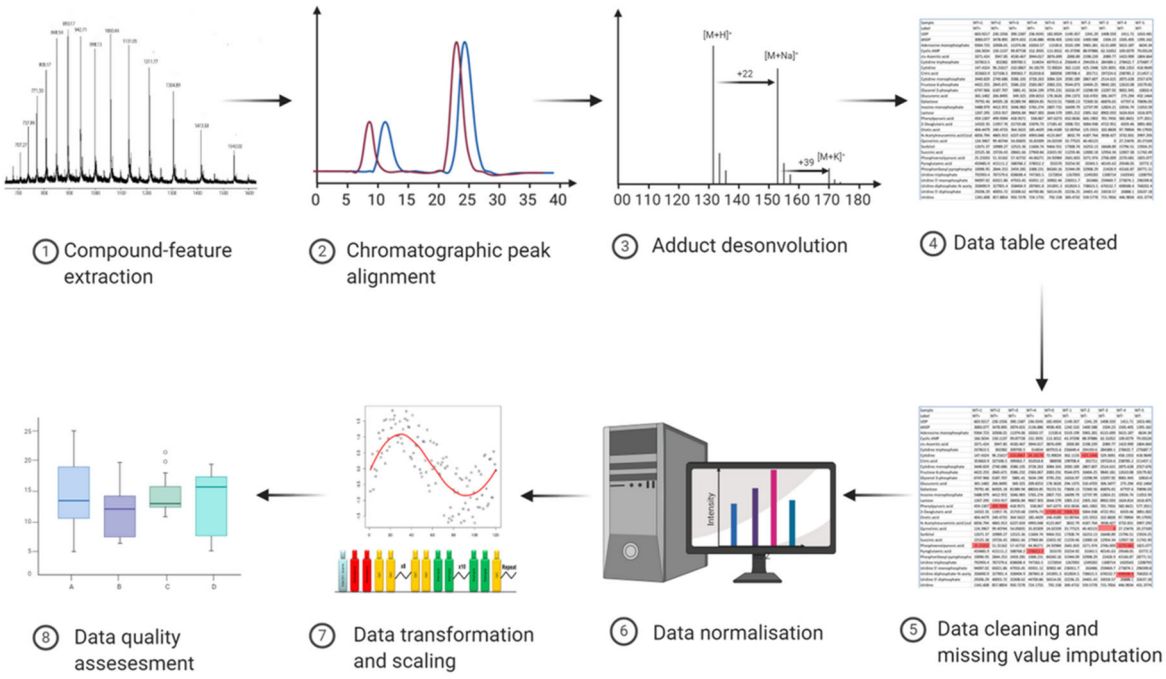
Lecture 1 Handout



Metabolomics Data Processing and Statistical Analysis

OVERVIEW

Part 1: Introduction to processing and analysis of metabolomics data

- The purpose of bioinformatics in metabolomics
- Bioinformatics workflow
- Pre-processing data (raw files to data tables)
- Final data processing data (data formatting for statistical analysis)
- Statistical tools: univariate
- Statistical tools: multivariate

Part 2: MetaboAnalyst – data processing and statistical analysis

- Introduction to MetaboAnalyst software
- Data processing (preparing data for statistical analysis)
- Statistical analysis: fold-change, significance, supervised and unsupervised multivariate analysis, correlations and classification.
- Data reporting

DEFINING METABOLOMICS

Metabolomics as an analytical approach which aims to ***comprehensively characterise metabolites in a biological system in order to investigate altered metabolic states***.

Targeted and untargeted approaches can be used which provide **hypothesis generating** and **hypothesis-led** experiments leading to complementary information that contributes to a systems-level understanding of the biological context and subsequently a molecular phenotype.

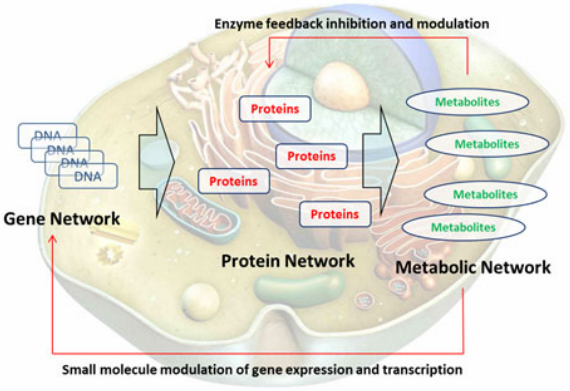
A number of **terms have evolved** that characterise these different elements and approaches within the metabolomics framework.

Metabolome: This refers to the entire set of metabolites found in a cell, tissue or organism.

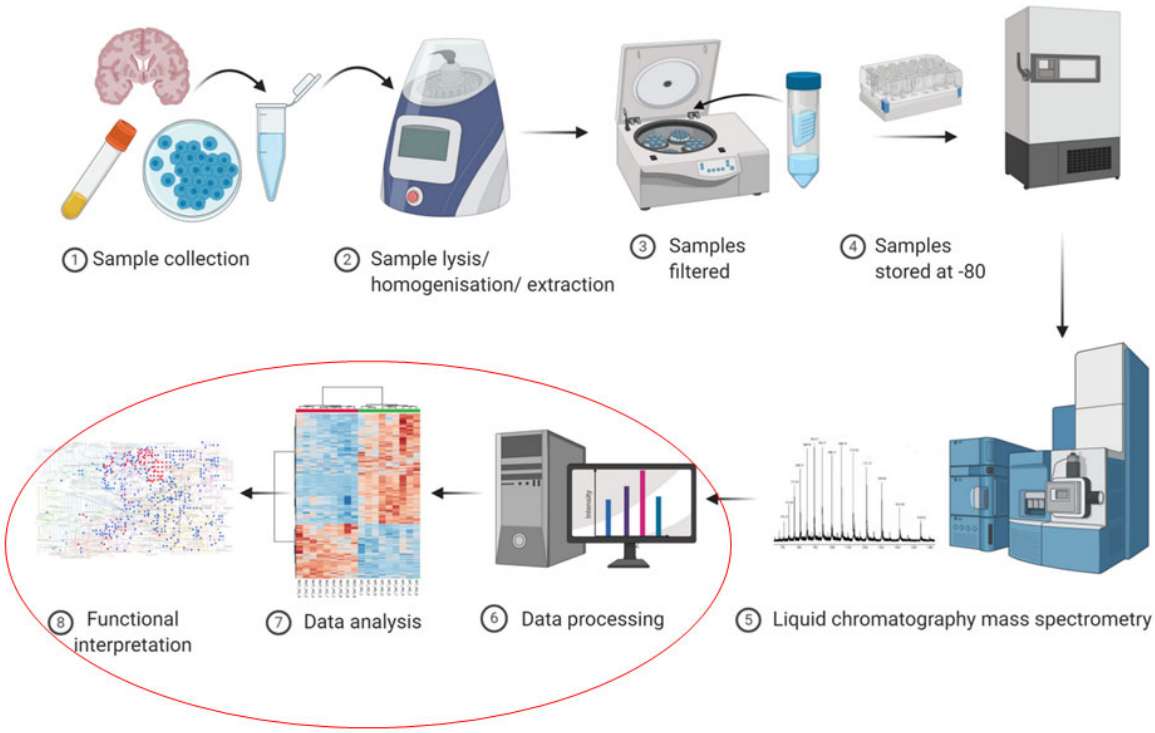
Metabolic profiling: The analysis of a selected number of pre-defined metabolites in a biochemical system.

Metabolic fingerprinting: The metabolites and their concentrations as a snapshot in time representing the state of a biological system or organism.

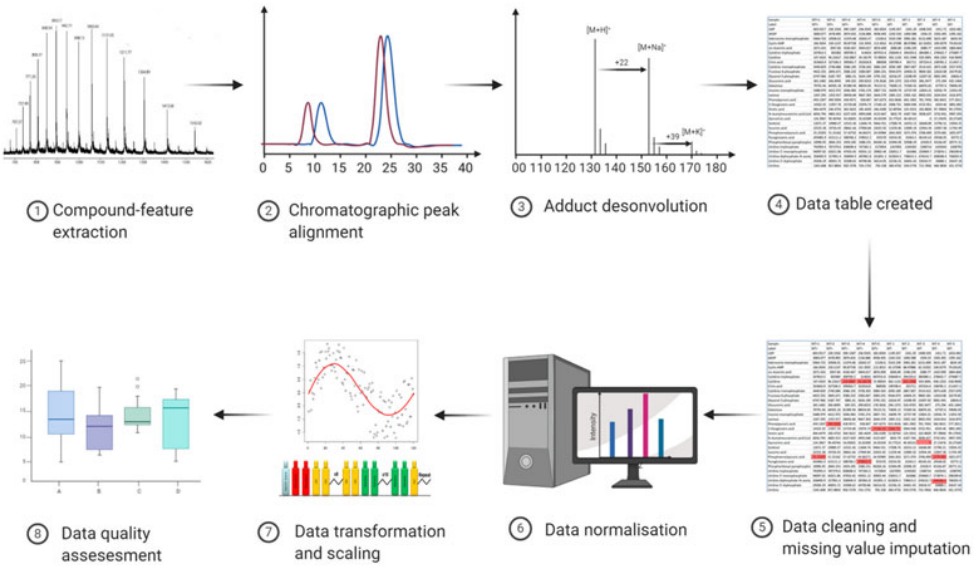
Metabolic footprinting: This refers to measuring what cells or a biological system excretes under controlled conditions, also known as the exometabolome.



OVERVIEW OF THE METABOLOMICS WORKFLOW

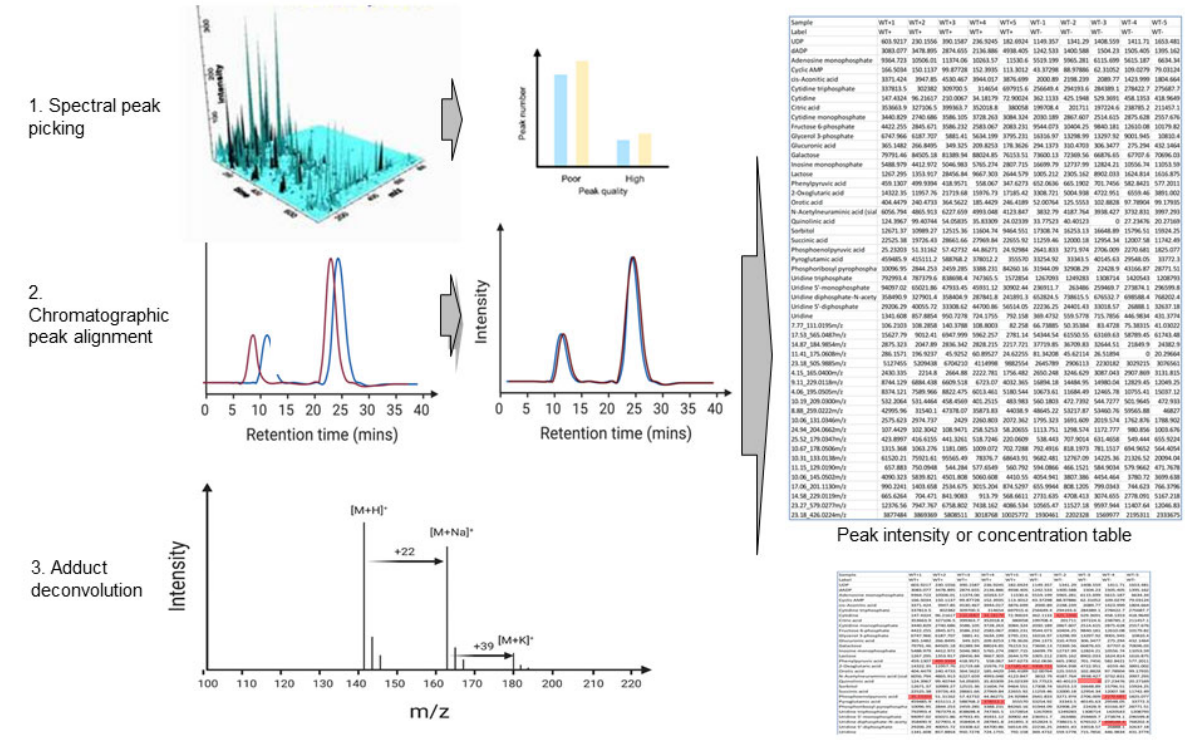


OVERVIEW OF DATA PROCESSING



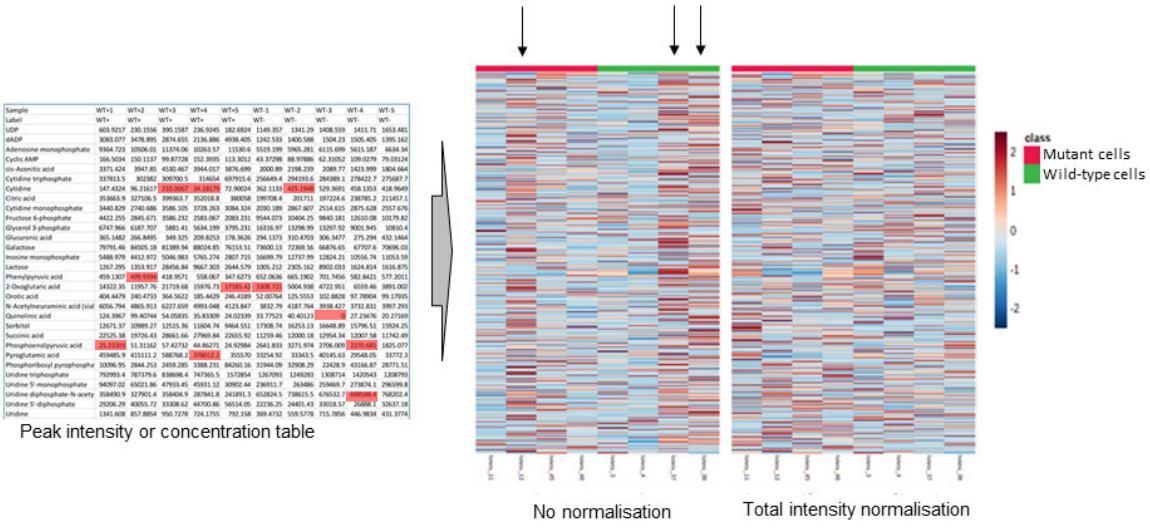
The **purpose of data processing** is to shape and format platform-specific raw data into a data table or matrix appropriate for statistical analysis. The process is split into two parts.

1. DATA (PRE) PROCESSING



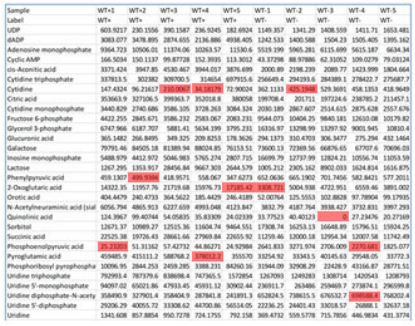
2. DATA PROCESSING: NORMALISATION

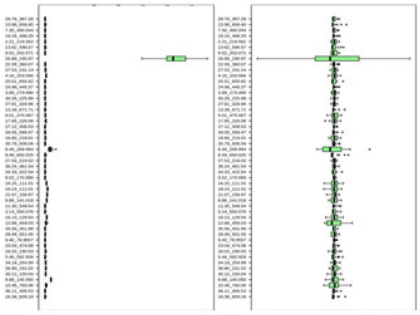
Normalisation of metabolomics data aims to **remove systematic variation between experimental conditions.** Data-driven normalisation is the final normalisation step before statistical analysis and adjusts each same by a factor. **Total intensity normalisation** for example is based on using the total ion abundance for each sample as the normalisation factor.



2. DATA PROCESSING: TRANSFORMATION AND SCALING

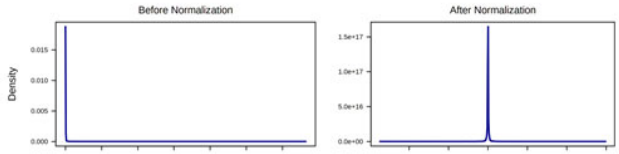
Data table





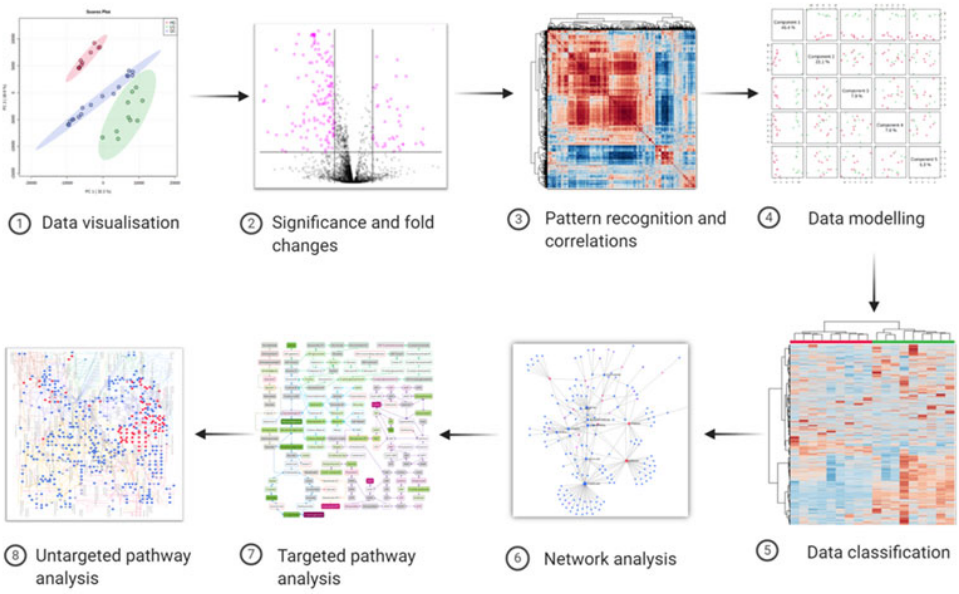
Data scaling: brings metabolite abundances onto a similar scale

Data transformation: ensures data is normally distributed



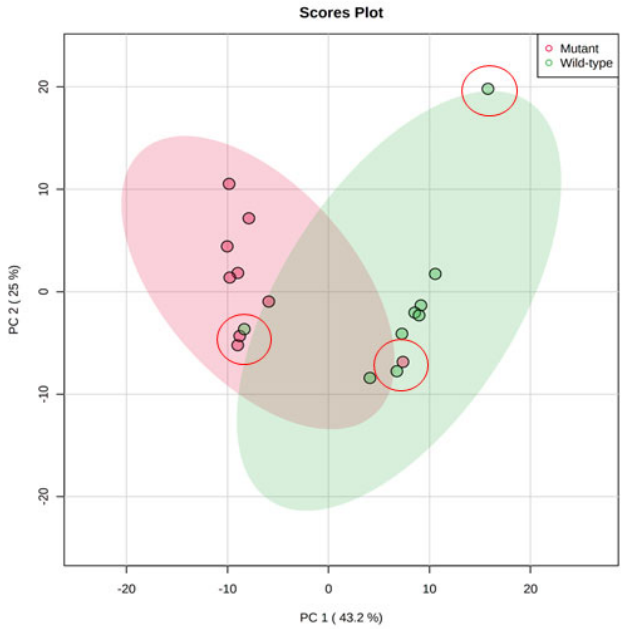
Class	Method	Formula	Unit	Goal	Advantages	Disadvantages
I	Centering	$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$	0	Focus on the differences and not the similarities in the data	Remove the offset from the data	When data is heteroscedastic, the effect of this pretreatment method is not always sufficient
II	Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$	(-)	Compare metabolites based on correlations	All metabolites become equally important	Inflation of the measurement errors
	Range scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{(x_{i_{max}} - x_{i_{min}})}$	(-)	Compare metabolites relative to the biological response range	All metabolites become equally important. Scaling is related to biology	Inflation of the measurement errors and sensitive to outliers
	Pareto scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	0	Reduce the relative importance of large values, but keep data structure partially intact	Stays closer to the original measurement than autoscaling	Sensitive to large fold changes
	Vast scaling	$\tilde{x}_{ij} = \frac{(x_{ij} - \bar{x}_i)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$	(-)	Focus on the metabolites that show small fluctuations	Aims for robustness, can use prior group knowledge	Not suited for large induced variation without group structure
	Level scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$	(-)	Focus on relative response	Suited for identification of e.g. biomarkers	Inflation of the measurement errors
III	Log transformation	$\tilde{x}_{ij} = {}^{10}\log(x_{ij})$ $\tilde{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	Log 0	Correct for heteroscedasticity, pseudo scaling. Make multiplicative models additive	Reduce heteroscedasticity, multiplicative effects become additive	Difficulties with values with large relative standard deviation and zeros
	Power transformation	$\tilde{x}_{ij} = \sqrt{(x_{ij})}$ $\tilde{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	√0	Correct for heteroscedasticity, pseudo scaling	Reduce heteroscedasticity, no problems with small values	Choice for square root is arbitrary.

DATA ANALYSIS AND INTERPRETATION OVERVIEW



The **purpose of data analysis** in metabolomics is to identify patterns and structures in the data which reveal biologically meaningful information.

DATA OVERVIEW: PRINCIPLE COMPONENTS ANALYSIS (PCA)



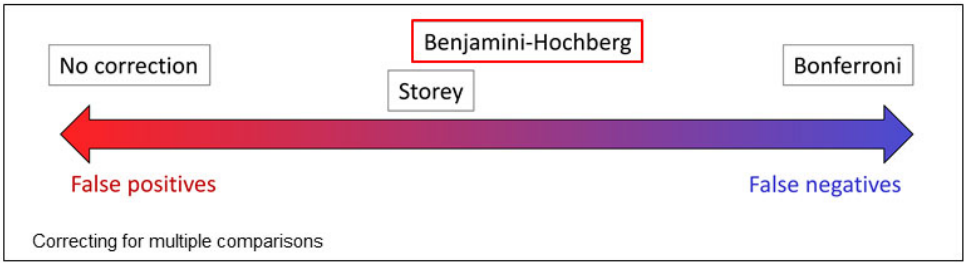
PCA in metabolomics

- Commonly applied as an overview of the dataset.
- Converts metabolite abundances into coordinates in multi-dimensional space with one principle component for each variable.
- Principle components maximise abundance variation in the data which generally decreases as the principle components increase.
- The first few principle components therefore usually show the maximal variation in the dataset.
- **Used to identify outlier samples or variables**.

IDENTIFYING THE SIGNIFICANCE OF METABOLITE CHANGES USING UNIVARIATE STATISTICS

$$Fold\ change\ (FC) = \frac{Final\ concentration}{Initial\ concentration}$$

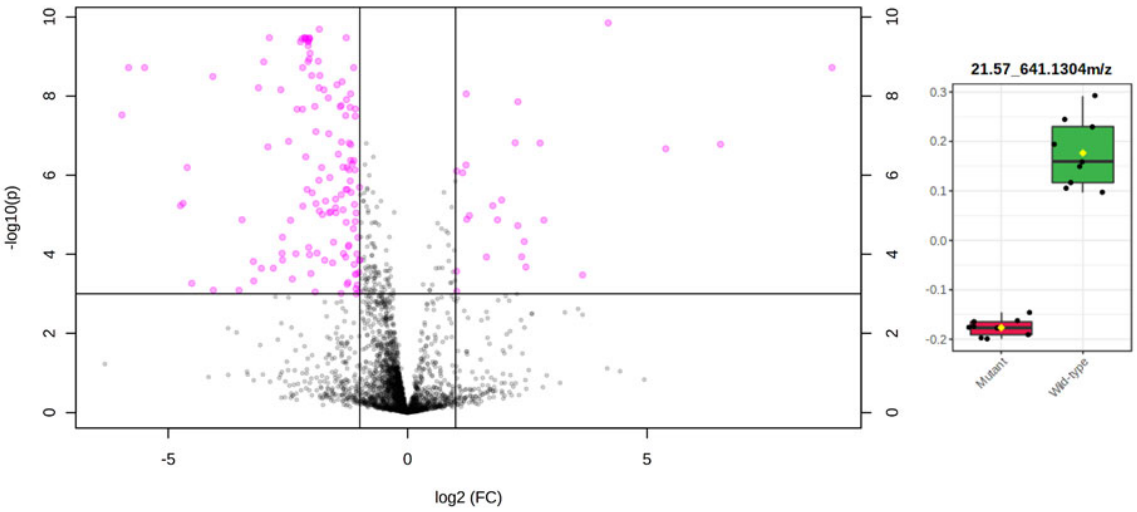
Type	Paired	Unpaired	Assumptions
Parametric	Student's t-test	Welch's t-test	The experimental population represents the 'true' population and is normally distributed
Non-parametric	Wilcoxin-signed rank	Mann-Whitney-U test	Does not assume the experimental population is normally distributed
ANOVA	Repeated measures ANOVA	One-way ANOVA	Analysis of variance (ANOVA) generalizes the t-test when the data belong to more than two groups. Assumes a normal distribution



"Have compound-features been modified by induced experimental differences and if so are these significant?"

VOLCANO PLOT

"How to select metabolites which are altered most significantly between experimental groups?"

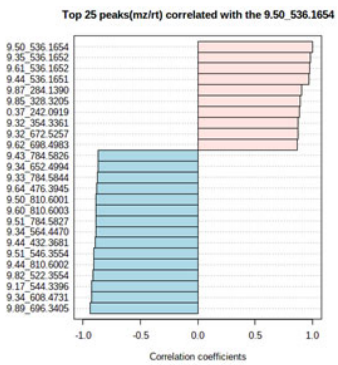


Some caution should be taken as **not all metabolites are regulated to the same extent**. Some pathways are more tightly regulated than others and hence a 1.5 fold change in one may be more important than a 2-fold change in another.

PATTERN RECOGNITION

Statistical approaches that calculate **correlation coefficients** can be used to identify the strength of linear relationships between variables. These commonly use **Pearson's r or Spearman's rank coefficients** to identify which variables correlated in the way they can in response to a change in biological condition. Correlation analysis can be used to answer simple questions such as ***which metabolites also increase (or decrease) significantly when citrate levels go up in hypoxic environments?*** The same approaches can also be used to identify more complicated relationships within the data, for examples which metabolites shows a similar pattern over a time-course.

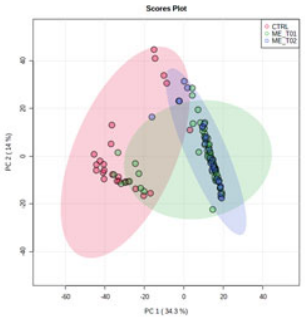
Although correlations can provide powerful information they are usually **sensitive to outliers, unequal variances and non-normality**. It should always be remembered than **correlation is not causation** and this approach does not provide any information about why the metabolite changes are correlated.



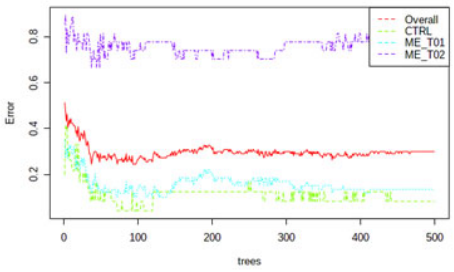
SUPERVISED MULTIVARIATE MODELLING OF METABOLOMICS DATA

Although univariate statistical analysis can help **identify metabolites that are significantly different** in abundance between experimental groups, univariate statistics ignores relationships between metabolites within datasets, they treat each compound-feature or metabolite as an independent variable.

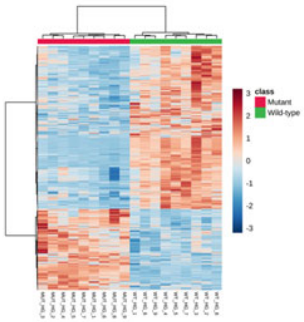
In reality we know this is not the case, that **many metabolites are inter-dependent** in a biological context and multivariate statistical approaches enables such relationships to be modelled as part of the data analysis: The benefit is that i) The analysis provides greater biological insight. ii) The reliability of the models can be statistically evaluated and ii) the models can be used prospectively for future studies.



Vector based approaches

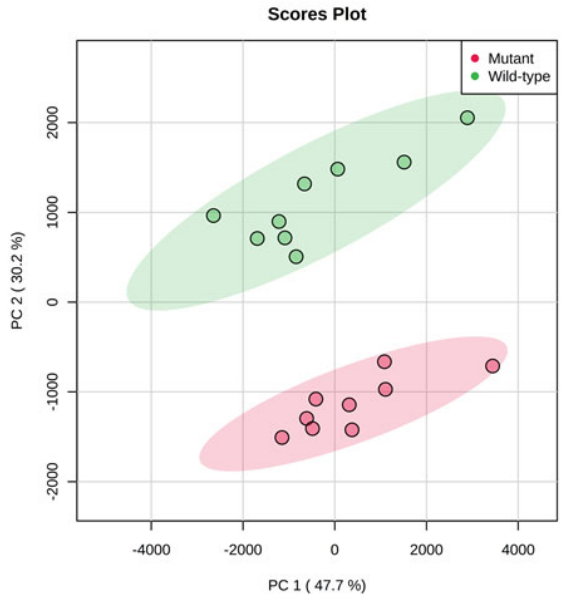


Random Forest classification



Hierarchical clustering

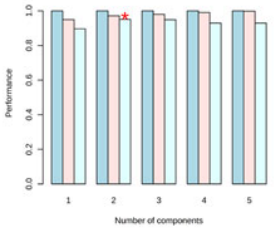
PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS (PLS-DA)

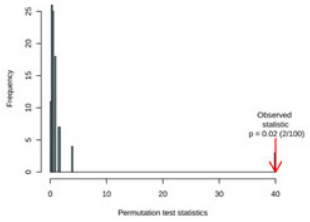


PLS-DA cross validation details:

Accuracy
R2
Q2

Measure	1 comps	2 comps	3 comps	4 comps	5 comps
Accuracy	1.0	1.0	1.0	1.0	1.0
R2	0.9494	0.97163	0.98006	0.99116	0.99821
Q2	0.89664	0.95087	0.94898	0.9297	0.92962

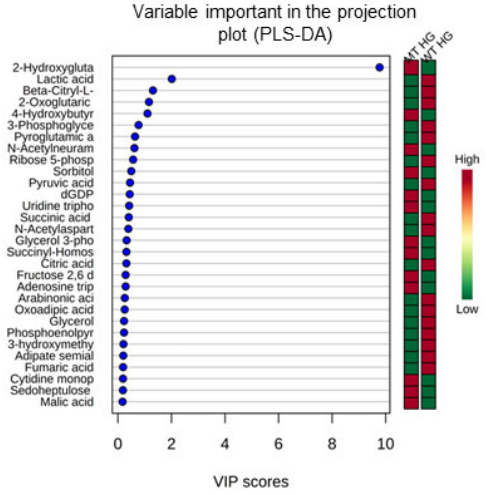




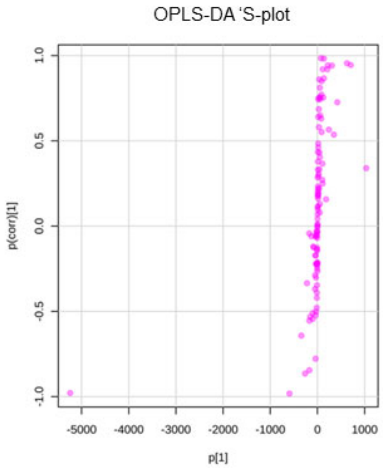
PLS-DA: A vector-based approach (like PCA) but rather than capturing only variance in the variables it also capture the covariance in the sample grouping.

Note: supervised multivariate techniques such as PLS-DA try to discriminate between experimental increasing the probability of doing so by chance. This can lead to **overfitting of data** and its important to **validate the models**.

PARTIAL LEAST SQUARES DISCRIMINANT ANALYSIS (PLS-DA)



When PLS-DA or OPLS-DA models have been constructed and validated they can be used to identify important variables.

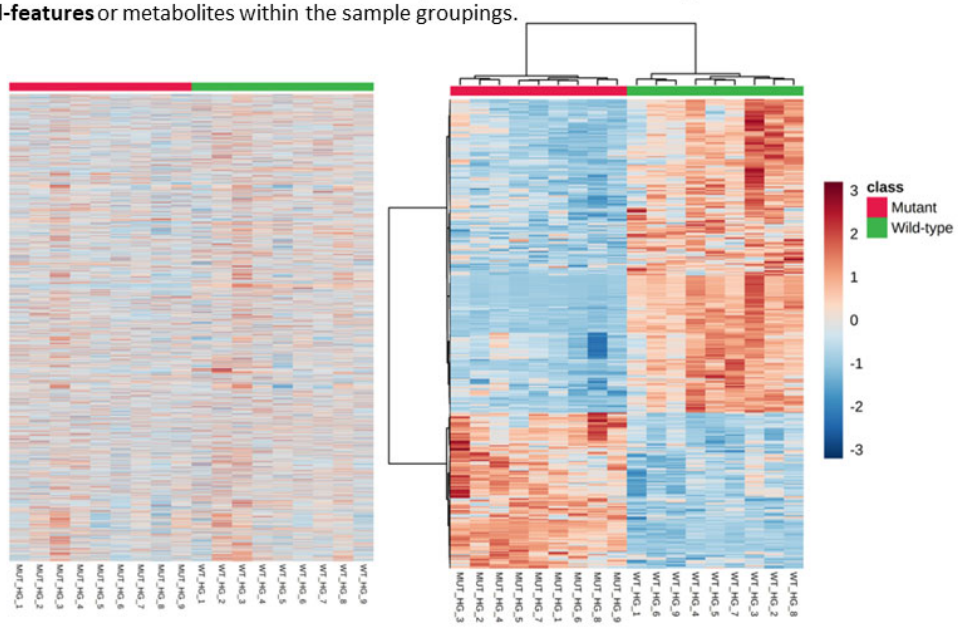


A variable importance in the projection plot (VIP plot) scores each variable according to the weighted sum of the squares of the PLS **loading determined by the amount of explained variance between the experimental groups.**

An S-plot combines the covariance and correlations. The Y-axis is the metabolite's **reliability score in the model and on the X-axis is the direction and magnitude of its deviation** from the control sample.

HIERARCHICAL CLUSTERING

Unsupervised multivariate statistical approach used to **organise complex datasets,** often visualised using heatmaps. **Dendograms are often associated with these heat-maps to illustrate the similarity between the different compound-features** or metabolites within the sample groupings.



Identifying relationships between metabolites

PUTTING DATA PROCESSING AND ANALYSIS INTO PRACTICE

MeaboAnalyst: https://www.metaboanalyst.ca/

MetaboAnalyst is a free online data processing and analysis platform provided for the metabolomics community by the Xia Lab @ McGill university in Canada. It provides software tools for data processing, analysis, visualisation and functional interpretation of targeted and untargeted metabolomics data.