



# Spectra Processing, Compound Annotation, Functional Insight and Causal Analysis using **MetaboAnalyst 6.0**

Jianguo (Jeff) Xia, Associate Professor

Canada Research Chair in Bioinformatics & Big Data Analytics

[jeff.xia@mcgill.ca](mailto:jeff.xia@mcgill.ca) | [www.xialab.ca](http://www.xialab.ca)

McGill University, Canada



**XiaLab.ca**

Empowering researchers through trainings, tools, and AI



**McGill**  
UNIVERSITY

# Schedule

## Part I: 2:15 p.m. – 4:15 p.m

- **2:15 – 3:00:** Background
  - ✓ General introduction
  - ✓ LC-MS & MS/MS spectral processing
  - ✓ From peaks to functions
- **3:00 – 3:20:** Live demo
- **3:20 – 4:15:** Hands on practice

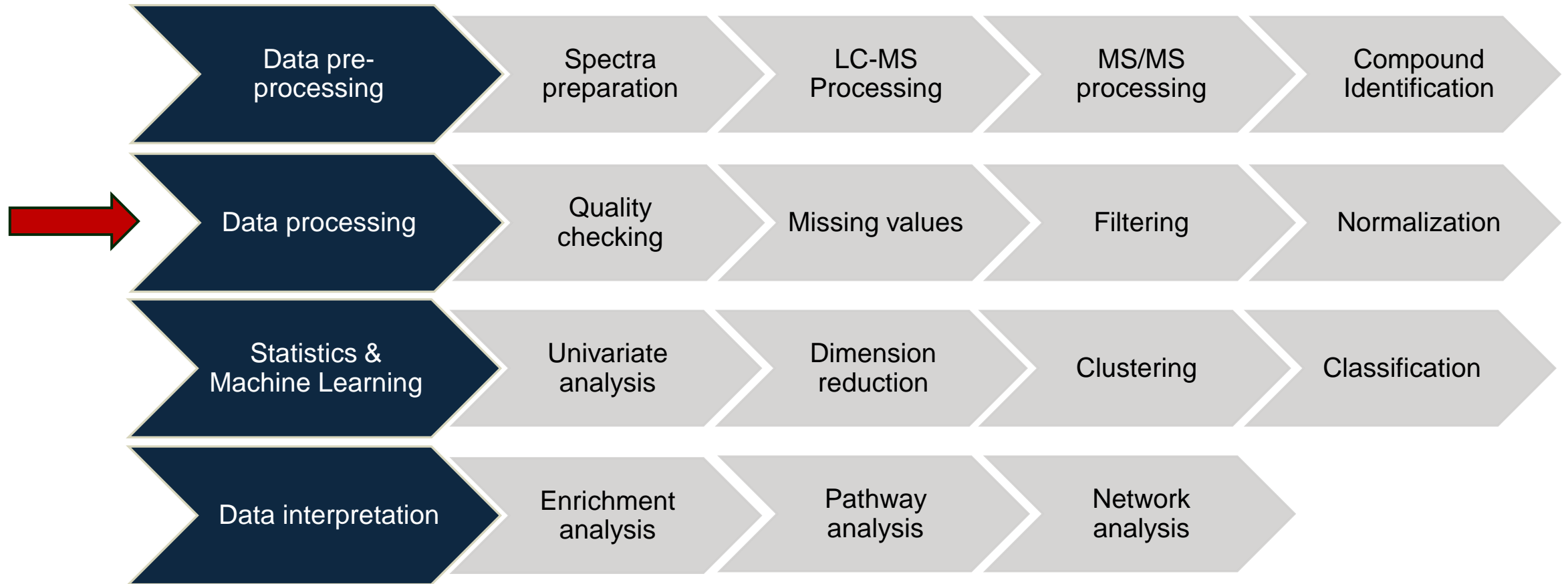
## Part II: 4:30 p.m. – 6:30 p.m.

- **4:30 – 5:10:** Background
  - ✓ Data processing
  - ✓ Statistical analysis
  - ✓ Causal analysis
- **5:10 – 5:40:** Live demo
- **5:40 – 6:15:** Hands on practice
- **6:15 – 6:30:** Summary & discussion

# Github Repository

- [https://github.com/xia-lab/Metabolomics\\_2024](https://github.com/xia-lab/Metabolomics_2024)
- Slides (in PDF format);
- Example data;
- Reference literatures;
- Contact information.

# Common Tasks for Metabolomics



# Data Processing

# MetaboAnalyst 6.0 Modules

Input Data Type	Available Modules (click on a module to proceed, or scroll down to explore a total of 18 modules including <a href="#">utilities</a> )				
LC-MS Spectra (mzML, mzXML or mzData)			Spectra Processing [LC-MS w/wo MS2]		
MS Peaks (peak list or intensity table)		Peak Annotation [MS2-DDA/DIA]	Functional Analysis [LC-MS]	Functional Meta-analysis [LC-MS]	
Generic Format (.csv or .txt table files)	Statistical Analysis [one factor]	Statistical Analysis [metadata table]	Biomarker Analysis	Statistical Meta-analysis	Dose Response Analysis
Annotated Features (metabolite list or table)		Enrichment Analysis	Pathway Analysis	Network Analysis	
Link to Genomics & Phenotypes (metabolite list)			Causal Analysis [Mendelian randomization]		

# Targeted metabolomics data

Sample	Label	Acetate	Acetone	Alanine	Betaine	Carnitine	Choline	Citrate	Creatine
Control_01	0	189.07	24.24	266.27	298.95	304.62	305.61	3969.16	366.7
Control_02	0	386.52	26.29	612.02	313.5	122.06	78.46	2075.6	435.99
Control_03	0	506.28	27.84	347.32	101.49	88.82	35.88	745.5	83.39
Control_04	0	51	177.56	468.42	172.65	133.28	0	10797.57	77.69
Control_05	0					21	187.98	1016.97	83.39
Disease_01	1					04	122.4	1486.6	152.5
Disease_02	1					86	44.93	3327.06	263.41
Disease_03	1	231.59	12.91	226.63	0	0	35.26	758.18	50.36
Disease_04	1	285.55	0	217.43	0	77.79	0	609.23	0
Disease_05	1	353.51	15.87	699.81	98.7	458.81	112.21	3415.49	229.79

Targeted metabolomics  
(samples in rows)

Sample	Contr_1	Contr_2	Contr_3	Contr_4	Contr_5	Disease_1	Disease_2	Disease_3	Disease_4	Disease_5
Label	0	0	0	0	0	1	1	1	1	1
Acetate	189.07	386.52	506.28	51	733.45	315.12	325.39	231.59	285.55	353.51
Acetoacetate	0	0	145.96	232.69	0	148.37	0	56.19	0	0
Acetone	24.24	26.29					29	12.91	0	15.87
Alanine	266.27	612.02					53	226.63	217.43	699.81
Betaine	298.95	313.5					5.1	0	0	98.7
Carnitine	304.62	122.06	88.82	133.28	89.21	32.04	89.36	0	77.79	458.81
Choline	305.61	78.46	35.88	0	187.98	122.4	44.93	35.26	0	112.21
Citrate	3969.16	2075.6	745.5	10797.57	1016.97	1486.6	3327.06	758.18	609.23	3415.49
Creatine	366.7	435.99	83.39	77.69	83.39	152.5	263.41	50.36	0	229.79

Targeted metabolomics  
(samples in columns)

# Untargeted metabolomics data

Sample	men_Pt004_RPLCpos	men_Pt007_RPLCpos	men_Pt008_RPLCpos	men_Pt010_RPLCpos	men_Pt013_RPLCpos	men_Pt020_RPLCpos
Label	Covid	Covid	Covid	Covid	Covid	Covid
68.9948_646.68	1158657.48719	1151894.23428	1189086.78073	1285945.06468	1196004.21814	1018581.17851
69.995_641.77	107433.72307	126812.99598	124835.38621	160081.36667	132737.27627	120183.30701
70.0663_32.7	10385.57069	13137.07876	15305.24912	5714.296	1831.73589	627.95873
81.5222_642.31	209861.50958	194875.48234	210983.40708	211579.43879	209288.2791	208780.21243
84.0461_30.06	53988.89851	44693.63728	54104.51066	34887.7869	23436.84666	46781.95438
84.9606_643.25	186784.7	181153.3814	285718.9142	291334.92371	189987.5361	134815.70457
84.9607_664.62					2.48874	124820.09035
84.9607_716.7	4				6.55494	749034.65826
85.0302_30.53					7.91671	2616.3663
86.0976_89.04					7.66029	129637.63897
87.5108_642.38	29017.3417	46430.91348	29443.40354	50376.17085	45349.23108	
88.5106_646.42	467439.8206	452376.17516	486908.8982	488320.52256	506518.4509	
89.5084_646.45	8408192.9225	8132241.81496	8477437.65518	8543761.98101	8679306.12814	
89.6056_646.48	101573.6793	89544.45735	97535.64361	116017.93274	127775.03719	
89.9406_643.85	169537.75028	169867.84414	201913.54163	156549.447	152568.62769	
90.0093_646.69	839967.52451	823526.37057	859947.96708	900633.07574	877001.83348	
90.5085_641.82	1539089.51759	1263856.98911	1444405.93286	1983990.84075	1470945.60964	
90.5086_482.02	14153278.42237	14091173.41408	14308006.37643	14770096.07378	14987402.61328	

Sample	Class	Bin.9.98	Bin.9.94	Bin.9.90	Bin.9.86	Bin.9.82	Bin.9.78	Bin.9.74	Bin.9.70	Bin.9.66
P002	patient	2.00E-05	2.00E-05	2.00E-05	2.00E-05	4.00E-05	2.00E-05	1.00E-05	0	2.00E-05
P012	patient	-3.00E-05	-3.00E-05	-3.00E-05	-1.00E-05	-1.00E-05	0	3.00E-05	3.00E-05	4.00E-05
P014	patient	0.00024	0.00017	0.00016	0.00018	0.00014	0.00015	0.00016	2.00E-04	0.00021
P027	patient	9.00E-05	0.00013	9.00E-05	9.00E-05	0.00014	1.00E-04	0.00014	0.00016	0.00013
P034	patient	0.00012	0.00015	2.00E-05	3.00E-05	7.00E-05	7.00E-05	3.00E-05	4.00E-05	0.00012
P037	patient	1.00E-05	-1.00E-05	2.00E-05	-3.00E-05	-6.00E-05	6.00E-05	4.00E-05	6.00E-05	-9.00E-05
P038	patient	1.00E-05	1.00E-05	1.00E-05	-5.00E-05	-1.00E-05	-5.00E-05	2.00E-05	5.00E-05	6.00E-05
P041	patient	-9.00E-05	-6.00E-05	-7.00E-05	1.00E-05	-0.00012	-6.00E-05	-1.00E-05	-1.00E-05	-3.00E-05
P042	patient	0.00015	8.00E-05	0.00017	3.00E-05	3.00E-05	-0.00011	2.00E-05	6.00E-05	0.00015
P049	patient								-0.00015	2.00E-05
P056	patient								-0.00015	0
P058	patient								-0.00015	-3.00E-05
P060	patient								-0.00015	2.00E-05
P064	patient								-0.00015	7.00E-05
P065	patient								-0.00015	5.00E-05
P070	patient	-1.00E-05	-1.00E-05	-1.00E-05	-3.00E-05	-1.00E-05	0	0	-2.00E-05	3.00E-05
P080	patient	0	1.00E-05	0	0	1.00E-05	0	2.00E-05	2.00E-05	3.00E-05
P085	patient	1.00E-05	7.00E-05	5.00E-05	5.00E-05	6.00E-05	3.00E-05	2.00E-05	1.00E-04	6.00E-05
P086	patient	-1.00E-05	1.00E-05	1.00E-05	-1.00E-05	3.00E-05	3.00E-05	-2.00E-05	7.00E-05	-2.00E-05
P089	patient	1.00E-05	0	1.00E-05	2.00E-05	3.00E-05	3.00E-05	4.00E-05	5.00E-05	6.00E-05
P092	patient	-3.00E-05	-3.00E-05	-3.00E-05	-1.00E-05	0	0	3.00E-05	-1.00E-05	3.00E-05
P099	patient	0	0	2.00E-05	-1.00E-05	-4.00E-05	3.00E-05	5.00E-05	1.00E-05	4.00E-05
P113	patient	-2.00E-05	1.00E-05	-4.00E-05	-2.00E-05	-7.00E-05	-4.00E-05	-1.00E-05	1.00E-05	-4.00E-05
P013b	patient	-2.00E-05	-2.00E-05	-1.00E-05	-1.00E-05	-2.00E-05	0	0	1.00E-05	0











# Metadata table

- Common in observational field studies
  - Clinical
  - Exposomics
  - Epidemiology
- Study design & context
  - Primary experimental factors
  - Covariates

Sample	Diagnosis	Gender	Treatment	Age
S1	COVID	Male	non_Treated	62
S2	COVID	Male	non_Treated	44
S3	COVID	Male	Treated	54
S4	COVID	Male	non_Treated	62
S5	COVID	Male	Treated	82
S6	COVID	Male	Treated	65
S7	COVID	Female	Treated	49
S8	COVID	Female	Treated	42
S9	COVID	Female	Treated	56
S10	COVID	Female	Treated	56
S11	COVID	Female	Treated	69
S12	HC	Male	non_Treated	24
S13	HC	Female	non_Treated	38
S14	HC	Female	non_Treated	42
S15	HC	Female	non_Treated	40
S16	HC	Female	non_Treated	56
S17	HC	Male	non_Treated	57
S18	HC	Male	non_Treated	57
S19	HC	Male	non_Treated	60
S20	HC	Male	non_Treated	62
S21	HC	Male	non_Treated	55

# Understanding & formatting metadata

- Essential for downstream analysis
- Categorize as “Categorical” or “Continuous”
  - For categorical, must have at least 2 groups with at least 3 replicates each
- No missing values
- First meta-data column will be considered primary variable by default

Name	Status	Type	Edit	Remove
TCE_Exp_Category	OK	Categorical ▼	Edit	
TCE_Exp_Conc	OK	Categorical	Edit	
Age	OK	Continuous ▼	Edit	
Sex	OK	Categorical ▼	Edit	
Smoking_Status	OK	Categorical ▼	Edit	
Alcohol_Use	OK	Categorical ▼	Edit	
BMI	OK	Continuous ▼	Edit	
Batch	OK	Categorical ▼	Edit	

# View and edit metadata

Edit metadata column

Reset

Sample ↑↓	Diagnosis ↑↓ Discrete	Age ↑↓ Continuous	OGTT ↑↓ Continuous	HbA1c ↑↓ Continuous	BMI ↑↓ Continuous		
DP063	IGT	84	NA	5.9	22.6562		
DP064	ND	59	NA	5.6	24.3375		
DP096	T3D	70	12.2	7	31.6404		
DP099	T2D	66	NA	8.4	29.7056	✓ ✗	
DP102	T2D	75	NA	7.3	27.9155		
DP106	T3D	70	NA	6.8	34.1598		
DP107	T2D	56	NA	7.4	42.0508		
DP108	IGT	71	5.29	5.8	27.7389		
DP112	T3D	63	13.25	5.2	23.4236		
DP114	T3D	62	9.23	6.7	25.5132		
DP117	T3D	84	NA	5.9	27.6361		
DP122	T2D	70	NA	8.8	26.0617		
DP123	IGT	56	9.81	5.6	31.1846		
DP124	T2D	71	NA	6.5	25.204		
DP126	T2D	54	NA	7.3	23.4509		

<< < 1 2 3 > >> 15

Edit

Edit metadata columns: ×

Include/Exclude

Primary metadata

Order (factor-level)

Only applicable to categorical metadata.

Select metadata: 

Diagnosis

↑

↓

↕

T3D

T2D

IGT

ND

Update

Cancel

# Feature filtering (I)

- ❖ Not all features are informative
- ❖ There are redundancies in omics data for most features
- ❖ Filtering non-informative features before statistical analysis can often significantly improve the power



# Feature filtering (II)

## Three types of filters

### ❖ Low reliability filter




- Too many missing values
- Hard to measure metabolites: low repeatability based on QC samples

### ❖ Low abundance filter

- Variables of very small values (close to baseline or detection limit).

### ❖ Low variance filter

- Variables that are near-constant values throughout the experiment conditions (housekeeping or homeostasis)

Reliability filter:	<input type="checkbox"/> Filtering features based on technical repeatability QC samples	RSDs greater than:  25%
Variance filter:	<input checked="" type="radio"/> Interquartile range (IQR) <input type="radio"/> Standard deviation (SD) <input type="radio"/> Median absolute deviation (MAD) <input type="radio"/> Relative standard deviation (RSD = SD/mean) <input type="radio"/> Non-parametric relative standard deviation (MAD/median)	Percentage to filter out:  5%
Abundance filter:	<input checked="" type="radio"/> Mean intensity value <input type="radio"/> Median intensity value	Percentage to filter out:  0%

**DO NOT filter features based on their p-values or fold changes at this stage**

# Normalization (I)

## Three types of normalization

- Sample-wise normalization aims to make each sample (row) comparable to each other (i.e. tissue volume, urine samples with different dilution effects)
- Data transformation (such as log, cubic root) transform individual values independently
  - Most metabolomics data are log-normal
- Data scaling (auto, pareto) takes into consideration of the distribution (range, SD) of individual features
  - Unique to the current data

**Sample normalization**

- ☒ None
- ☐ Sample-specific normalization (i.e. weight, volume) [Specify](#)
- ☐ Normalization by sum
- ☐ Normalization by median
- ☐ Normalization by a reference sample (PQN) [Specify](#)
- ☐ Normalization by a pooled sample from group (group PQN) [Specify](#)
- ☐ Normalization by reference feature [Specify](#)
- ☐ Quantile normalization (suggested only for > 1000 features)

**Data transformation**

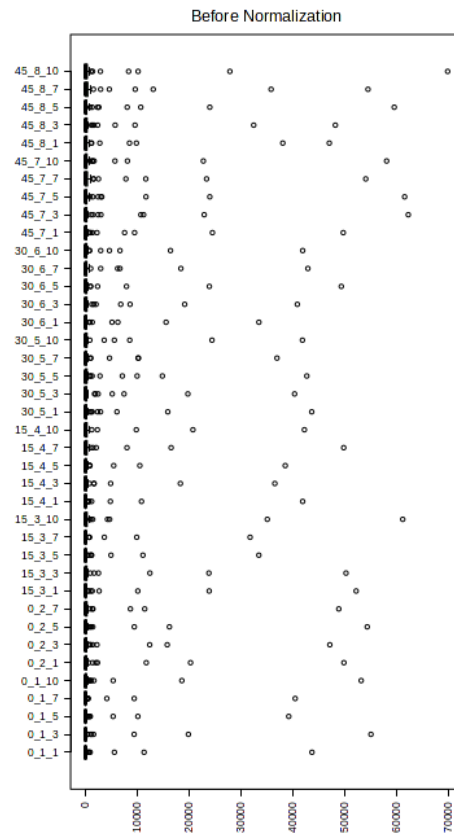
- ☒ None
- ☐ Log transformation (base 10)
- ☐ Square root transformation (square root of data values)
- ☐ Cube root transformation (cube root of data values)

**Data scaling**

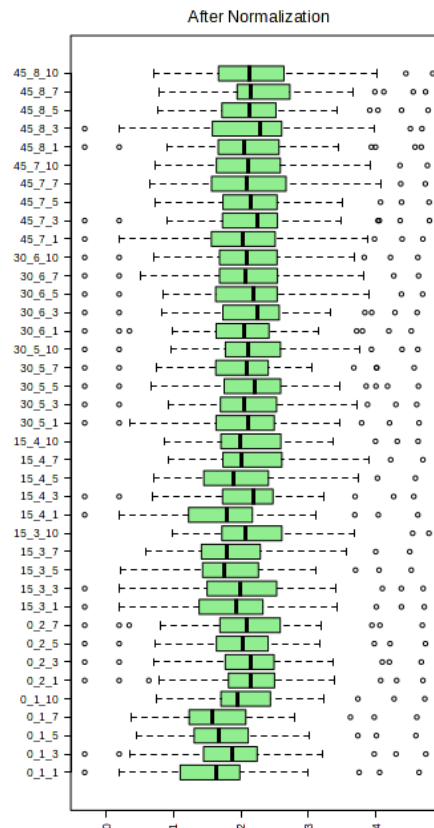
- ☒ None
- ☐ Mean centering (mean-centered only)
- ☐ Auto scaling (mean-centered and divided by the standard deviation of each variable)
- ☐ Pareto scaling (mean-centered and divided by the square root of the standard deviation of each variable)
- ☐ Range scaling (mean-centered and divided by the range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

# Sample-level normalization

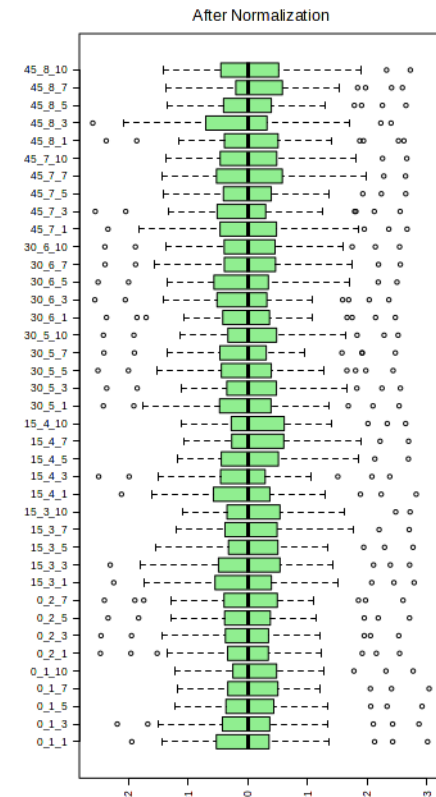
- Adjusting technical or biological bias or inconsistencies



Raw



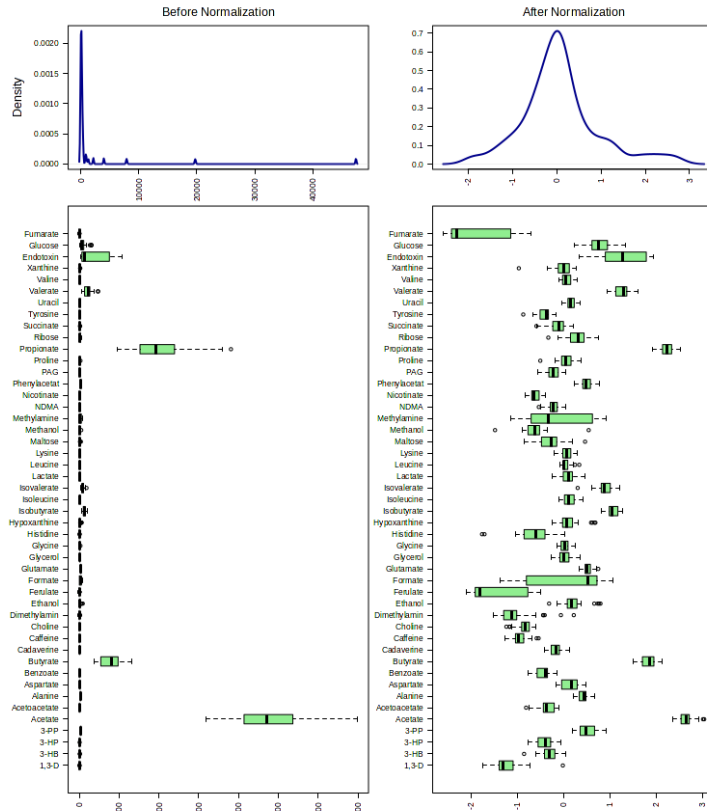
Log



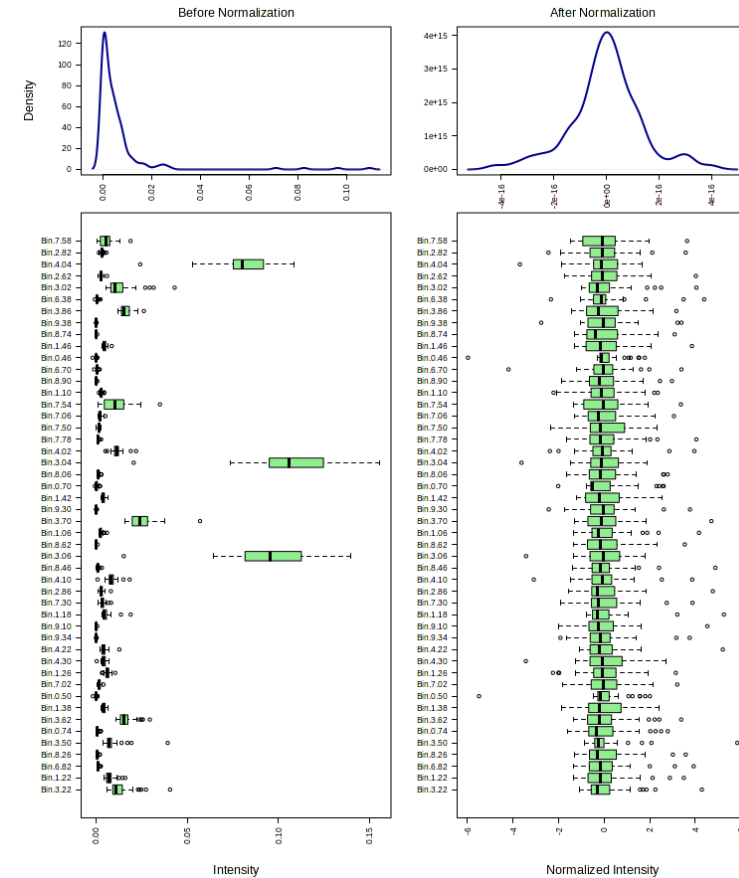
Median norm + Log

# Feature-level normalization

- Making features more “normal” and comparable



Log transformation



Auto-scaling



## Statistical Analysis

- identify significant features & patterns

# Univariate Analysis

Test each feature individually (ignore their correlations)

## 1.T-tests & Fold Change Analysis

- Compare the means between 2 conditions

## 2.ANOVA & post-hoc analysis

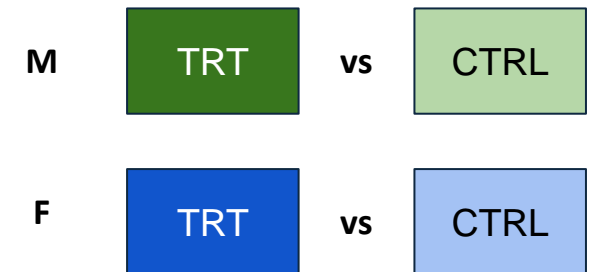
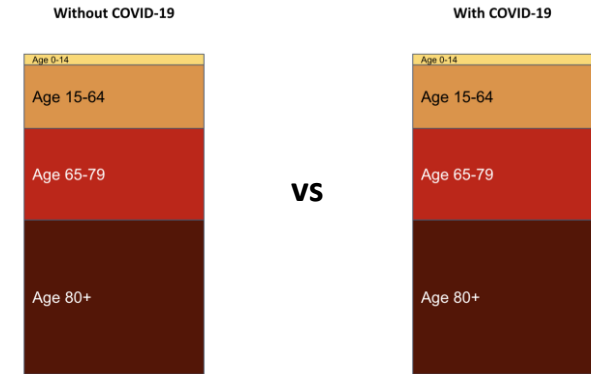
- One factor with more than 2 levels (One-way ANOVA)
- Two factors (Two-way ANOVA)

## 3.Linear modeling (i.e., limma): more flexible analysis

- Multiple factors (metadata table)
  - Time series
  - Covariates analysis

# Dealing with covariates

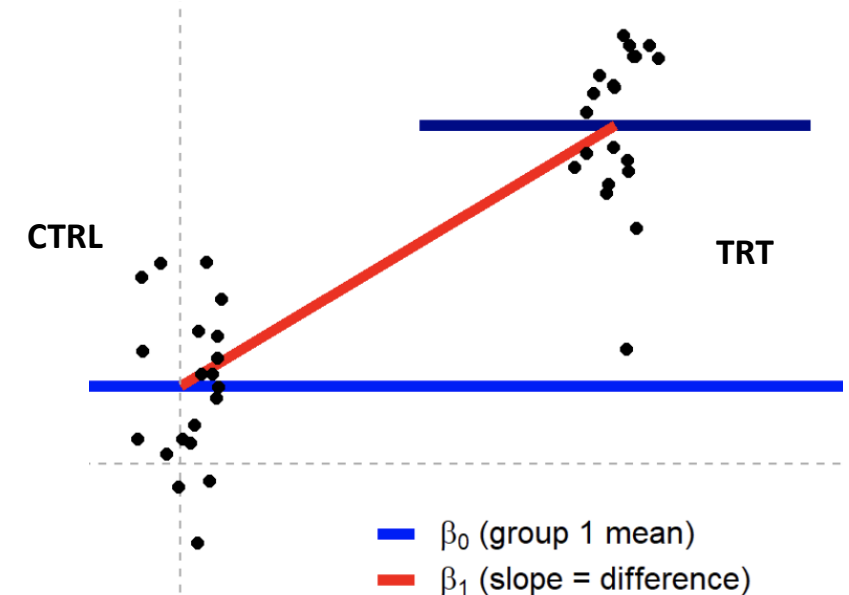
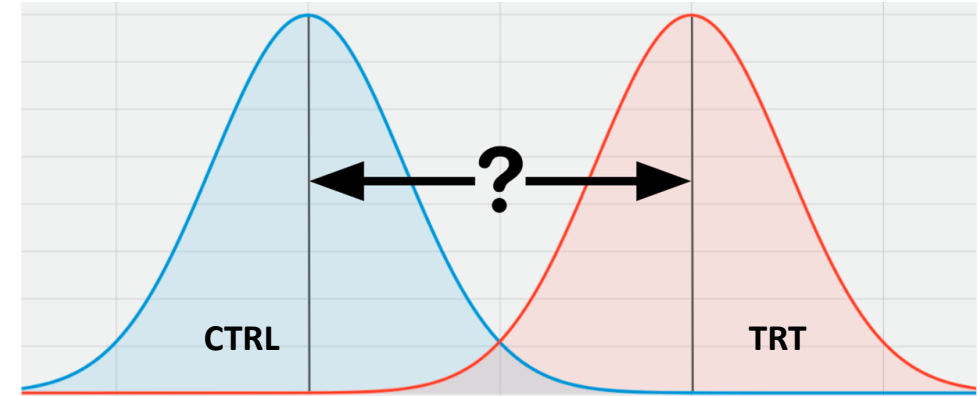
- Experimental design:
  - Control group that 'matches' your group of interest
- Data analysis (limited by sample size):
  1. Reduce the number of factors
    - use PCA + other tools to choose only factors with most influence on the data
  2. Stratification - for factors with few classes, split up the data and analyze separately
    - Analyze male / female separately
  3. Taking into account
    - Covariates



$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n$$

# T-test vs. linear regression

- You can do t-test with linear regression:
- $y = B_0 + B_1 * x$ 
  - $y$ : level of metabolite A
  - $x$ : variable of interest
  - Categorical variables expressed using 'dummy variables'
  - Null hypothesis:  $B_1 = 0$
  - Alternative hypothesis:  $B_1 > 0$



# Linear regression is flexible

- Linear regression is more flexible than the classical t-test:
  - Predictor variables (x) can be continuous or categorical
  - You can have **multiple predictor variables**
- Multiple linear regression:
  - $y = B_0 + B_1 * x_{\text{diagnosis}} + B_2 * x_{\text{age}}$
  - Coefficient estimates: relationship between  $x_i$  and  $y$  *with all other variables held constant*
  - We can generate a t-stat for any coefficient using the same formula from before:  $t_i = (B_i \text{ coefficient estimate}) / (\text{standard error of } B_i \text{ coefficient estimate})$
- Coefficients are unstable when the predictors are highly correlated

# Method overview

Overview to understand  
structure of data & metadata

Simple yet effective univariate  
statistical analysis

Advanced multivariate statistics  
& machine learning

## Data and Metadata Overview

### [Metadata Visualization](#)

Users can explore the metadata patterns and correlations through intuitive graphics. It is very useful for users to identify highly dependent metadata and quickly assess the overall patterns of the metadata.

### [Interactive PCA Visualization](#)

Users can visualize data using different colors or shapes based on selected metadata in an 2D and 3D (interactive) PCA plots. It is very useful to detect overall patterns of data with regard to different metadata.

### [Hierarchical Clustering and Heatmap Visualization](#)

This method displays data and metadata in the form of colored cells. It provides direct visualization of feature abundances across different samples and metadata.

## Univariate Analysis

### [Linear Models with Covariate Adjustment](#)

This approach uses linear models (limma or lm) to perform significance testing with covariate adjustments. Users can choose different metadata to be included in the analysis.

### [Correlation and Partial Correlation Analysis](#)

This approach allows users to explore the correlations or partial correlations (with covariate adjustments) between metabolomics features and different metadata of interest.

### [Two-way ANOVA \(ANOVA2\)](#)

This approach provides classical two-way ANOVA based on the two factors selected by users. For time-series data, users should choose within-subjects ANOVA.

## Multivariate Analysis

### [ANOVA Simultaneous Component Analysis \(ASCA\)](#)

This approach is designed to identify major patterns with regard to the two given factors and their interaction. The implementation was based on the algorithm described by [AK Smilde, et al.](#) with additional improvements on feature selection and model validation.

### Multivariate Empirical Bayes Analysis of Variance (MEBA) for Time Series

This approach is designed to compare temporal profiles across different biological conditions. It is based on the timecourse method described by [YC Tai, et al.](#)

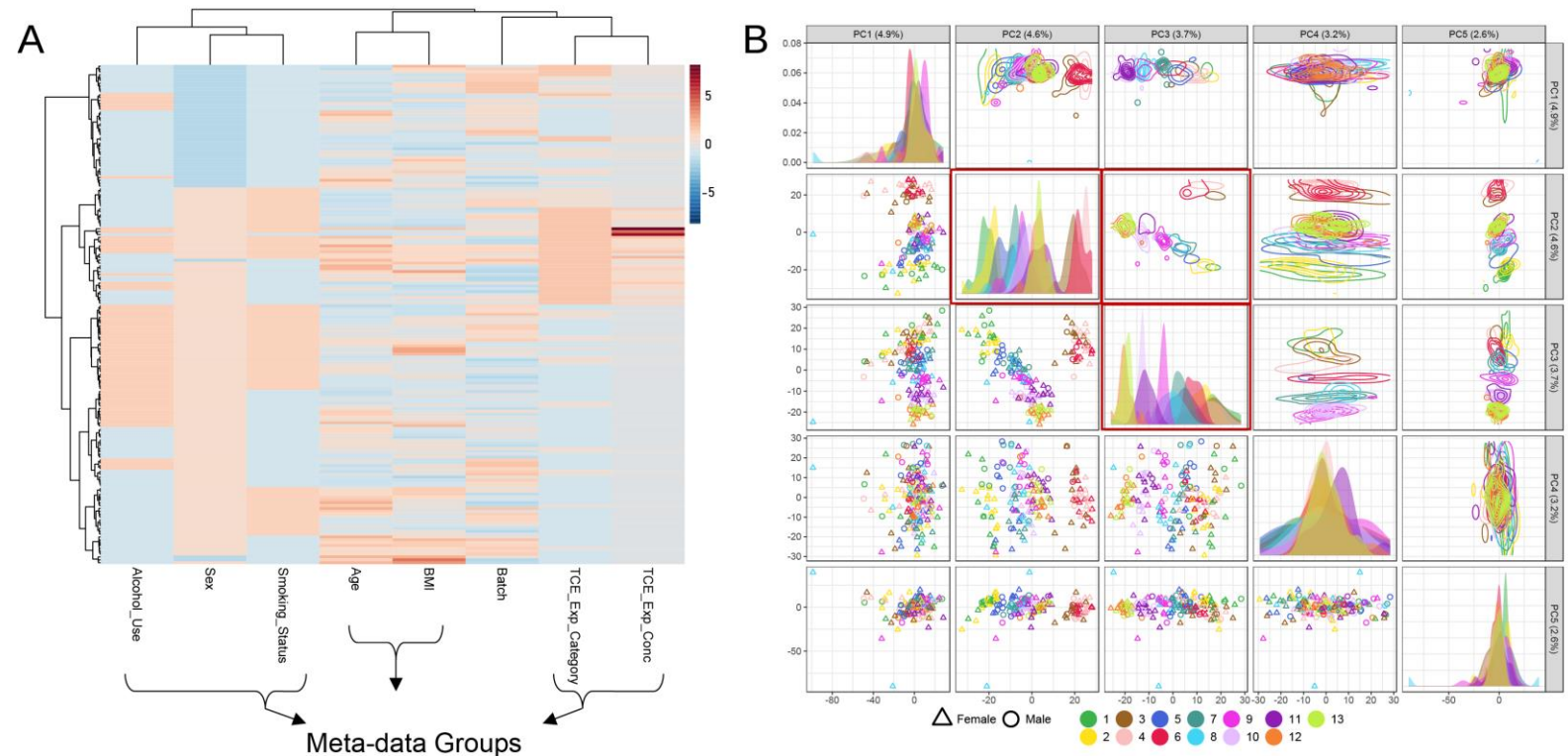
## Supervised Classification

### [Random Forest](#)

This machine learning approach is designed to perform classification and feature selection analysis. Users can also test contribution of meta-data to class prediction.

# Reducing metadata factors in analysis

- Factors maybe redundant (colinear)
  - ➔ Keep only one
- Overlay factors on omics data using PCA
  - ➔ Keep those with effects



# Linear models for complex design

Simple Metadata

Complex Metadata

Primary metadata:

Diagnosis

Reference group:

T3D

Contrast:

All contrasts (ANOVA-style)

Covariates (control for):

Blocking factor:

☐ --- Not Available ---

Adjust using robust trend

☐ Diagnosis

☐ Age

☐ OGTT

☐ HbA1c

☐ BMI

Submit

Did you know?

To perform multi-factor comparison analysis for complex metadata, we leverage the linear models with covariate adjustments of [limma](#) for its high-performance implementation. Some data may include some form of blocking in the study design, which can be modeled as either fixed or random effects. Please note that although you can model random effects, we in general recommend keeping the fixed effect model not only is computationally more efficient but also is more consistent with the interpretation of differential expression analysis (2009) for more technical discussions.

a

b



# Fixed vs. random effects

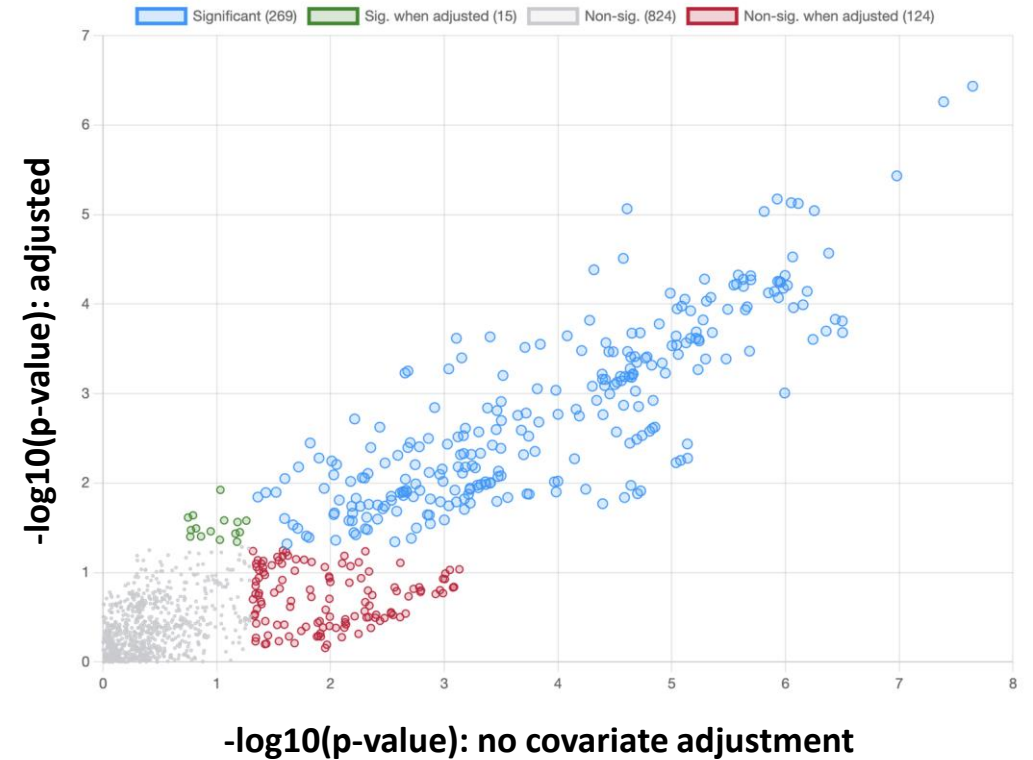
- Fixed effects = covariates
  - Simply variables included in the regression model
  - Age, sex, tissue, etc
- Random effects = blocking factor
  - Accounted for with multi-level modeling
  - Batch, subjects
- In general, we **recommend using fixed effects for simplicity & computational efficiency in most cases**. Treating the blocks as a fixed effect has the huge advantage of being able to quantify block-by-factor interactions which is perhaps the best method to quantify the robustness of your design structure.

Primary metadata:	Diagnosis ▼
Covariates (control for):	Gender X ▼
Blocking factor:	-- Unspecified -- ▼

# Interpretation of coefficient results

Multiple linear regression example:

- $y = B_0 + B_1 * x_{\text{diagnosis}} + B_2 * x_{\text{age}}$
- By including  $B_2 * x_{\text{age}}$  in the model, we account for effects of age
- Extract  $B_1$  from the model:
  - $B_1$  value = magnitude & direction of relationship between metabolite 'y' and  $x_{\text{diagnosis}}$
  - $B_1$  p-value = statistical significance of relationship



**Linear model with covariate adjustment**

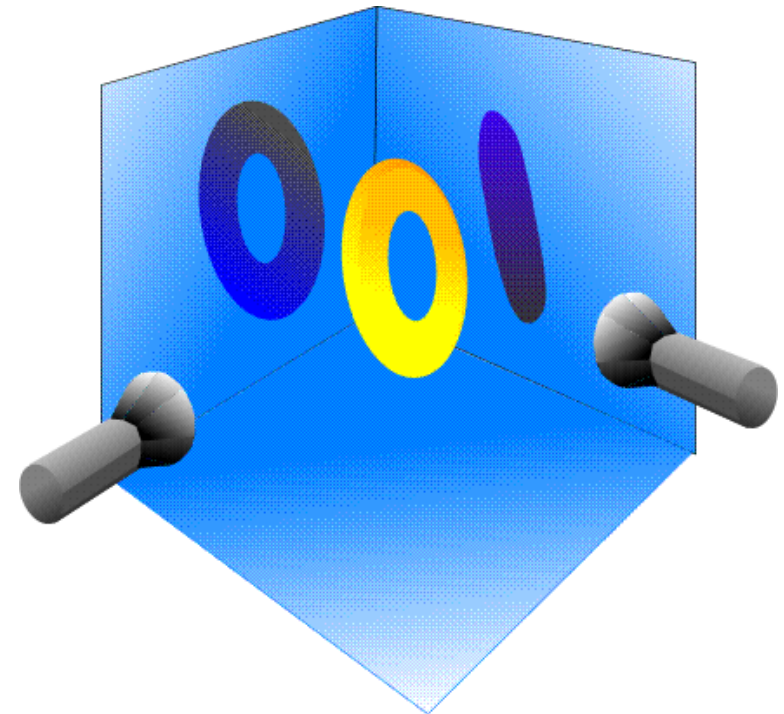
# Multivariate Analysis

# Principal Component Analysis (PCA)

- Project high-dimensional omics data into low dimensions (two or three PCs)
- Works best when there are a large number of variables are correlated
  - This is particularly relevant for untargeted metabolomics where single metabolites can generate several peaks
- PCA is very useful for:
  - Data overview
  - Outlier detection
  - Look at relationships between variables

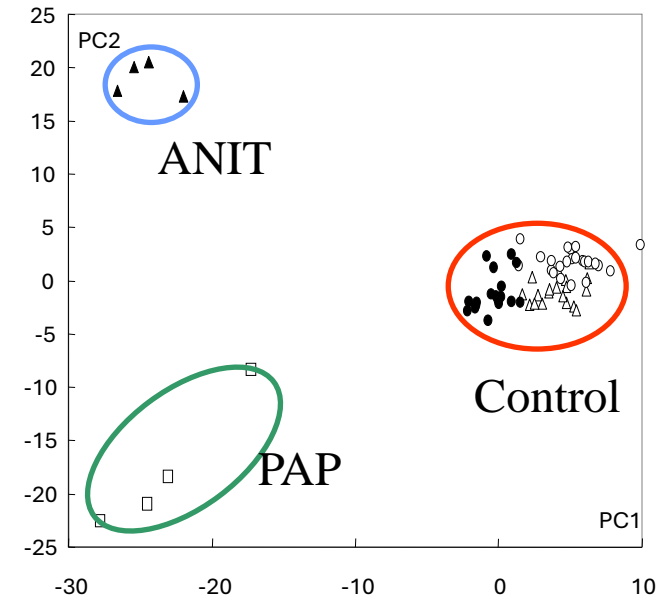
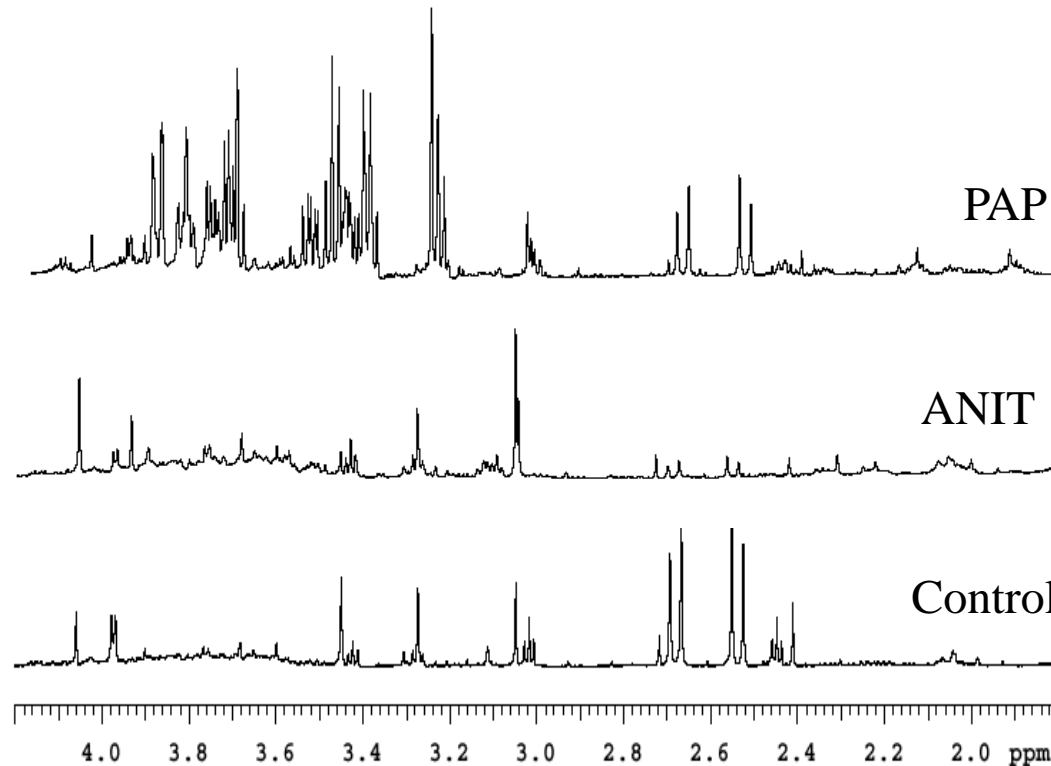
# Principal Component Analysis (PCA)

- Project high-dimensional data into lower dimensions that capture the **most variance** of the data
- Assumption:  
**Main directions of variance**  
 **$\approx$  major data characteristics**



# PCs capture the main variance in omics data

Hundreds of variables  $\longrightarrow$  2 components



Scores plot

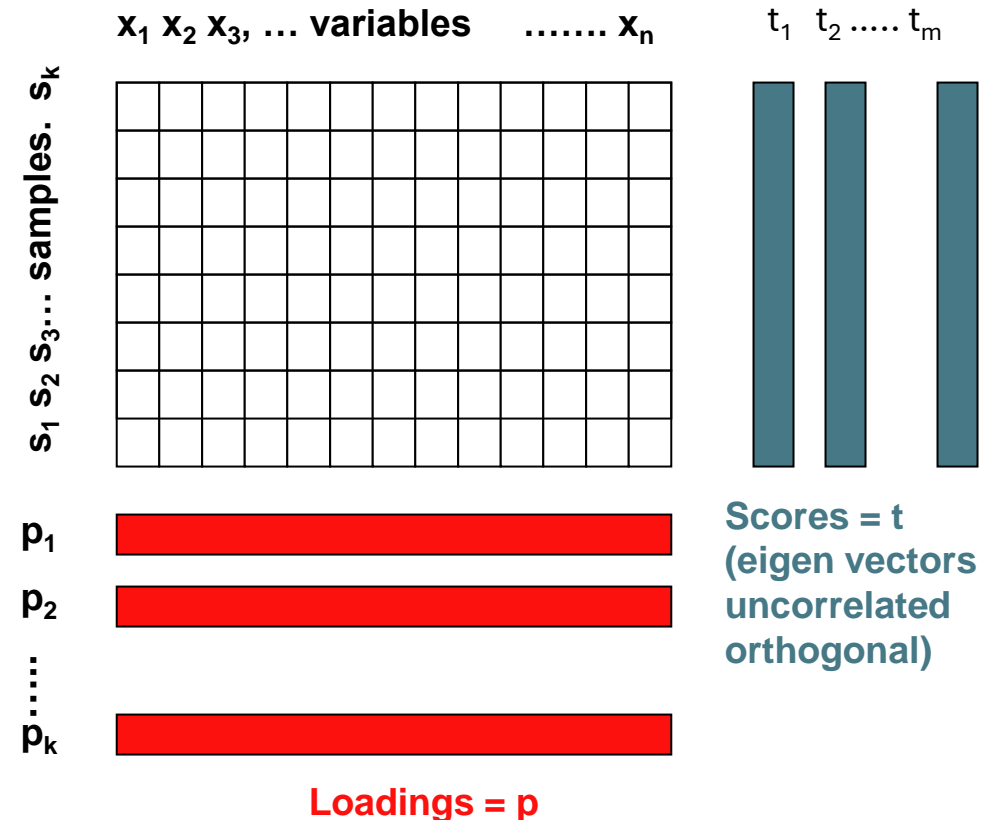
# PCA – under the hood

- Orthogonal linear transformation
- PCA transforms data to a new coordinate system so that the greatest variance of the data comes to lie on the first coordinate (1st PC), the second greatest variance on the 2nd PC etc.

**Scores = Loadings x Data**

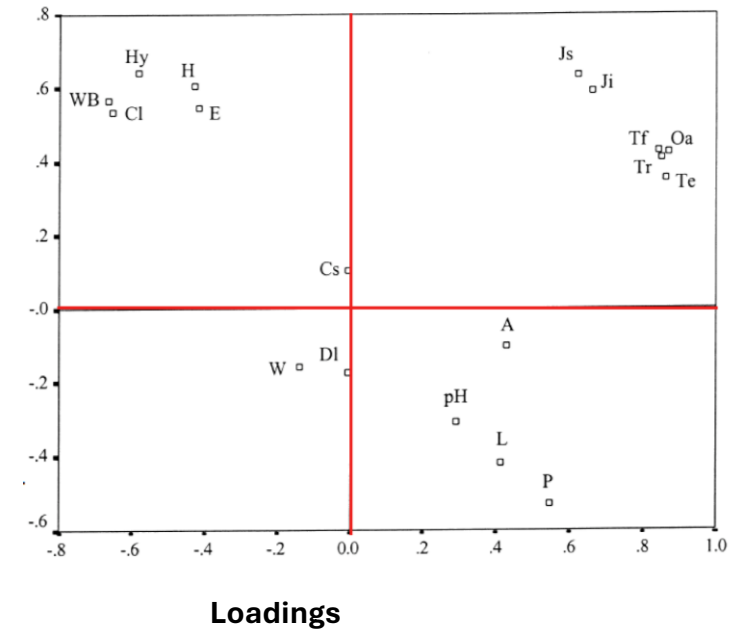
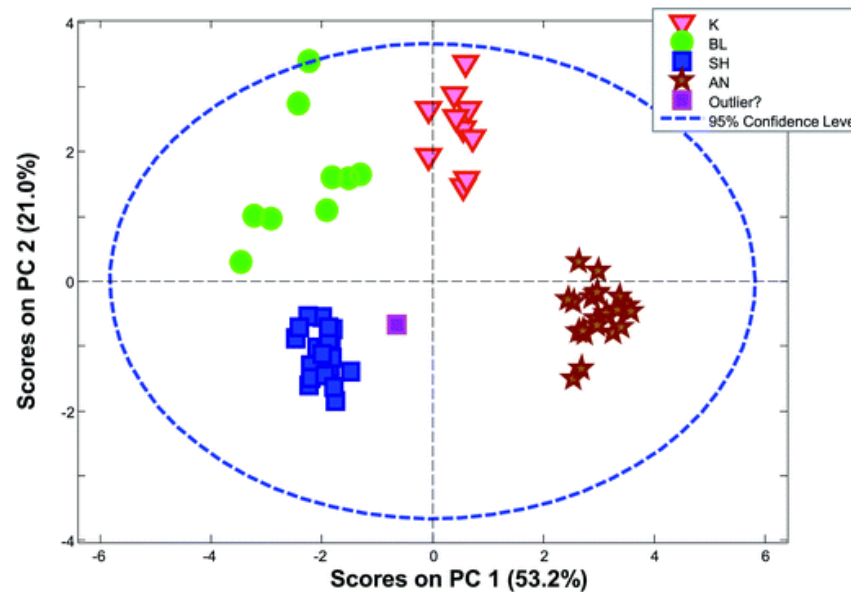
$$t_1 = p_1x_1 + p_2x_2 + p_3x_3 + \dots + p_nx_n$$

.....



# Scores & loadings plots

- Sample patterns (scores) are directly related to feature patterns (loadings)



**Scores = Loadings x data**

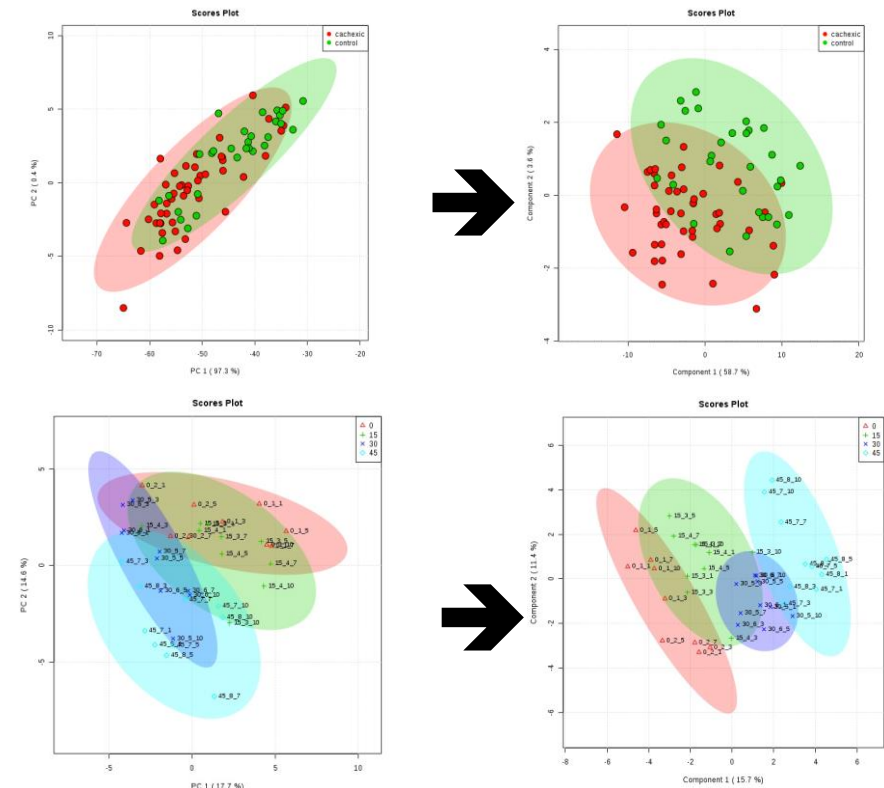
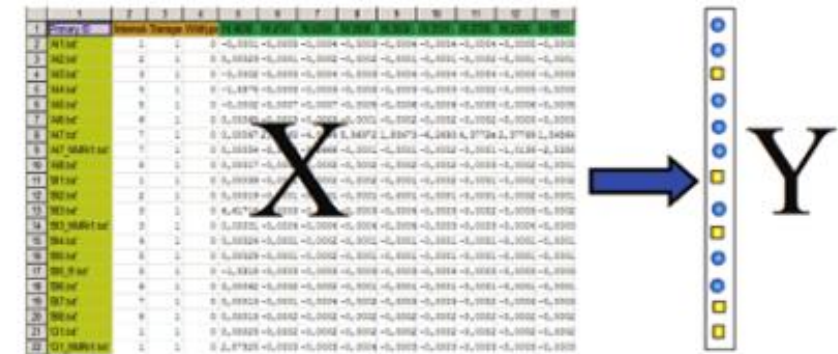
$$t_1 = p_1x_1 + p_2x_2 + p_3x_3 + \dots + p_nx_n$$

.....



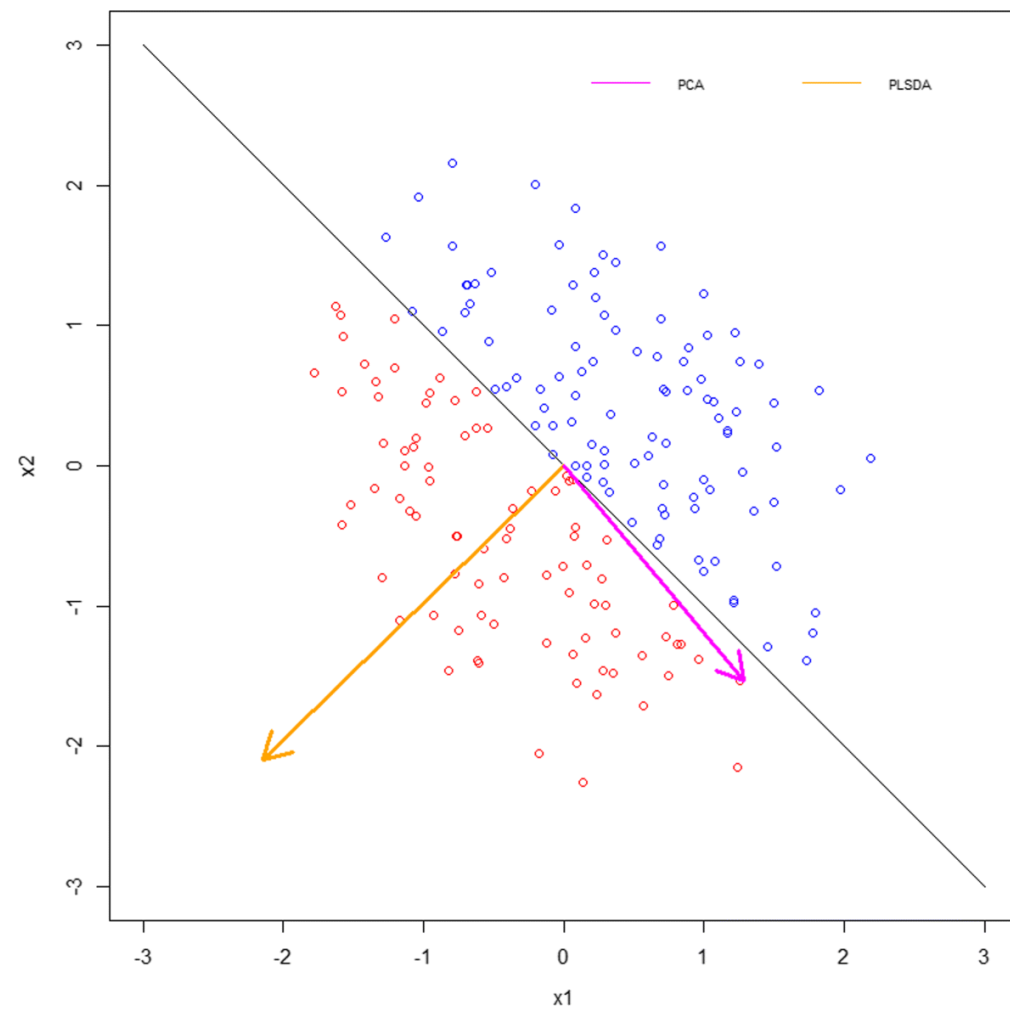
# Partial least squares-discriminant analysis (PLS-DA)

- When the experimental effects are subtle or moderate, PCA will not show good separation patterns
- 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)



PCA → PLS-DA

# Variance vs co-variance



# Max covariance may not explain max variance

- Variance explained by top components from PLS-DA
- In some cases, the 2<sup>nd</sup> component may explain more data variance (not covariance!) than the 1<sup>st</sup> one

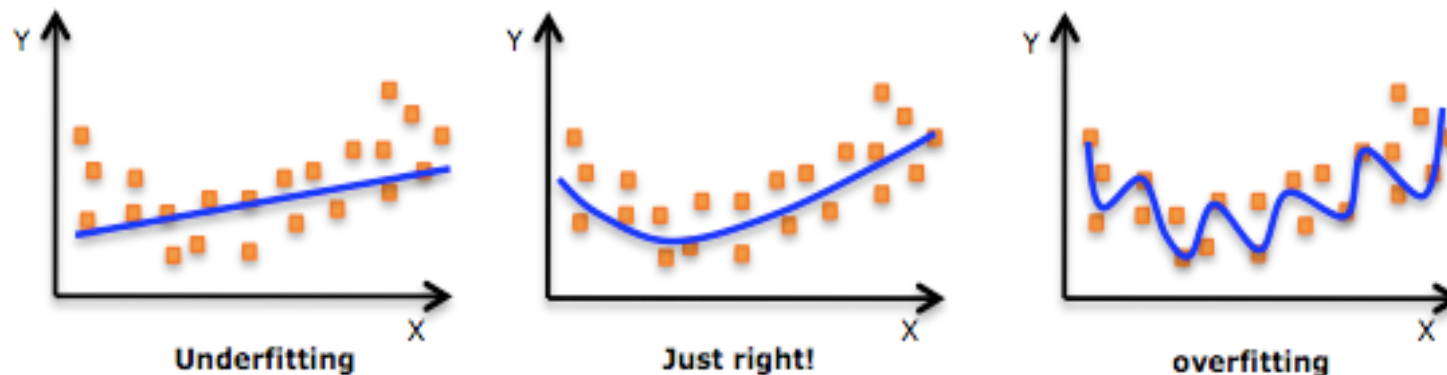


Explained variance by top 5 PLS-DA components

# Working with supervised approaches - overfitting

**Caution!** PLS-DA **always** produces certain separation patterns with regard the conditions even there is no real difference between them!

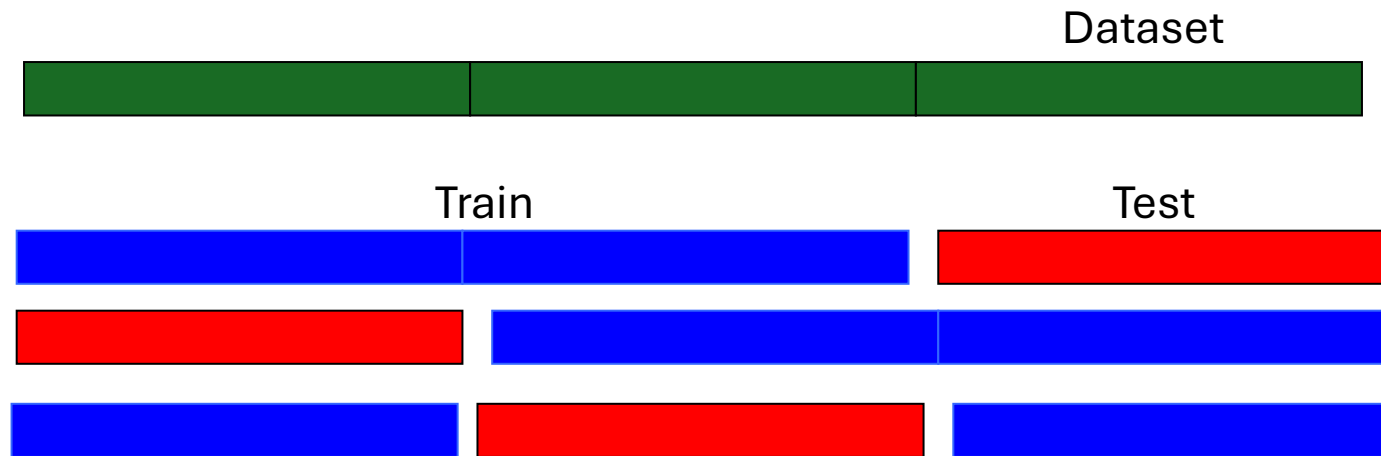
- Fitted model performs well for the current data
- Fitted model is not good for prediction of new data – prediction error is underestimated
- Model is too elaborate, models “noise” that will not be the same for new data



# Performance evaluation for PLS-DA

**Cross validation** – whether the model can predict on new events

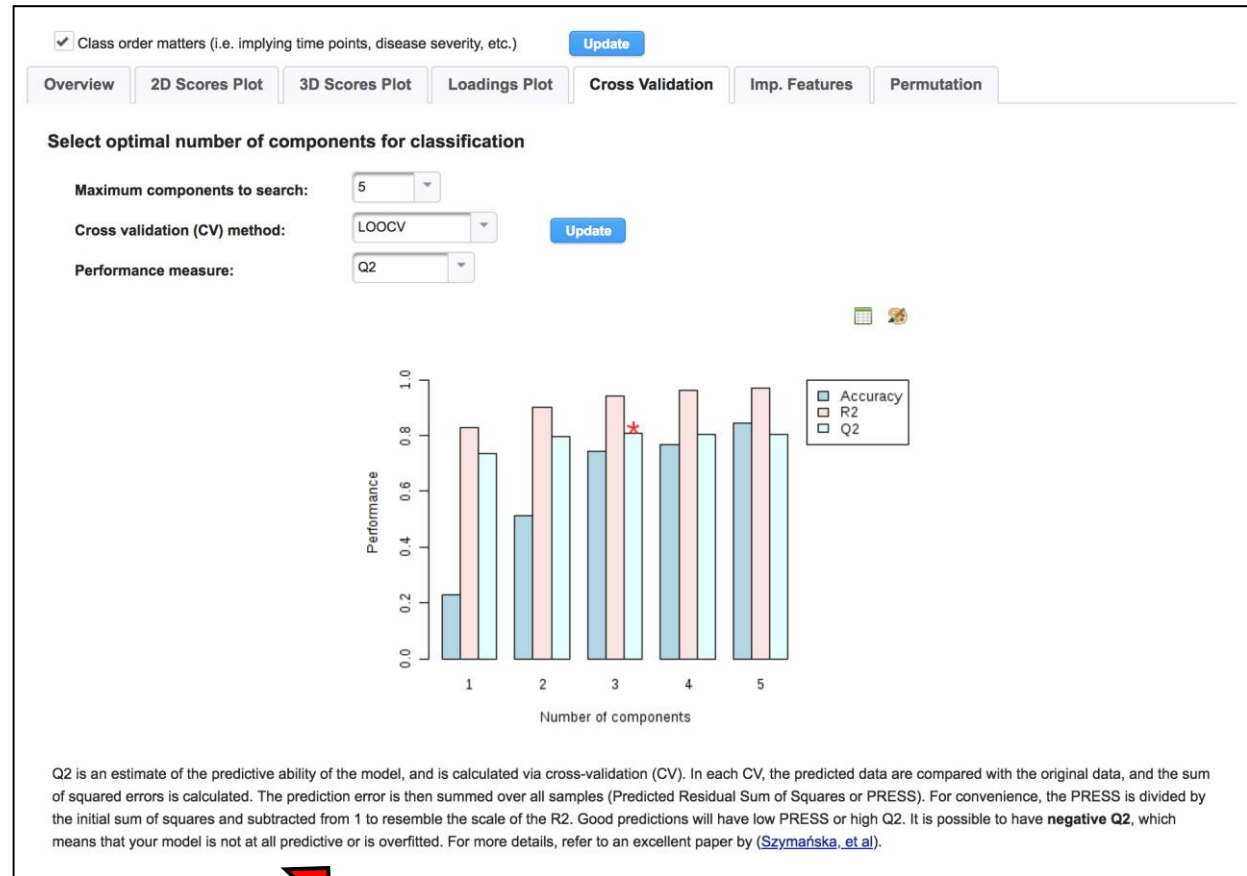
- Prediction accuracy
- Sum of squares captured by the model ( $R^2$ )
- Cross-validated  $R^2$  (also known as  $Q^2$ )



**Permutation tests** – whether the model captures real signals compared to the null models (those with group labels randomly assigned)

# Evaluation of PLS-DA Model

- PLS-DA model can be evaluated by cross validation,  $R^2$  and  $Q^2$
- Using too many components can over-fit
- 3 component model seems to be a good compromise here
- Good  $R^2/Q^2$  ( $>0.7$ )

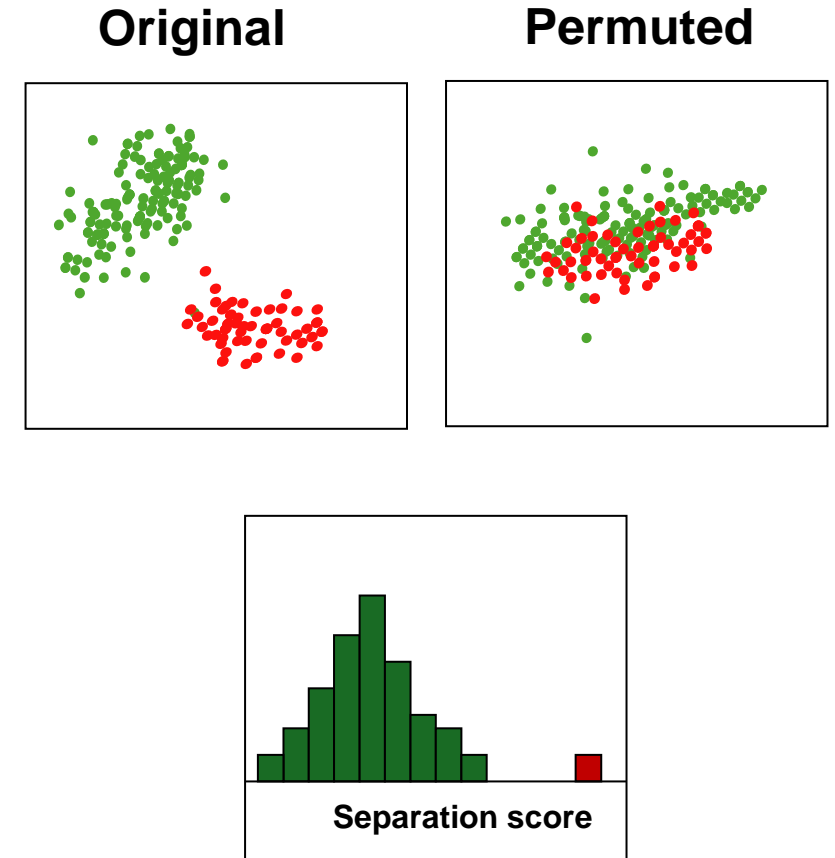


Can  $Q^2$  be negative?!

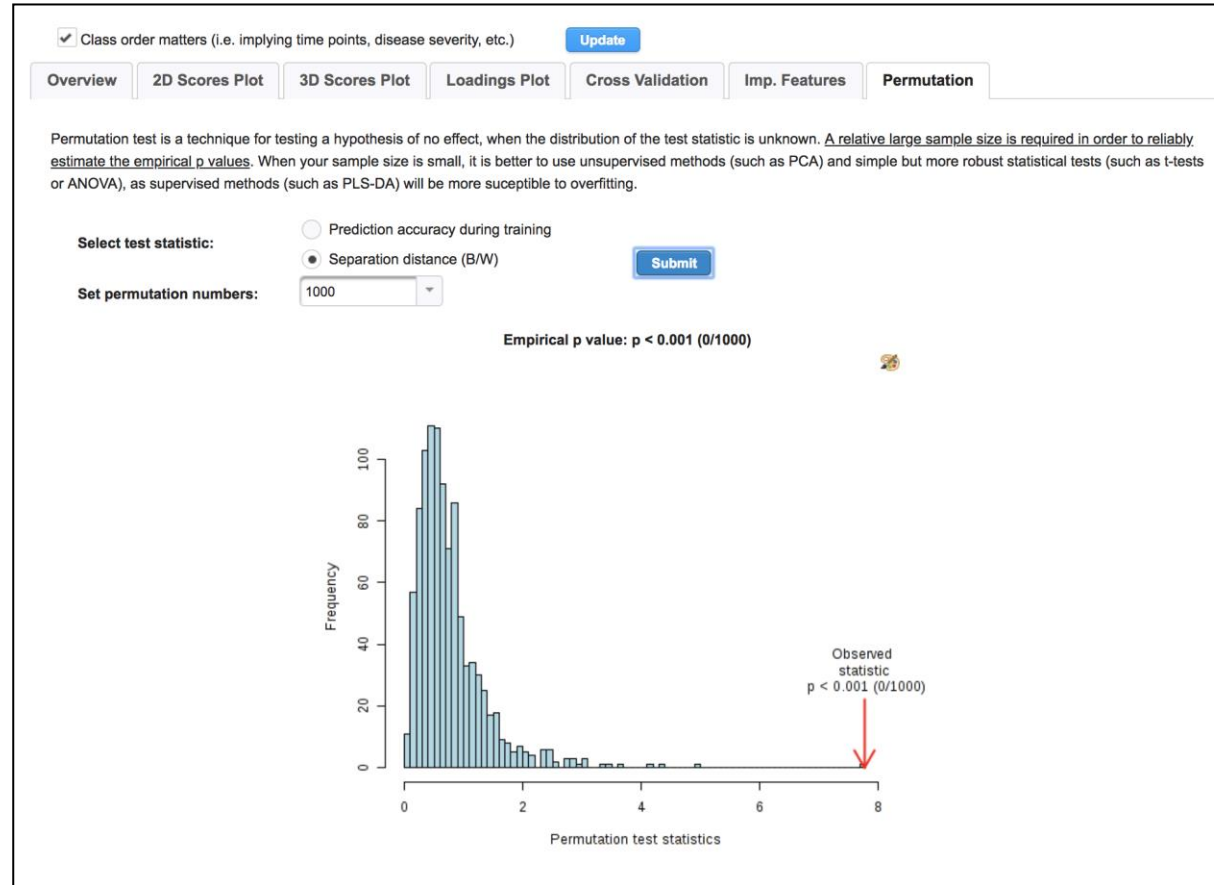
# Permutation Tests

To test whether the model is significantly different from the null models

1. Randomly shuffle the class labels ( $y$ ) and build the (null) model between new  $y$  and  $x$ ;
2. Test whether there is still the similar performance (i.e. distances in separation patterns);
3. We can compute empirical  $p$  values
  - If the result is similar as the permuted results (i.e. null model), then we can **not** say  $y$  and  $x$  is significantly correlated



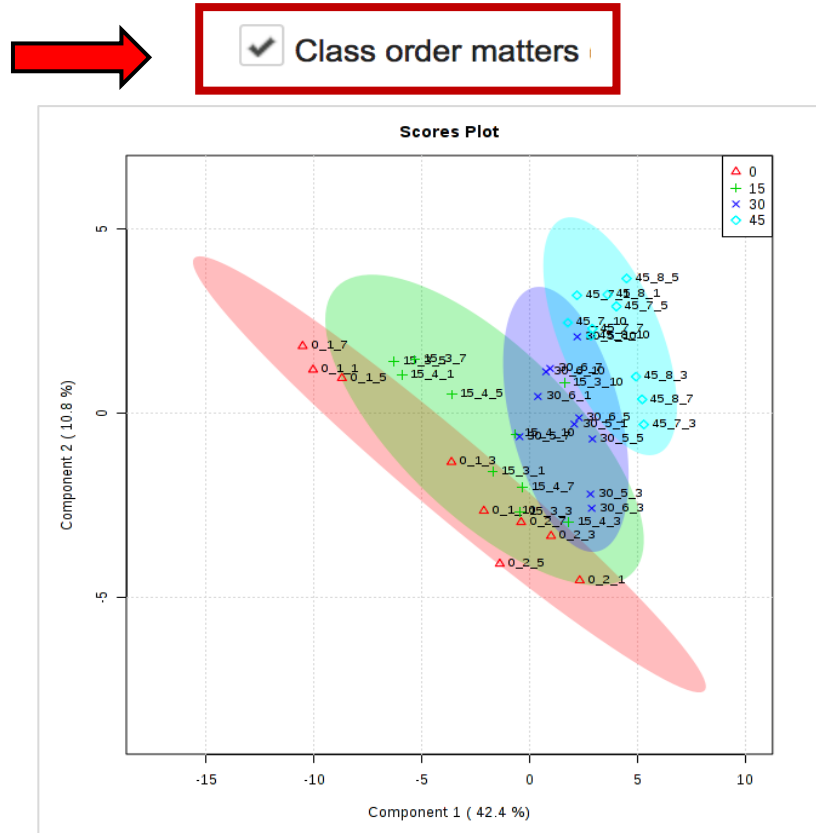
# Model validation by permutations



Permutation is computationally intensive. It is not performed by default. Users need to set the permutation number and press the submit button



# PLS-DA for multi-group: order matters



Different group names can potentially change separation patterns!



Different group names will not matter