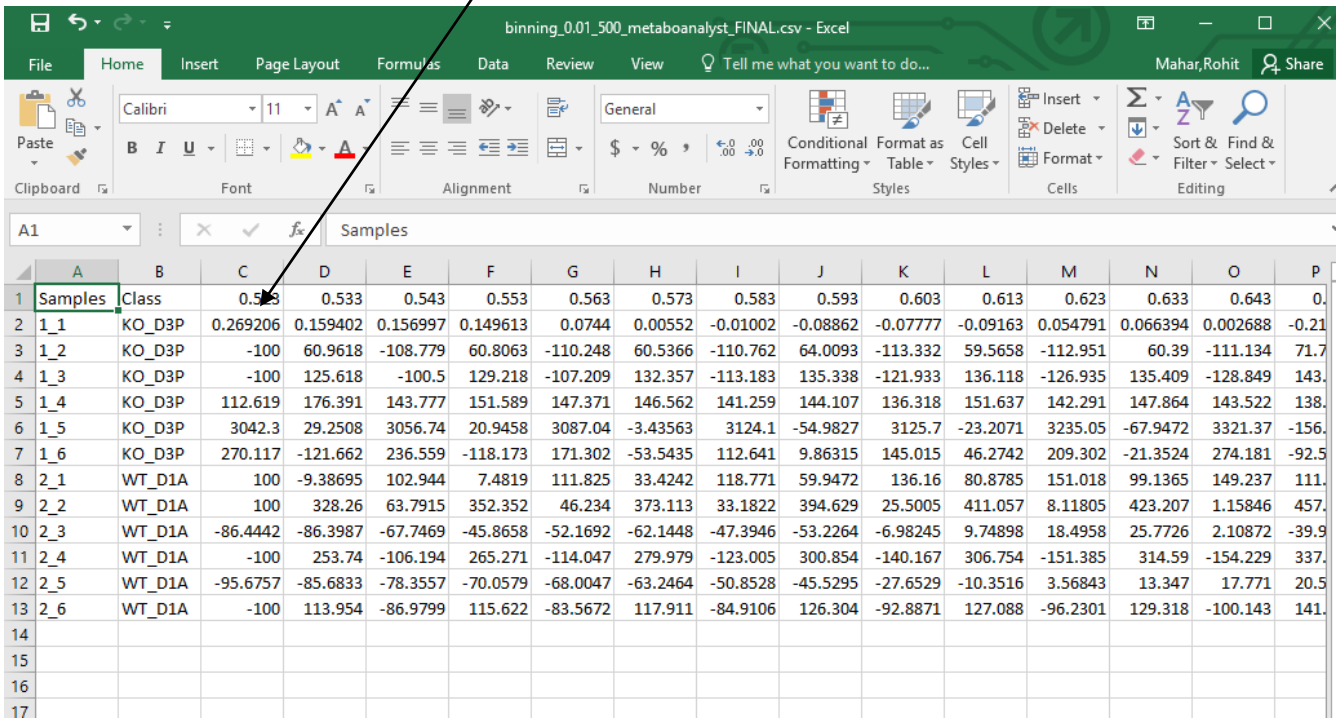
NMR spectra after excluded areas

Binning (bucketing) of Data

Bucket width = 0.01 ppm

Range of spectrum = 0.50 to 9.5 ppm

bucket (variable)



binning_0.01_500_metaboanalyst_FINAL.csv - Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Samples	Class	0.5	0.533	0.543	0.553	0.563	0.573	0.583	0.593	0.603	0.613	0.623	0.633	0.643	0.
2	1_1	KO_D3P	0.269206	0.159402	0.156997	0.149613	0.0744	0.00552	-0.01002	-0.08862	-0.07777	-0.09163	0.054791	0.066394	0.002688	-0.21
3	1_2	KO_D3P	-100	60.9618	-108.779	60.8063	-110.248	60.5366	-110.762	64.0093	-113.332	59.5658	-112.951	60.39	-111.134	71.7
4	1_3	KO_D3P	-100	125.618	-100.5	129.218	-107.209	132.357	-113.183	135.338	-121.933	136.118	-126.935	135.409	-128.849	143.
5	1_4	KO_D3P	112.619	176.391	143.777	151.589	147.371	146.562	141.259	144.107	136.318	151.637	142.291	147.864	143.522	138.
6	1_5	KO_D3P	3042.3	29.2508	3056.74	20.9458	3087.04	-3.43563	3124.1	-54.9827	3125.7	-23.2071	3235.05	-67.9472	3321.37	-156.
7	1_6	KO_D3P	270.117	-121.662	236.559	-118.173	171.302	-53.5435	112.641	9.86315	145.015	46.2742	209.302	-21.3524	274.181	-92.5
8	2_1	WT_D1A	100	-9.38695	102.944	7.4819	111.825	33.4242	118.771	59.9472	136.16	80.8785	151.018	99.1365	149.237	111.
9	2_2	WT_D1A	100	328.26	63.7915	352.352	46.234	373.113	33.1822	394.629	25.5005	411.057	8.11805	423.207	1.15846	457.
10	2_3	WT_D1A	-86.4442	-86.3987	-67.7469	-45.8658	-52.1692	-62.1448	-47.3946	-53.2264	-6.98245	9.74898	18.4958	25.7726	2.10872	-39.9
11	2_4	WT_D1A	-100	253.74	-106.194	265.271	-114.047	279.979	-123.005	300.854	-140.167	306.754	-151.385	314.59	-154.229	337.
12	2_5	WT_D1A	-95.6757	-85.6833	-78.3557	-70.0579	-68.0047	-63.2464	-50.8528	-45.5295	-27.6529	-10.3516	3.56843	13.347	17.771	20.5
13	2_6	WT_D1A	-100	113.954	-86.9799	115.622	-83.5672	117.911	-84.9106	126.304	-92.8871	127.088	-96.2301	129.318	-100.143	141.
14																
15																
16																
17																

The uploaded data table contains 12 (samples) by 771 (spectra bins) data matrix.

MetaboAnalyst

Work flow of the MetaboAnalyst

http://www.metaboanalyst.ca

MetaboAnalyst - statistical, functional and integrative analysis of metabolomics data

Welcome >> click here to start <<

News & Updates

- PLEASE TAKE OUR SURVEY >> Help us to help you!
- Upcoming metabolomics conference - **Metabolomics 2018** (June 24-28 Seattle, Washington). A summary of the program is available here; NEW
- Upcoming metabolomics workshop - **Informatics and Statistics for Metabolomics** (June 11-12, 2018 Montreal QC) NEW
- Check out our **MicrobiomeAnalyst** for comprehensive analysis of microbiome data;
- Fixed the issue for sample hold-out analysis in biomarker analysis (04/23/2018); NEW
- Fixed the issue with time-series group ordering based on numeric values (04/19/2018); NEW
- Enhanced support for SVG export for KEGG global network (04/04/2018); NEW
- Updated Resources on various local installation options including **Docker** (03/10/2018); NEW
- Minor bug fixes and feature enhancements based on user feedback (03/02/2018); NEW
- Updated data formats and other documentations (02/15/2018);
- Updated tutorials and FAQs for the three new modules (02/09/2018);
- Release of MetaboAnalyst 4.0 together with a companion R package MetaboAnalystR. You can still access **version 3.0 here** (01/29/2018);

Please Cite:

Xia, J. and Wishart, D.S. (2016) Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis Current Protocols in Bioinformatics, 55:14.10.1-14.10.91.

Xia, J., Sinelnikov, I., Han, B., and Wishart, D.S. (2015) MetaboAnalyst 3.0 - making metabolomics more meaningful . Nucl. Acids Res. 43, W251-257.

Xia, J., Mandal, R., Sinelnikov, I., Broadhurst, D., and Wishart, D.S. (2012) MetaboAnalyst 2.0 - a comprehensive server for metabolomic data analysis . Nucl. Acids Res. 40, W127-133.

Xia, J. and Wishart, D.S. (2011) Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst Nature Protocols 6 (6), 743-760.

Xia, J. and Wishart, D.S. (2011) Metabolomic data processing, analysis, and interpretation using MetaboAnalyst Current Protocols in Bioinformatics, 14.10.1-14.10.48.

Xia, J., Psychogios, N., Young, N. and Wishart, D.S. (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucl. Acids Res. 37, W652-660.

Xia, J., Broadhurst, D., Wilson, M. and Wishart, D. (2013) Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics, 9, 280-299. (*biomarker analysis*)

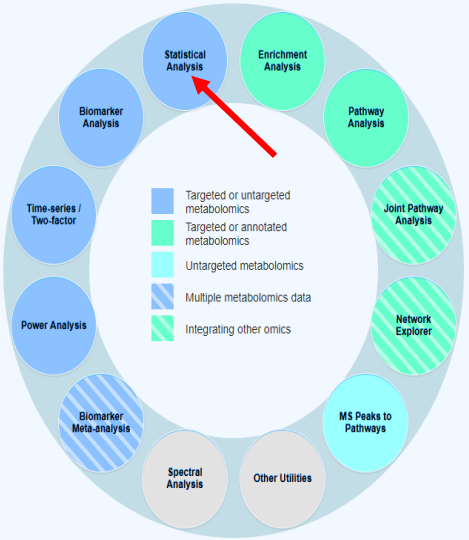
Xia, J., Sinelnikov, I., and Wishart, D.S. (2011) MetATT - a web-based metabolomic tool for analyzing time-series and two-factor data sets. Bioinformatics, 27, 2455-2456 (*time-series and two-factor analysis*)

Xia, J. and Wishart, D.S. (2010) MetPA: a web-based metabolomics tool for pathway analysis and visualization. Bioinformatics, 26, 2342-2344 (*metabolic pathway analysis*)

Xia Lab @ McGill (last updated 2018-04-26)

MetaboAnalyst - statistical, functional and integrative analysis of metabolomics data

Click a module to proceed, or **scroll down** for more details:



1) Upload your data

Tab-delimited text (.txt) or comma-separated values (.csv) file:

Data Type: ○ Concentrations ◉ Spectral bins ○ Peak intensity table

Format: Samples in rows (unpaired)

Data File: Choose File WT_KO_600_Cryo.csv

Submit

Zipped Files (.zip) :

Data Type: ◉ NMR peak list ○ MS peak list ○ MS spectra

Data File: Choose File No file chosen

Pair File: Choose File No file chosen

Submit

2) Try our test data :

	Data Type	Description
○	Concentrations	Metabolite concentrations of 77 urine samples from cancer patients measured by 1H NMR (Eisner R, et al.). Group 1- cachexic; group 2 - control
○	Concentrations	Metabolite concentrations of 39 rumen samples measured by proton NMR from dairy cows fed with different proportions of barley grain (Ametaj BN, et al.). Group label - 0, 15, 30, or 45 - indicating the percentage of grain in diet.
○	NMR spectral bins	Binned 1H NMR spectra of 50 urine samples using 0.04 ppm constant width (Psihogios NG, et al.) Group 1- control; group 2 - severe kidney disease.
○	NMR peak lists	Peak lists and intensity files for 50 urine samples measured by 1H NMR (Psihogios NG, et al.). Group 1- control; group 2 - severe kidney disease.

MetaboAnalyst - statistical, functional and integrative analysis of metabolomics data

Data Integrity Check:

1. Checking the class labels - at least three replicates are required in each class.
2. If the samples are paired, the pair labels must conform to the specified format.
3. The data (except class labels) must not contain non-numeric values.
4. The presence of missing values or features with constant values (i.e. all zeros)

Data processing information:

Checking data content ...passed

Samples are in rows and features in columns

The uploaded file is in comma separated values (.csv) format.

The uploaded data file contains 12 (samples) by 8990 (spectra bins) data matrix.

Samples are not paired.

2 groups were detected in samples.

Only English letters, numbers, underscore, hyphen and forward slash (/) are allowed.

Other special characters or punctuations (if any) will be stripped off.

All data values are numeric.

1113 columns with constant or a single value were found and deleted.

A total of 0 (0%) missing values were detected.

By default, these values will be replaced by a small value.

Click **Skip** button if you accept the default practice

Or click **Missing value imputation** to use other methods

Missing value estimation Skip

MetaboAnalyst - statistical, functional and integrative analysis of metabolomics data

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 is strongly recommended for untargeted metabolomics datasets (i.e. spectral binning data, peak lists) with large number of variables, many of them are from baseline noises. Filtering can usually improve the results. For details, please refer to the paper by Hackstadt, et al

Non-informative variables can be characterized in three groups: 1) variables of **very small values** (close to baseline or detection limit) - these variables can be detected using mean or median; 2) variables that are **near-constant values** throughout the experiment conditions (housekeeping or homeostasis) - these variables can be detected using standard deviation (SD); or the robust estimate such as interquantile range (IQR); and 3) variables that show **low repeatability** - this can be measured using QC samples using the relative standard deviation(RSD = SD/mean). Features with high percent RSD should be removed from the subsequent analysis (the suggested threshold is 20% for LC-MS and 30% for GC-MS). For data filtering based on the first two categories, the following empirical rules are applied during data filtering:

- **Less than 250 variables**: 5% will be filtered;
- **Between 250 - 500 variables**: 10% will be filtered;
- **Between 500 - 1000 variables**: 25% will be filtered;
- **Over 1000 variables**: 40% will be filtered;

Please note, in order to reduce the computational burden to the server, the **None** option is only for less than 4000 features. Over that, if you choose None, the IQR filter will still be applied. In addition, the maximum allowed number of variables is 8000. If over 8000 variables were left after filtering, only the top 8000 will be used in the subsequent analysis.

Filtering features if their RSDs are > 25 % in QC samples

- None (less than 5000 features)
- Interquantile range (IQR)
- Standard deviation (SD)
- Median absolute deviation (MAD)
- Relative standard deviation (RSD = SD/mean)
- Non-parametric relative standard deviation (MAD/median)
- Mean intensity value
- Median intensity value

Submit Proceed

MetaboAnalyst - statistical, functional and integrative analysis of metabolomics data

Normalization overview:

The normalization procedures are grouped into three categories. The sample normalization allows general-purpose adjustment for differences among your sample; data transformation and scaling are two different approaches to make individual features more comparable. You can use one or combine them to achieve better results.

Sample normalization

- None
- **Sample-specific normalization (i.e. weight, volume)** Click here to specify
- **Normalization by sum**
- **Normalization by median**
- **Normalization by reference sample (PQN)** 1_4
- **Normalization by a pooled sample from group** KO_PM
- **Normalization by reference feature** 0.511
- **Quantile normalization**

Data transformation

- None
- **Log transformation** (generalized logarithm transformation or glog)
- **Cube root transformation** (take cube root of data values)

Data scaling

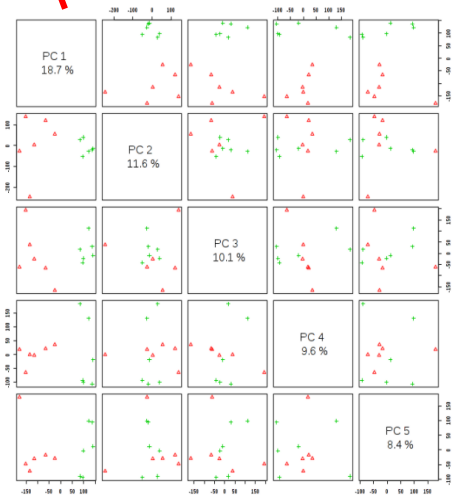
- None
- **Mean centering** (mean-centered only)
- **Auto scaling** (mean-centered and divided by the standard deviation of each variable)
- **Pareto scaling** (mean-centered and divided by the square root of standard deviation of each variable)
- **Range scaling** (mean-centered and divided by the range of each variable)

Normalize View Result Proceed

MetaboAnalyst - statistical, functional and integrative analysis of metabolomics data

Overview | Scree Plot | 2D Scores Plot | 3D Scores Plot | Loadings Plot | Biplot

Display pairwise score plot for top 5 PCs Update

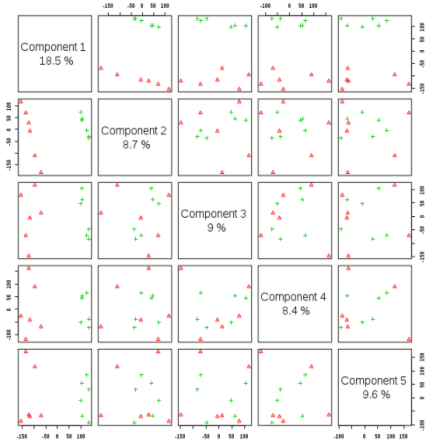


Upload
Processing
Normalization
Statistics
Fold change
T-test
Volcano plot
ANOVA
Correlations
PatternHunter
PCA
PLSDA
sPLSDA
OrthoPLSDA
SAM
EBAM
Dendrogram
Heatmap
SOM
K-means
RandomForest
SVM
Download
Exit

MetaboAnalyst - statistical, functional and integrative analysis of metabolomics data

Overview | 2D Scores Plot | 3D Scores Plot | Loadings Plot | Cross Validation | Imp. Features | Permutation

Display pairwise score plot for top 5 components Update



Upload
Processing
Normalization
Statistics
Fold change
T-test
Volcano plot
ANOVA
Correlations
PatternHunter
PCA
PLSDA
sPLSDA
OrthoPLSDA
SAM
EBAM
Dendrogram
Heatmap
SOM
K-means
RandomForest
SVM
Download
Exit

MetaboAnalyst - statistical, functional and integrative analysis of metabolomics data

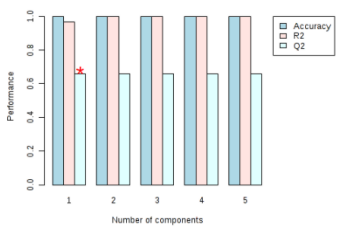
Overview | 2D Scores Plot | 3D Scores Plot | Loadings Plot | Cross Validation | Imp. Features | Permutation

Select optimal number of components for classification

Maximum components to search: 5

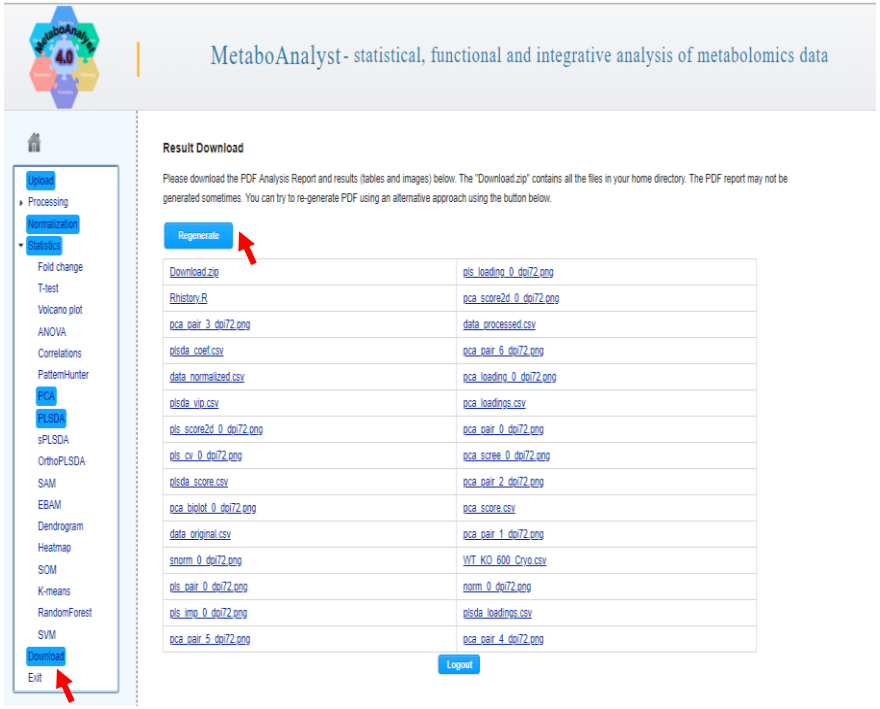
Cross validation (CV) method: LOOCV Update

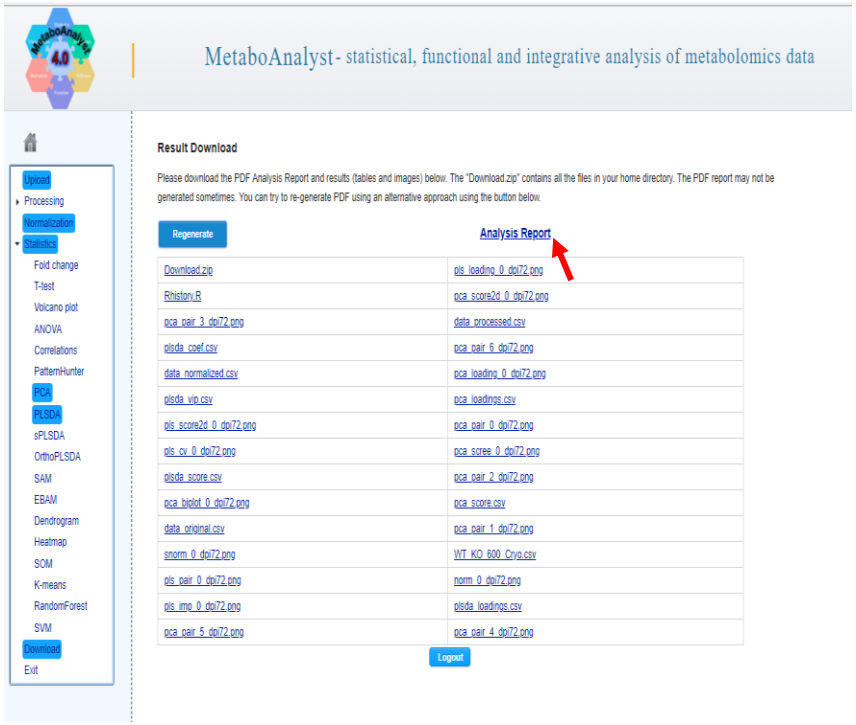
Performance measure: Q2



Q2 is an estimate of the predictive ability of the model, and is calculated via cross-validation (CV). In each CV, the predicted data are compared with the original data, and the sum of squared errors is calculated. The prediction error is then summed over all samples (Predicted Residual Sum of Squares or PRESS). For convenience, the PRESS is divided by the initial sum of squares and subtracted from 1 to resemble the scale of the R2. Good predictions will have low PRESS or high Q2. It is possible to have **negative Q2**, which means that your model is not at all predictive or is overfitted. For more details, refer to an excellent paper by (Szymańska, et al).

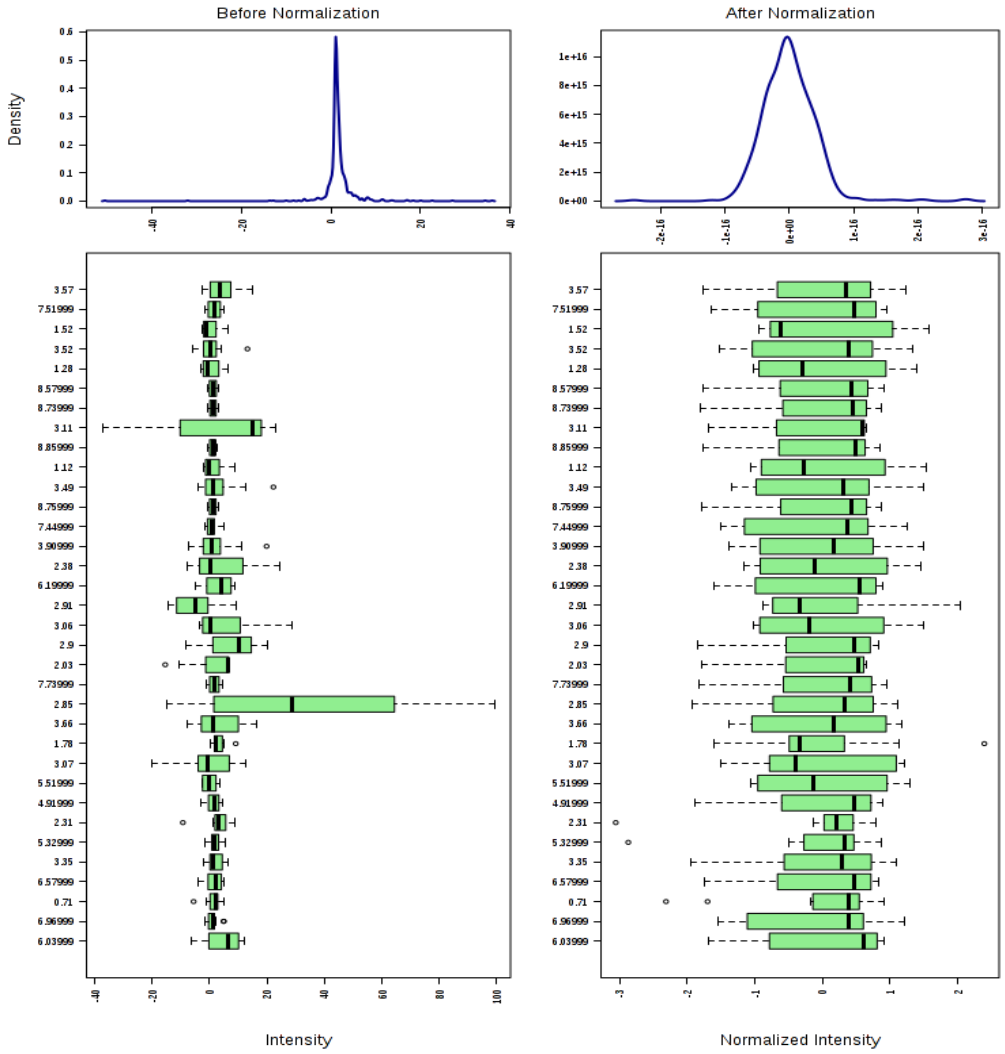
Upload
Processing
Normalization
Statistics
Fold change
T-test
Volcano plot
ANOVA
Correlations
PatternHunter
PCA
PLSDA
sPLSDA
OrthoPLSDA
SAM
EBAM
Dendrogram
Heatmap
SOM
K-means
RandomForest
SVM
Download
Exit


MetaboAnalyst - statistical, functional and integrative analysis of metabolomics data
Upload
Processing
Normalization
Statistics
Fold change
T-test
Volcano plot
ANOVA
Correlations
PatternHunter
PCA
PLSDA
sPLSDA
OrthoPLSDA
SAM
EBAM
Dendrogram
Heatmap
SOM
K-means
RandomForest
SVM
Download
Exit
Result Download
Please download the PDF Analysis Report and results (tables and images) below. The "Download.zip" contains all the files in your home directory. The PDF report may not be generated sometimes. You can try to re-generate PDF using an alternative approach using the button below.
Regenerate
Download.zip
Rhistory.R
pca_pair_3_dpi72.png
plsda_coef.csv
data_normalized.csv
plsda_vip.csv
pls_score2d_0_dpi72.png
pls_cv_0_dpi72.png
plsda_score.csv
pca_biplot_0_dpi72.png
data_original.csv
snorm_0_dpi72.png
pls_pair_0_dpi72.png
pls_imp_0_dpi72.png
pca_pair_5_dpi72.png
pls_loading_0_dpi72.png
pca_score2d_0_dpi72.png
data_processed.csv
pca_pair_6_dpi72.png
pca_loading_0_dpi72.png
pca_loadings.csv
pca_pair_0_dpi72.png
pca_scree_0_dpi72.png
pca_pair_2_dpi72.png
pca_score.csv
pca_pair_1_dpi72.png
WT_KO_600_Cryo.csv
norm_0_dpi72.png
plsda_loadings.csv
pca_pair_4_dpi72.png
Logout

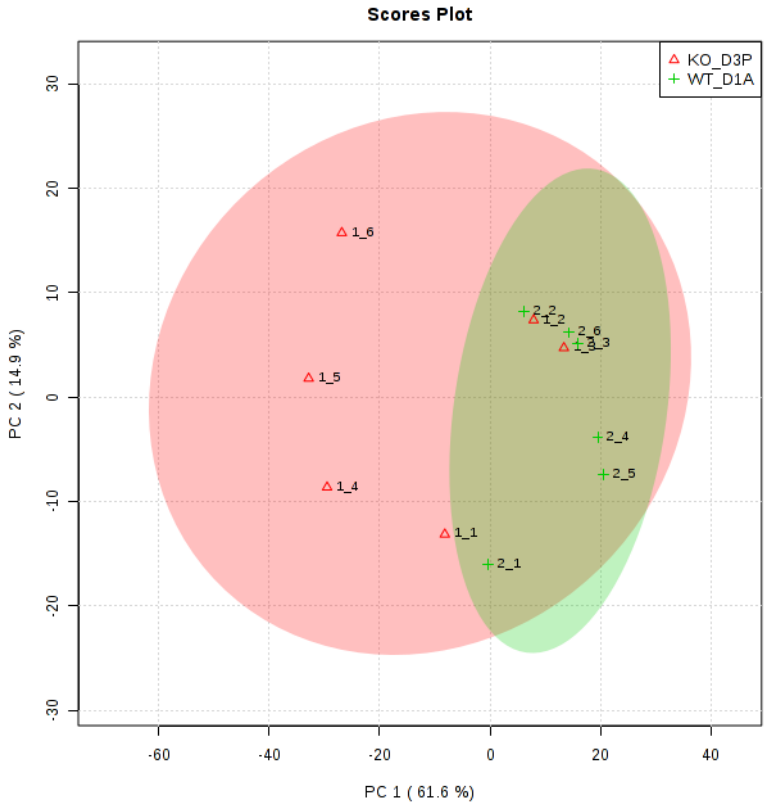

MetaboAnalyst - statistical, functional and integrative analysis of metabolomics data
Upload
Processing
Normalization
Statistics
Fold change
T-test
Volcano plot
ANOVA
Correlations
PatternHunter
PCA
PLSDA
sPLSDA
OrthoPLSDA
SAM
EBAM
Dendrogram
Heatmap
SOM
K-means
RandomForest
SVM
Download
Exit
Result Download
Please download the PDF Analysis Report and results (tables and images) below. The "Download.zip" contains all the files in your home directory. The PDF report may not be generated sometimes. You can try to re-generate PDF using an alternative approach using the button below.
Regenerate
Analysis Report
Download.zip
Rhistory.R
pca_pair_3_dpi72.png
plsda_coef.csv
data_normalized.csv
plsda_vip.csv
pls_score2d_0_dpi72.png
pls_cv_0_dpi72.png
plsda_score.csv
pca_biplot_0_dpi72.png
data_original.csv
snorm_0_dpi72.png
pls_pair_0_dpi72.png
pls_imp_0_dpi72.png
pca_pair_5_dpi72.png
pls_loading_0_dpi72.png
pca_score2d_0_dpi72.png
data_processed.csv
pca_pair_6_dpi72.png
pca_loading_0_dpi72.png
pca_loadings.csv
pca_pair_0_dpi72.png
pca_scree_0_dpi72.png
pca_pair_2_dpi72.png
pca_score.csv
pca_pair_1_dpi72.png
WT_KO_600_Cryo.csv
norm_0_dpi72.png
plsda_loadings.csv
pca_pair_4_dpi72.png
Logout

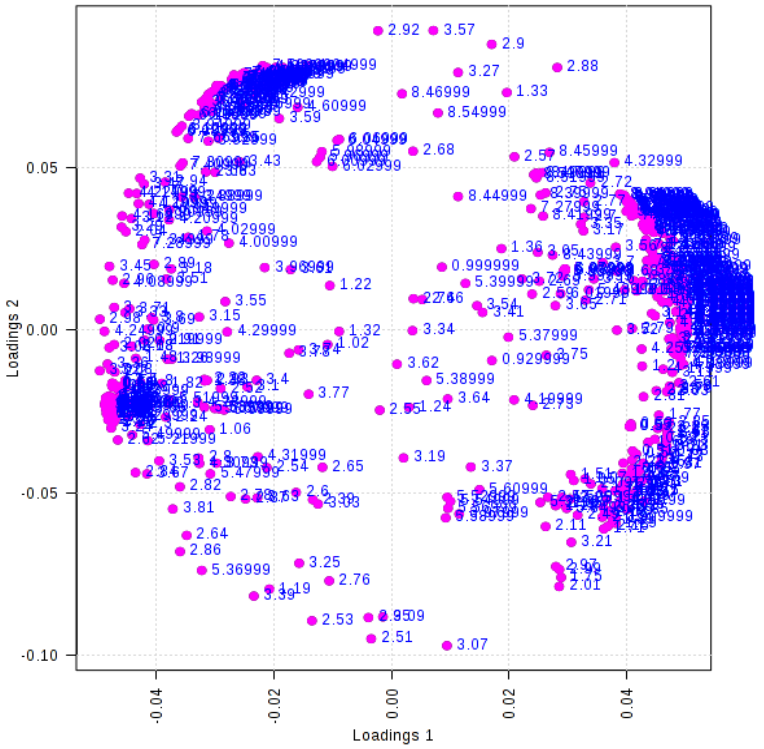
Statistical Data Analysis of Urine NMR-data recorded

On 600 MHz

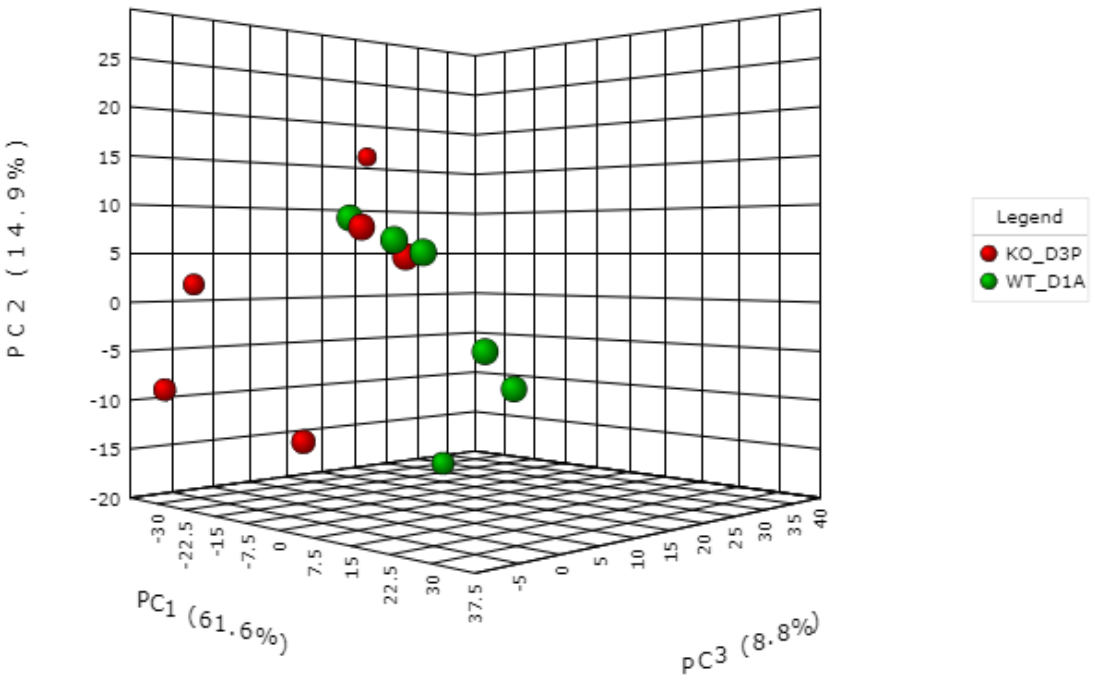


Box plots and kernel density plots before and after normalization. The density plots are based on all samples.

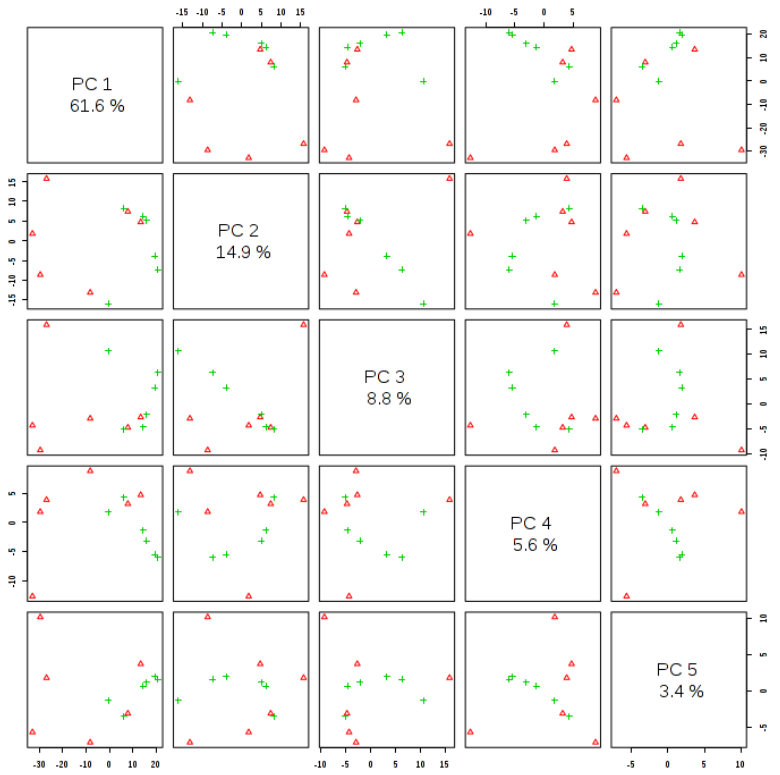


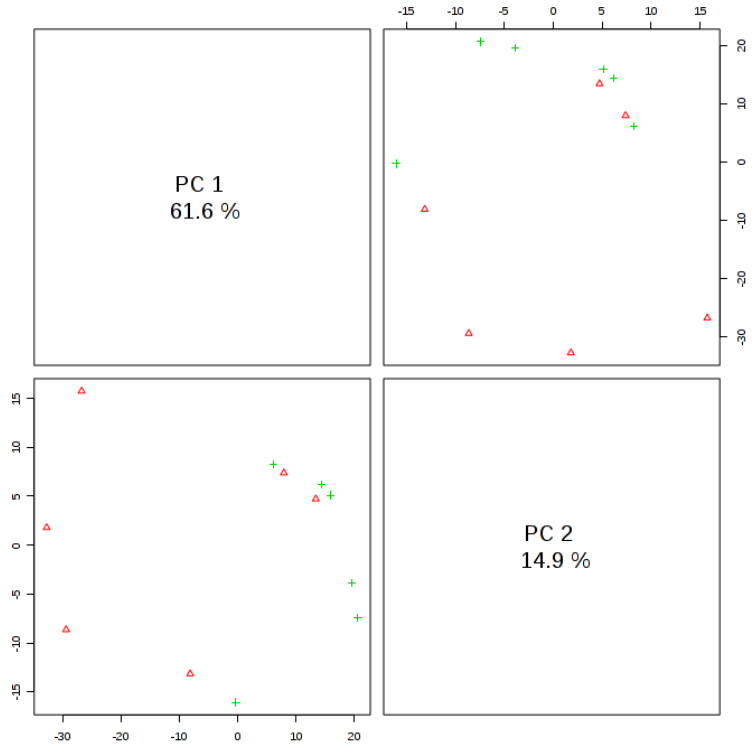


Scores plot between the selected PCs. The explained variances are shown in brackets and corresponding Loadings plot for the selected PCs..

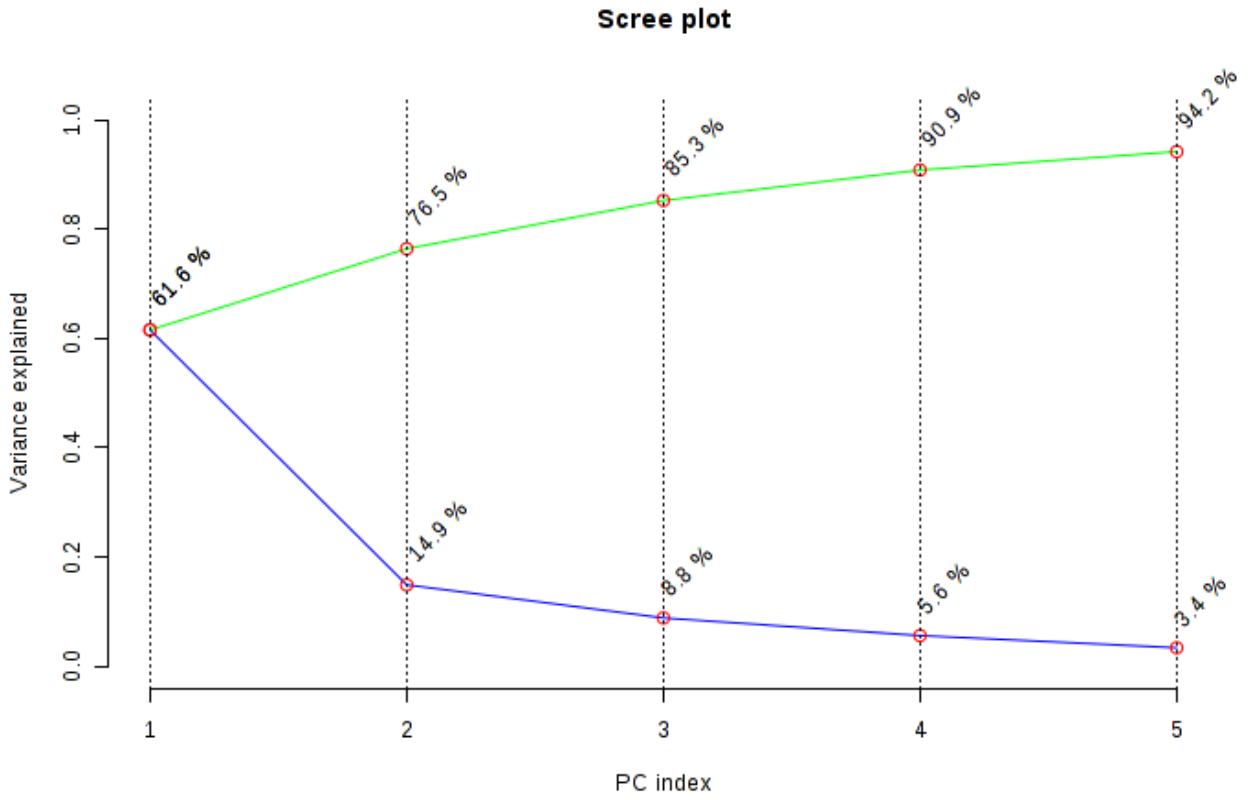


3D score plot between the selected PCs. The explained variances are shown in brackets

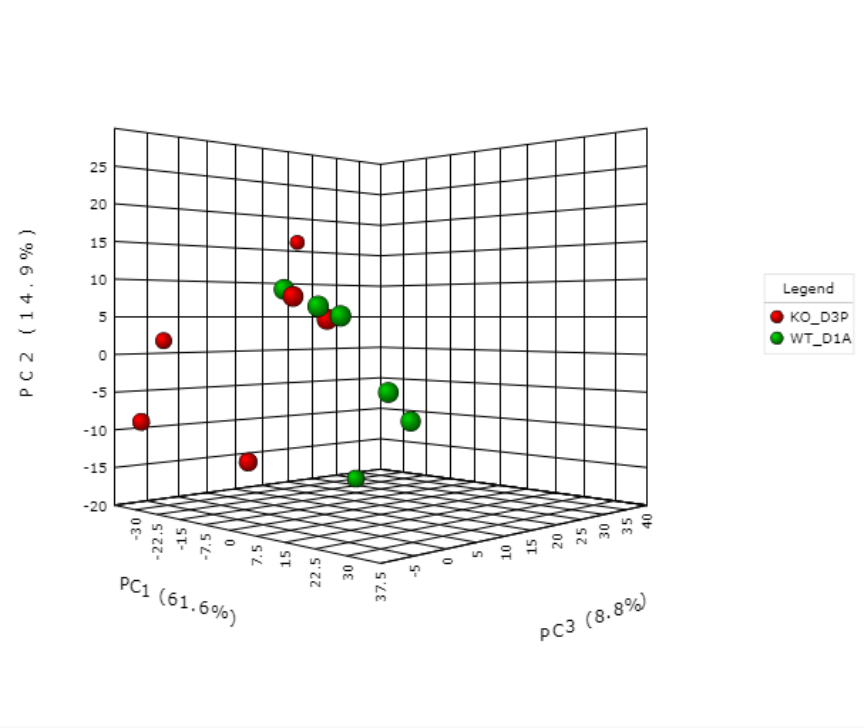


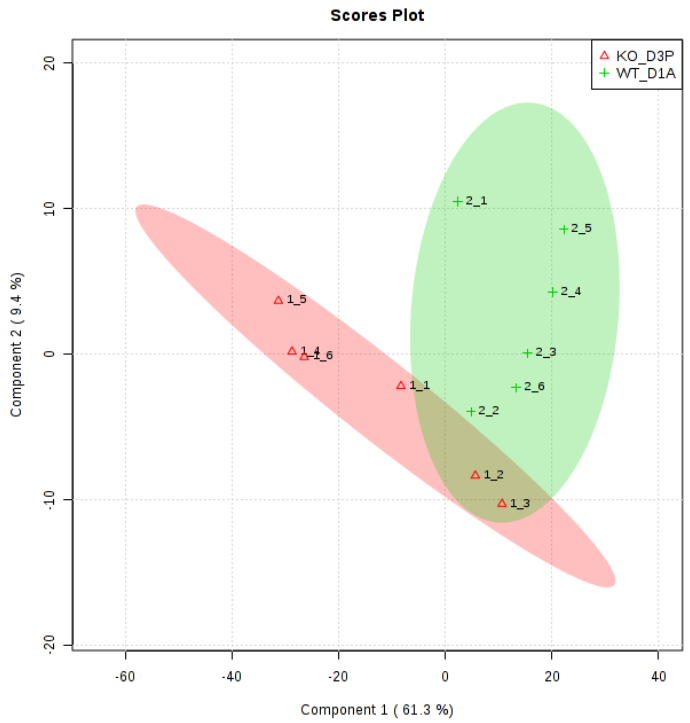


Pairwise score plots between the selected PCs. The explained variance of each PC is shown in the corresponding diagonal cell.

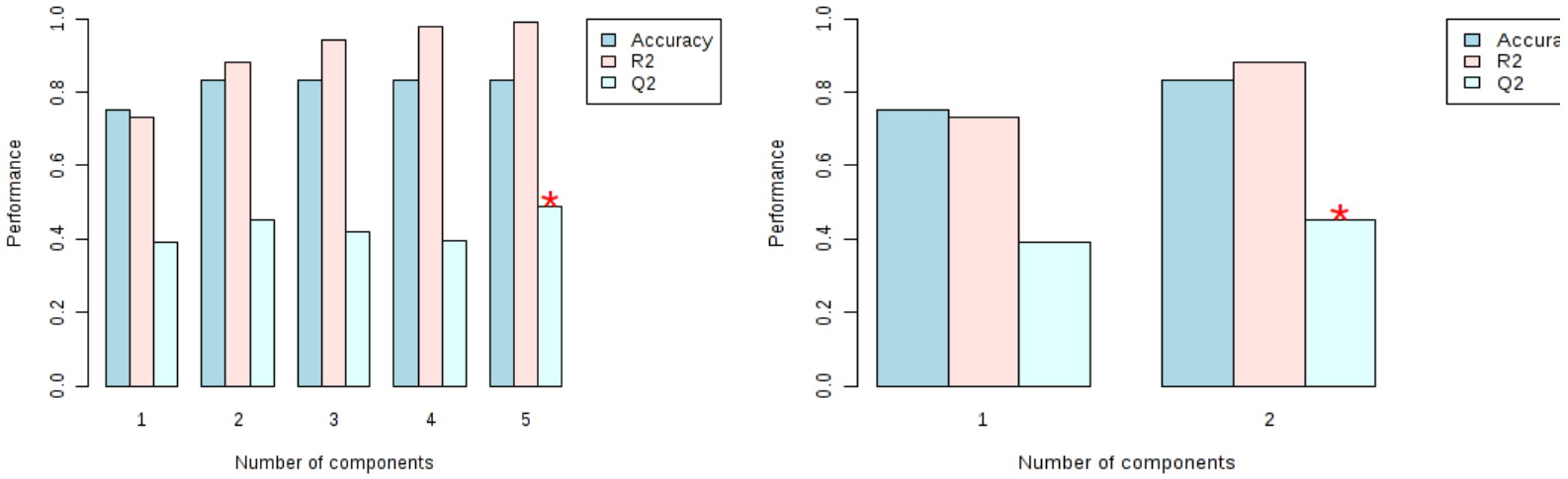


The green line on top shows the accumulated variance explained; the blue line underneath shows the variance explained by individual PC





PLS-DA 3D and 2D Scores plot and corresponding explained variance in the bracket

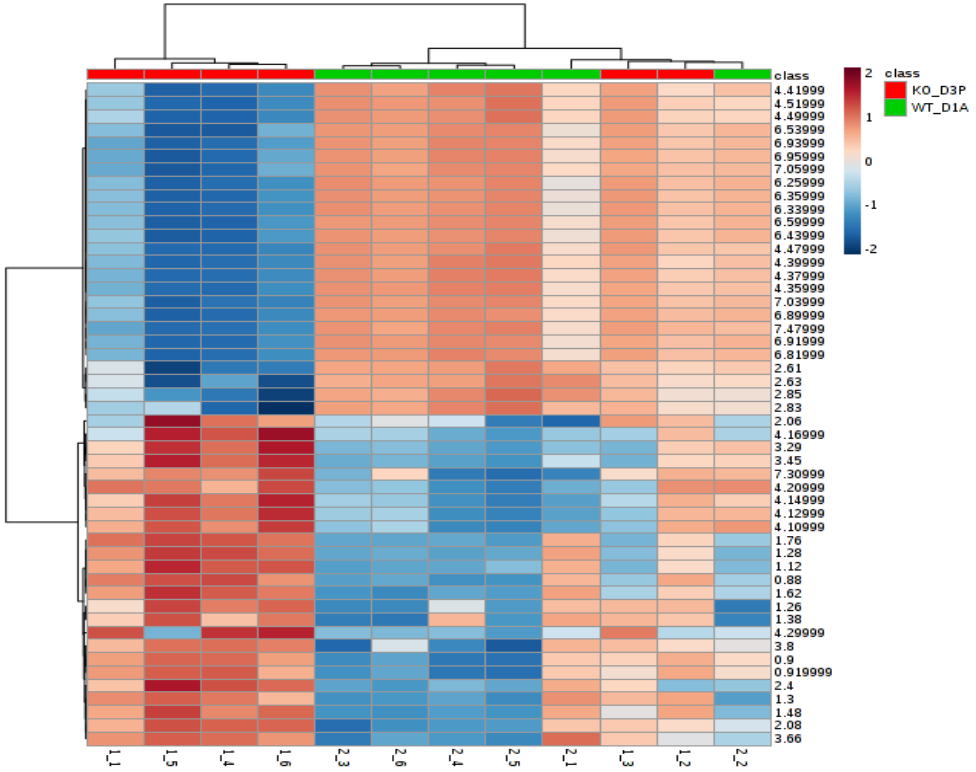


PLS-DA classification using different number of components. The red star indicates the best classier.

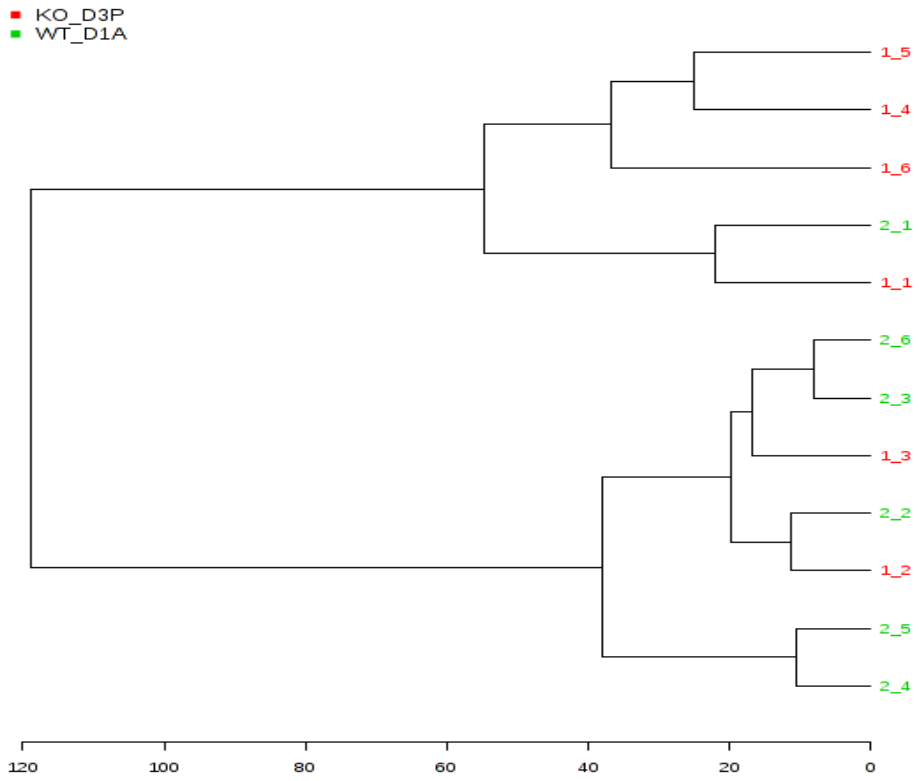
Hierarchical Clustering: Hierarchical clustering is commonly used for unsupervised clustering.

In hierarchical cluster analysis, each sample begins as a separate cluster and the algorithm proceeds to combine them until all samples belong to one cluster.

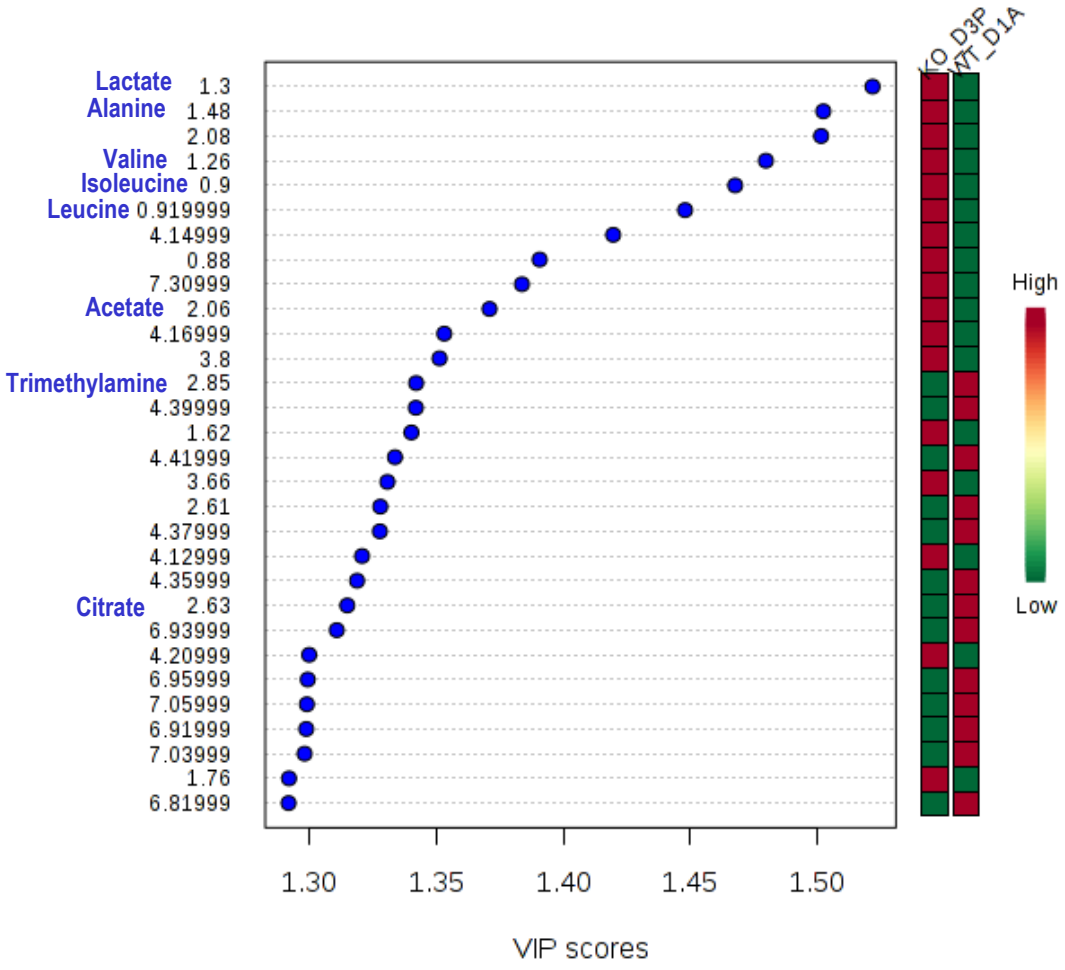
1. The first parameters is similarity measure - Euclidean distance, Pearson's correlation, Spearman's rank correlation.
2. The second parameter is clustering algorithms, including average linkage.



Clustering result shown as heatmap



Clustering result shown as dendrogram

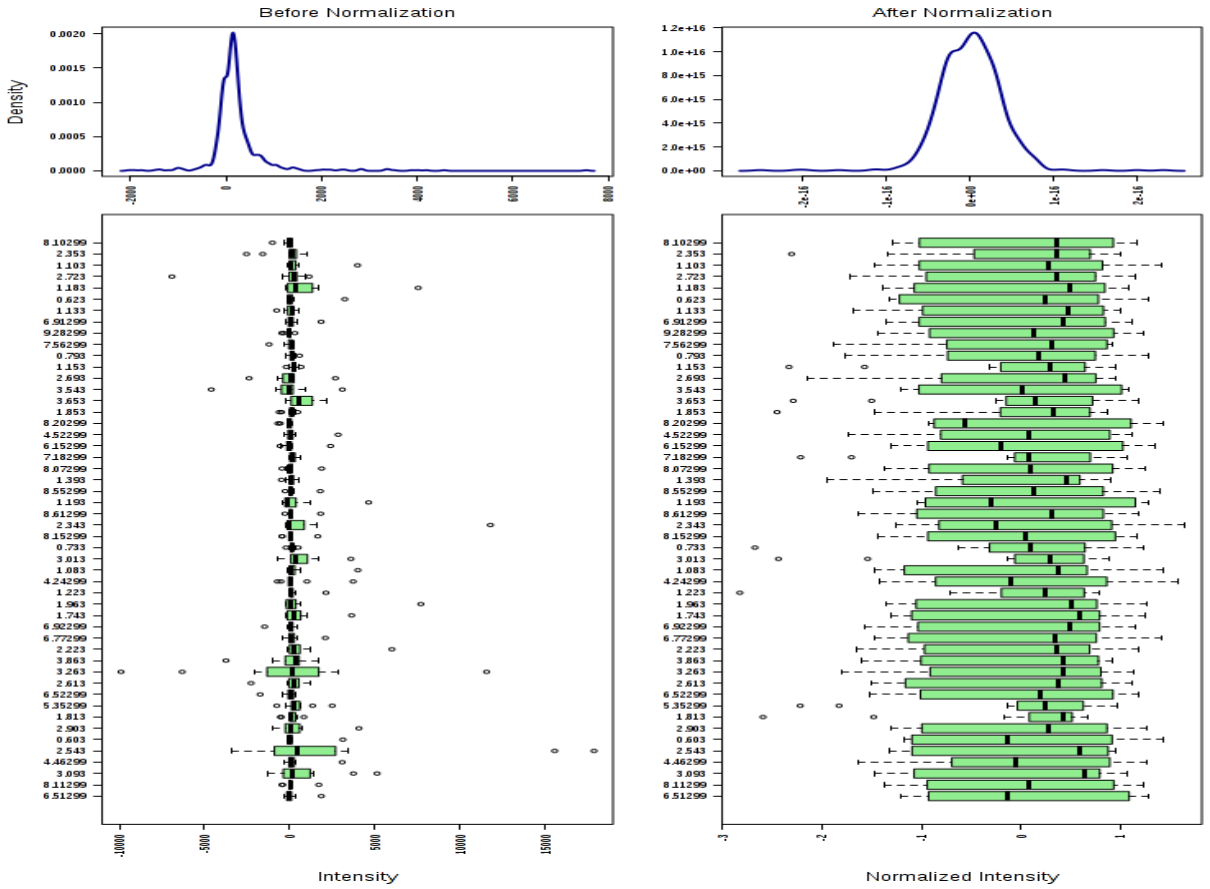


Important features identified by PLS-DA. The colored boxes on the right indicate the relative concentrations of the corresponding metabolite in each group under study.

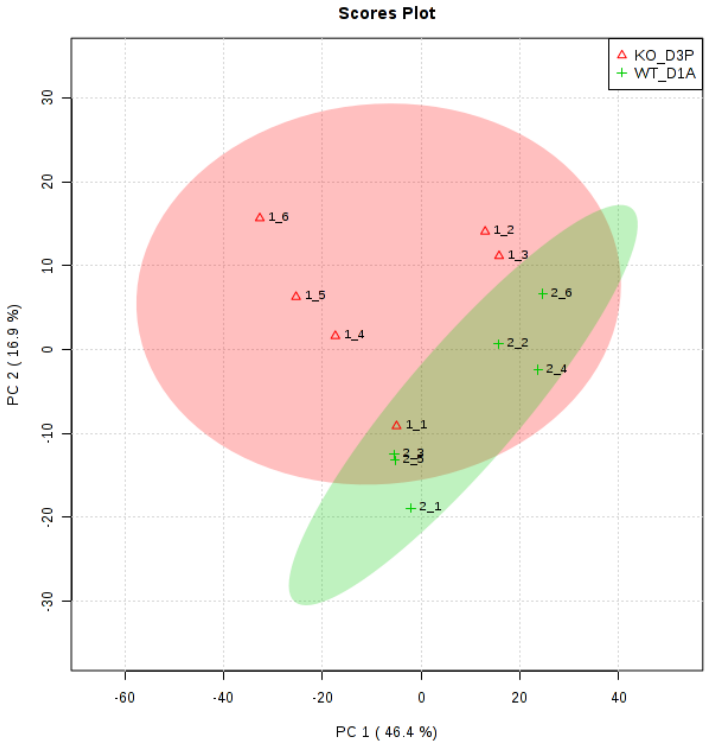
The Variable Importance in Projection (VIP) is a weighted sum of squares of the PLS loadings taking into account the amount of explained Y-variation in each dimension.

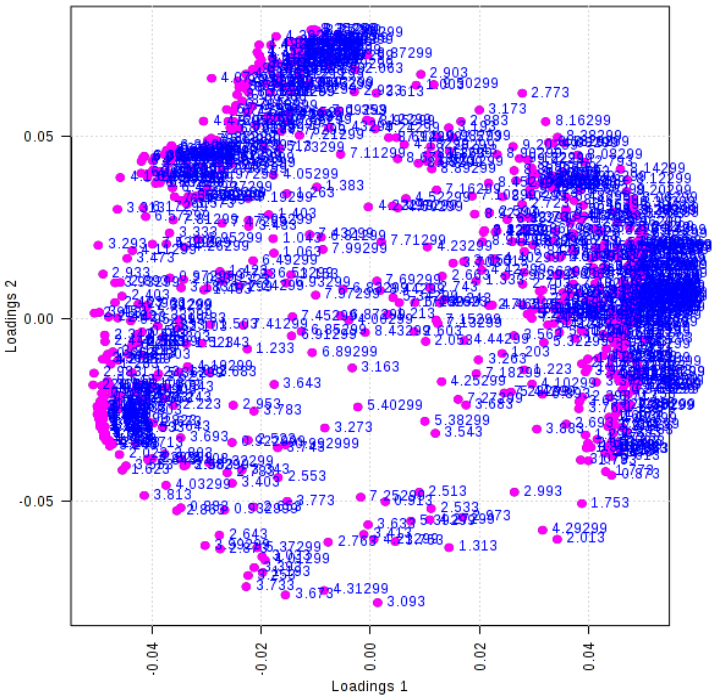
Statistical Data Analysis of Urine NMR-data recorded

On 500 MHz

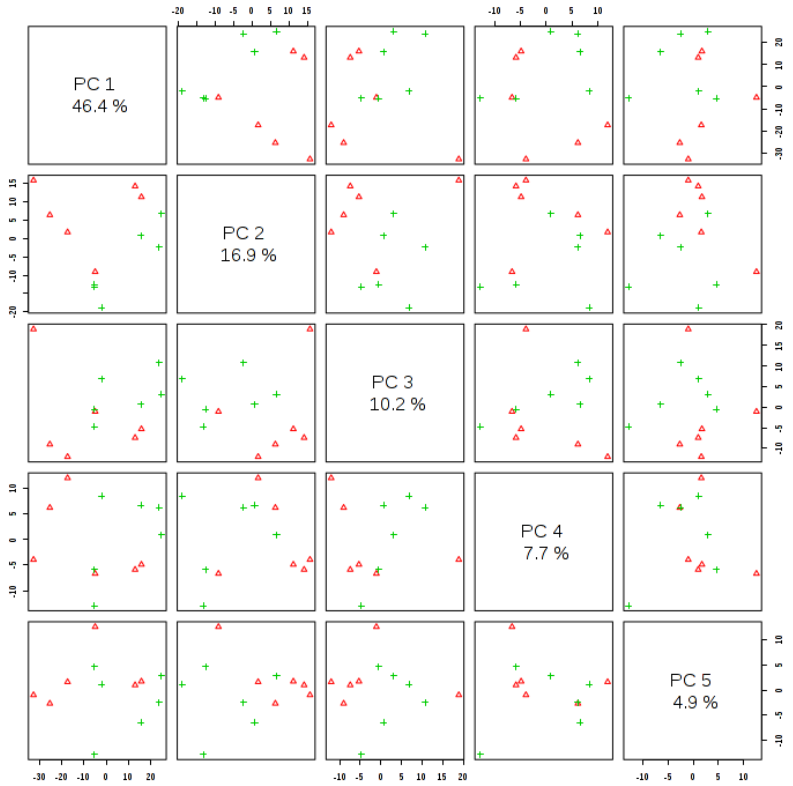


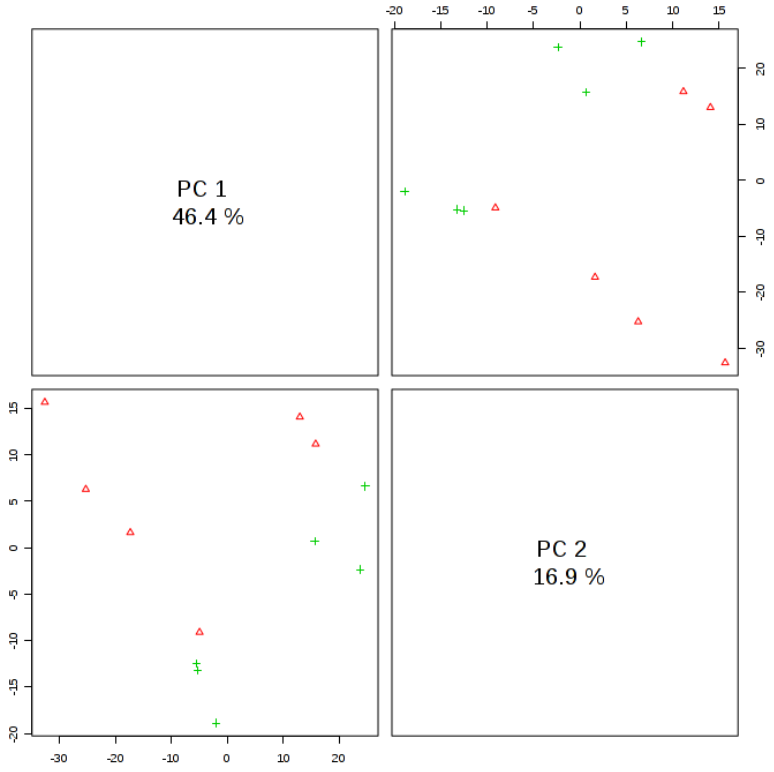
Box plots and kernel density plots before and after normalization. The density plots are based on all samples.



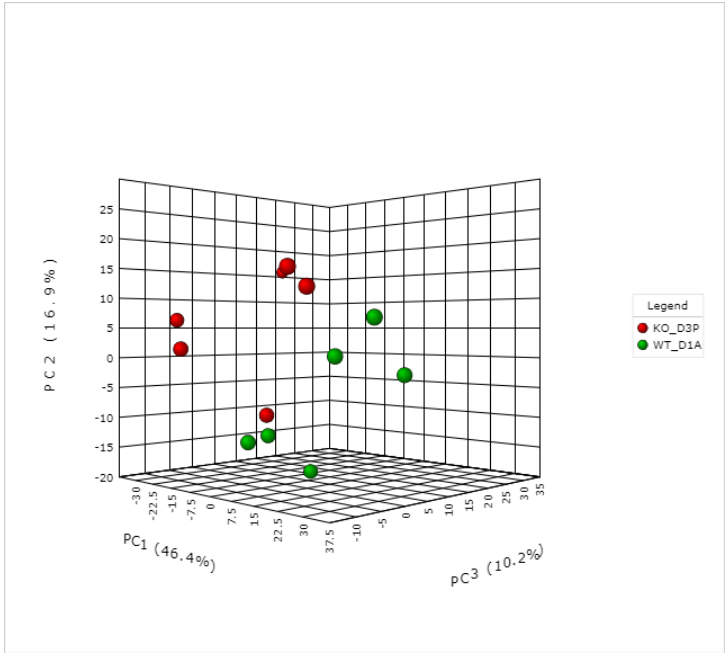


Scores plot between the selected PCs (The explained variances are shown in brackets) and corresponding Loadings plot for the selected PCs.

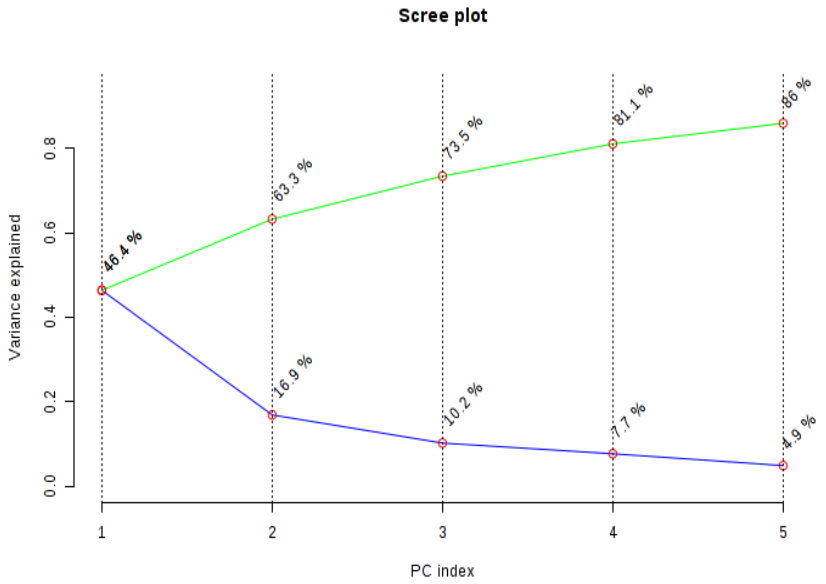




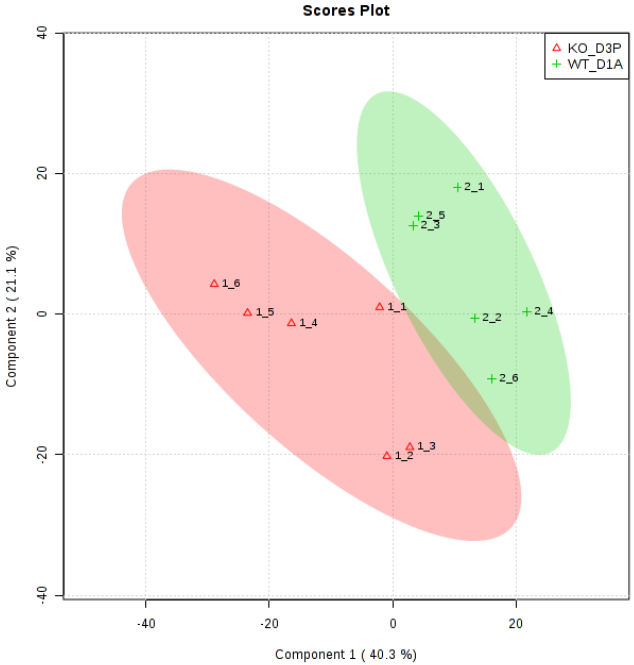
Pairwise score plots between the selected PCs. The explained variance of each PC is shown in the corresponding diagonal cell.

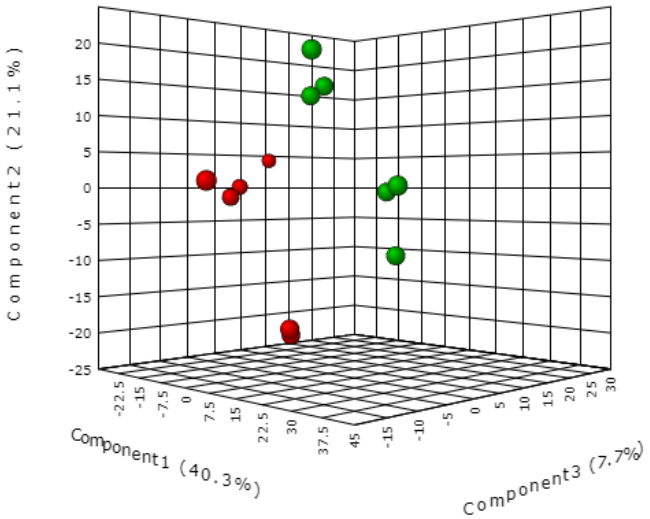


3D scores plot between the selected PCs. The explained variances are shown in brackets.

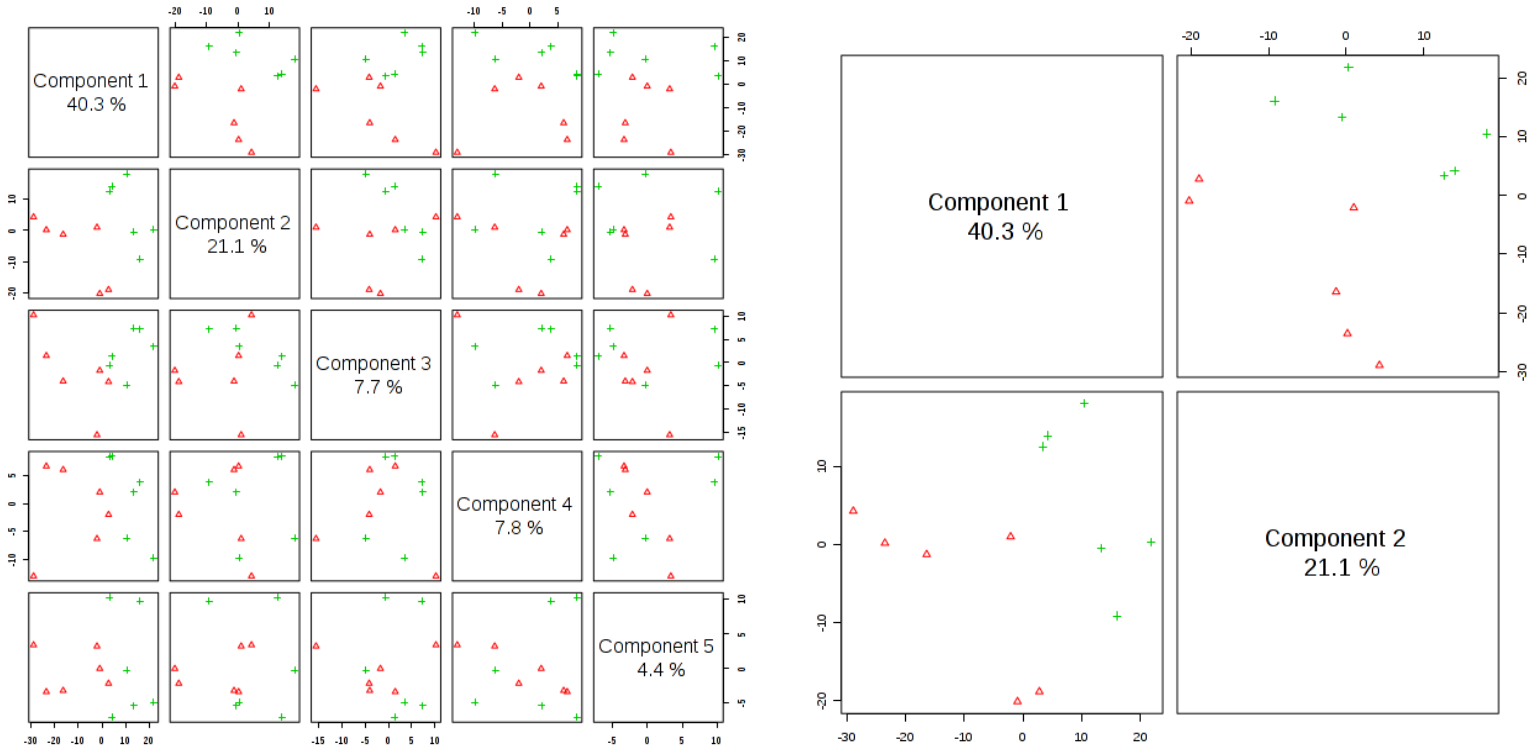


Scree plot shows the variance explained by PCs. (accumulated variance explained along with the variance explained by individual PC.

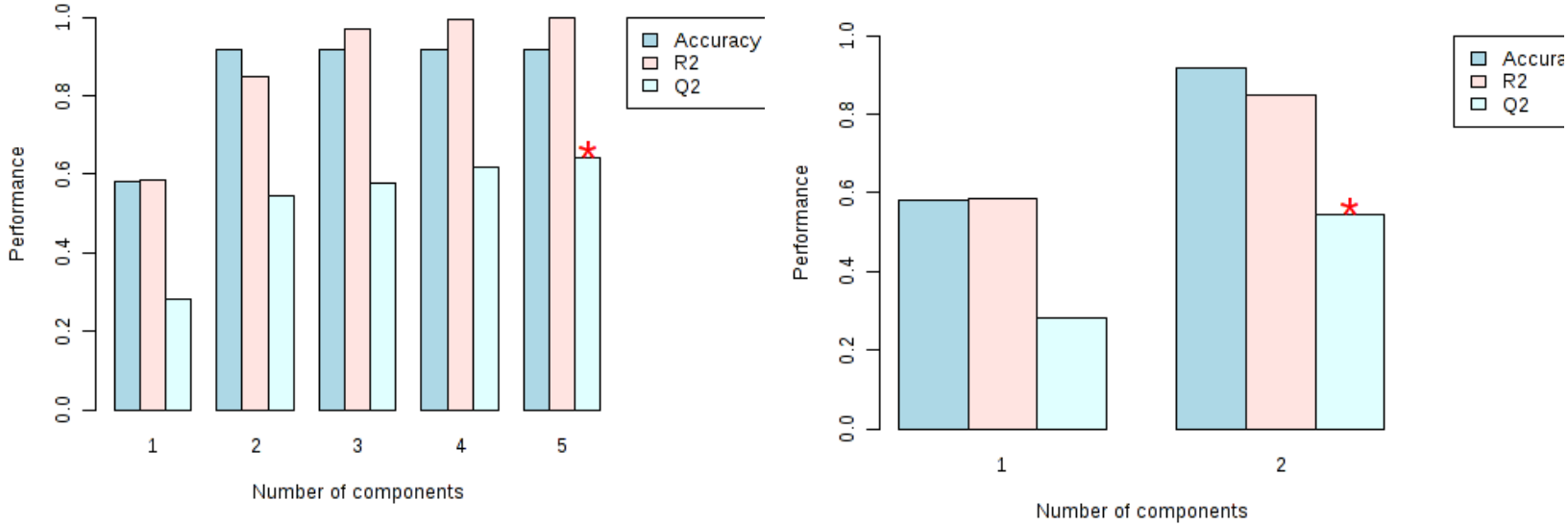




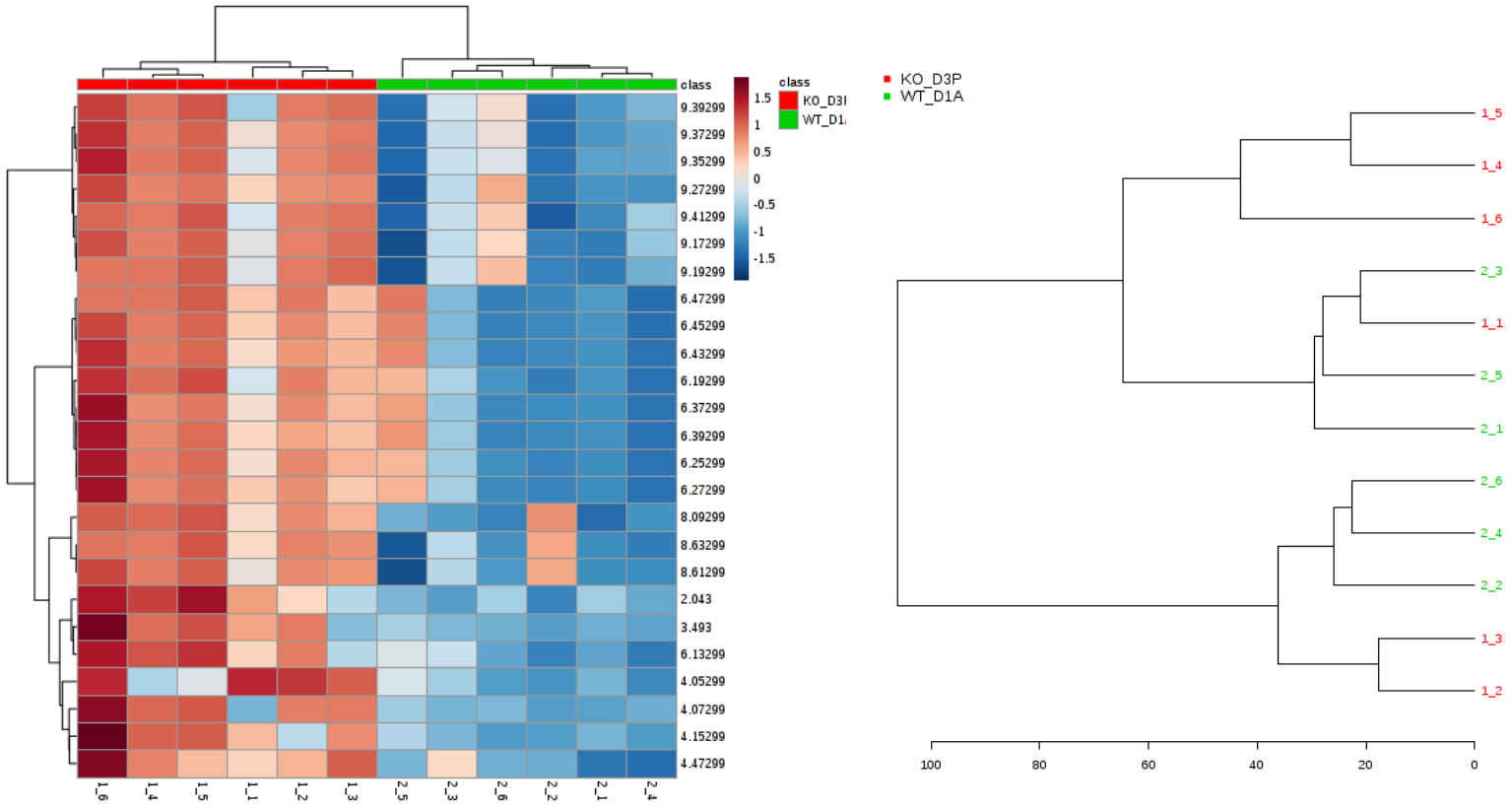
PLS-DA 2D and 3D scores plot between the selected PCs.



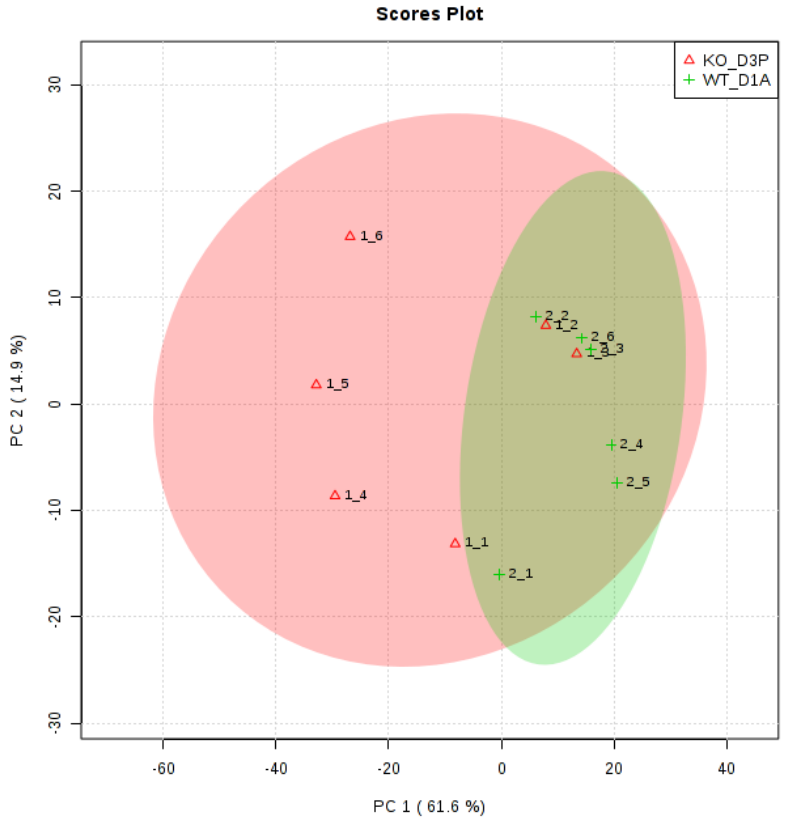
PLS-DA pairwise score plots between the selected PCs. The explained variance of each PC is shown in the corresponding diagonal cell.



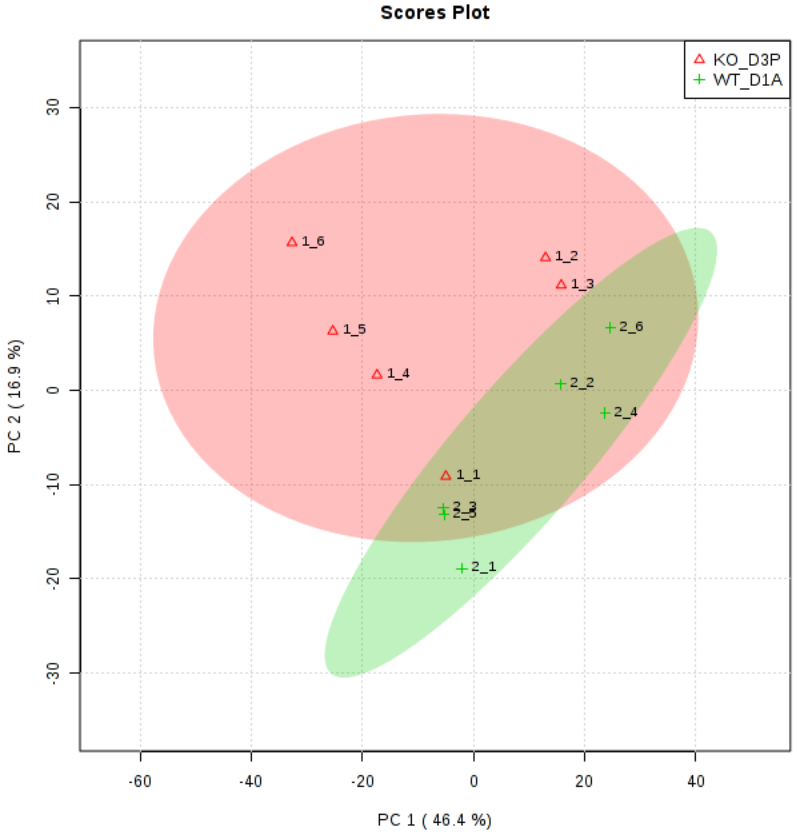
PLS-DA classification using different number of components. The red star indicates the best classier.



Clustering result shown as heatmap and dendrogram



PCA of 600 MHz data



PCA of 500 MHz data

Thank you

Workshop

Upload your expt. no 256 for all samples from 600 MHz (6 KO and 6 WT) to MestReNova in your NMRBox

1. Cleaning data (zf to same values, lb to same values, bc, alignment, remove water and urea) normalize spectrum to DSS.
2. Bin 600 data to 0.005, 0.01, 0.02, and 0.04 ppm buckets
3. Do full metabonalyst pipeline to look for differences between WT and KO
4. (How does binning effect our results)
5. Do 500 MHz data (Expt. No. 256) in same way
6. Add 500 and 600 MHz data with chosen binning dimensions
7. Run Metaboanalyst to see if the 500 vs 600 causes separation between groups