



Batch Effect Management

Yiwen (Eva) Wang

Melbourne Integrative Genomics, University of Melbourne



WIFI

WIFI: Uniwireless

USER: florey-guest

PASSWORD: Neur@lc0nnex@x0N

Objectives of this workshop

- To recognise the importance of addressing batch effects for research reproducibility.
- To understand assumptions, applications and limitations of existing methods handling batch effects.
- To gain practical skills in managing batch effects and evaluating correction effectiveness.



Batch effects

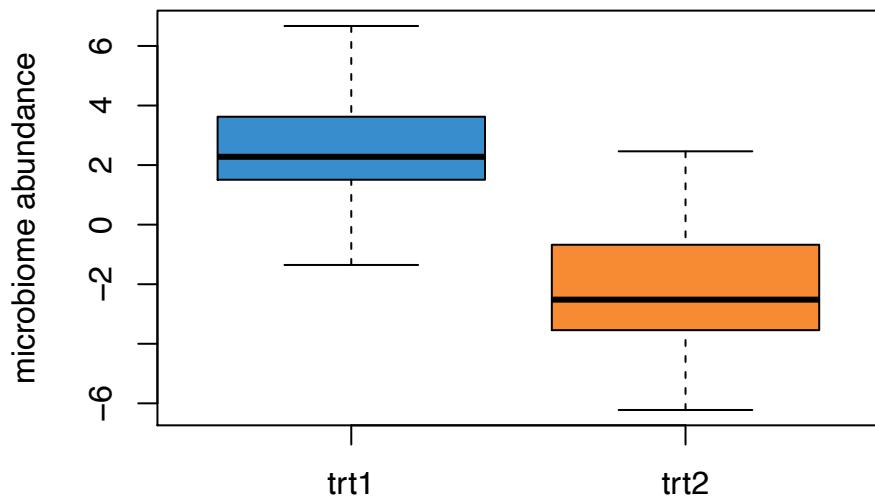
Definition: unwanted variation unrelated to but obscuring the biological factors of interest (e.g., treatments).

- Batch effects are associated with the outcome independently of treatments
- If batch effects correlate with treatment effects → confounders
- If batch effects don't correlate with treatment effects → prognostic variables
- Batches are usually categorical variables

Batch effects

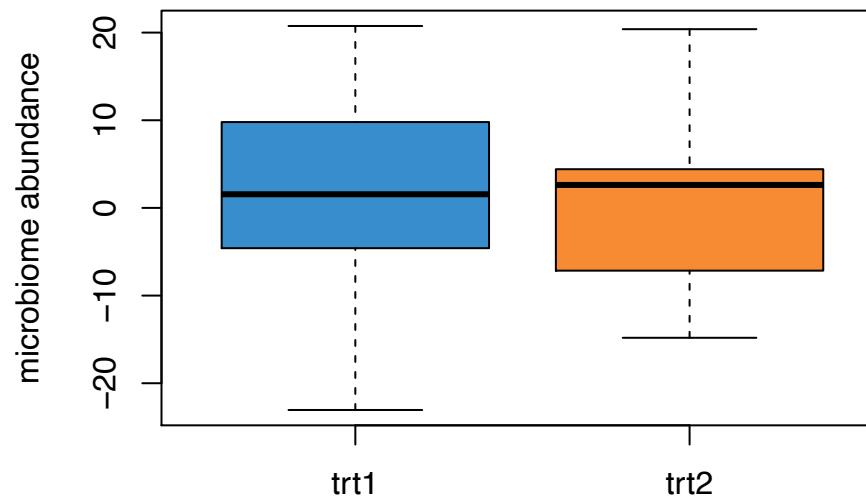
Consequence:

Data without batch effects



P < 0.001 of the treatment effect in T-test

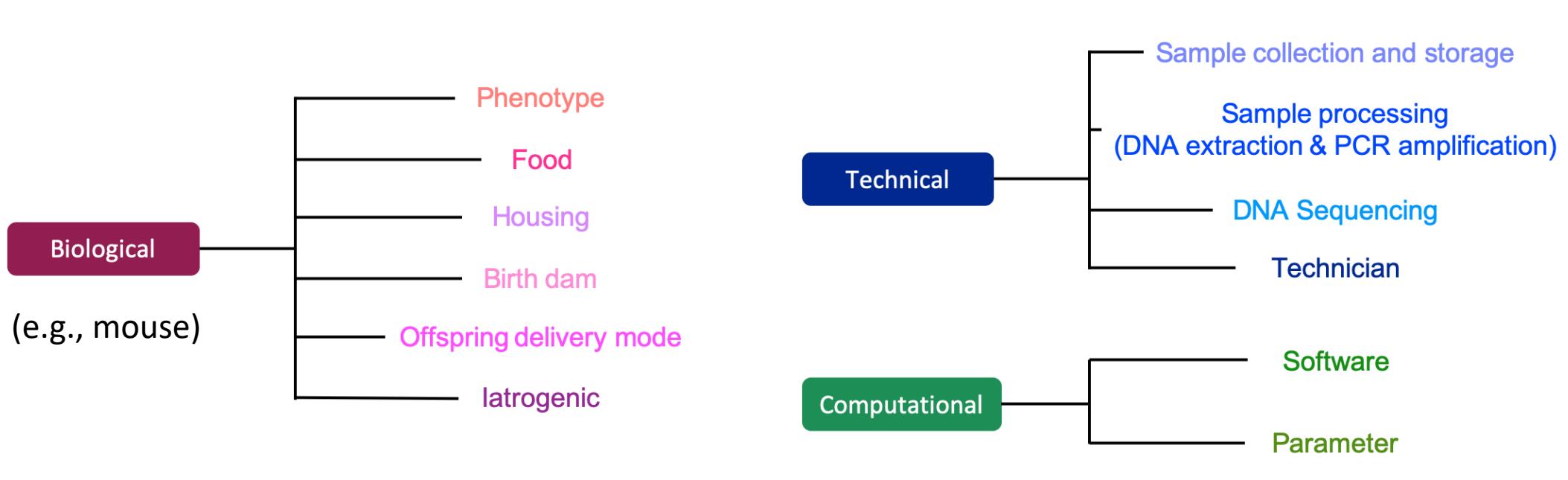
Data with batch effects



P > 0.05 of the treatment effect in T-test

Potential batch sources

Batch effects may happen in **any** step of experiments.



Assumptions about batch effects

Most statistical methods that **correct for** batch effects assume **balanced** batch x treatment designs, which means batch effects are independent of the effects of interest.

Balanced (prognostic)

	Treat 1	Treat 2
Batch 1	10	10
Batch 2	10	10

Unbalanced (confounding)

	Treat 1	Treat 2
Batch 1	4	16
Batch 2	16	4

Nested (confounding)

	Treat 1	Treat 2
Batch 1	10	0
Batch 2	0	10
Batch 3	10	0
Batch 4	0	10

Nested X

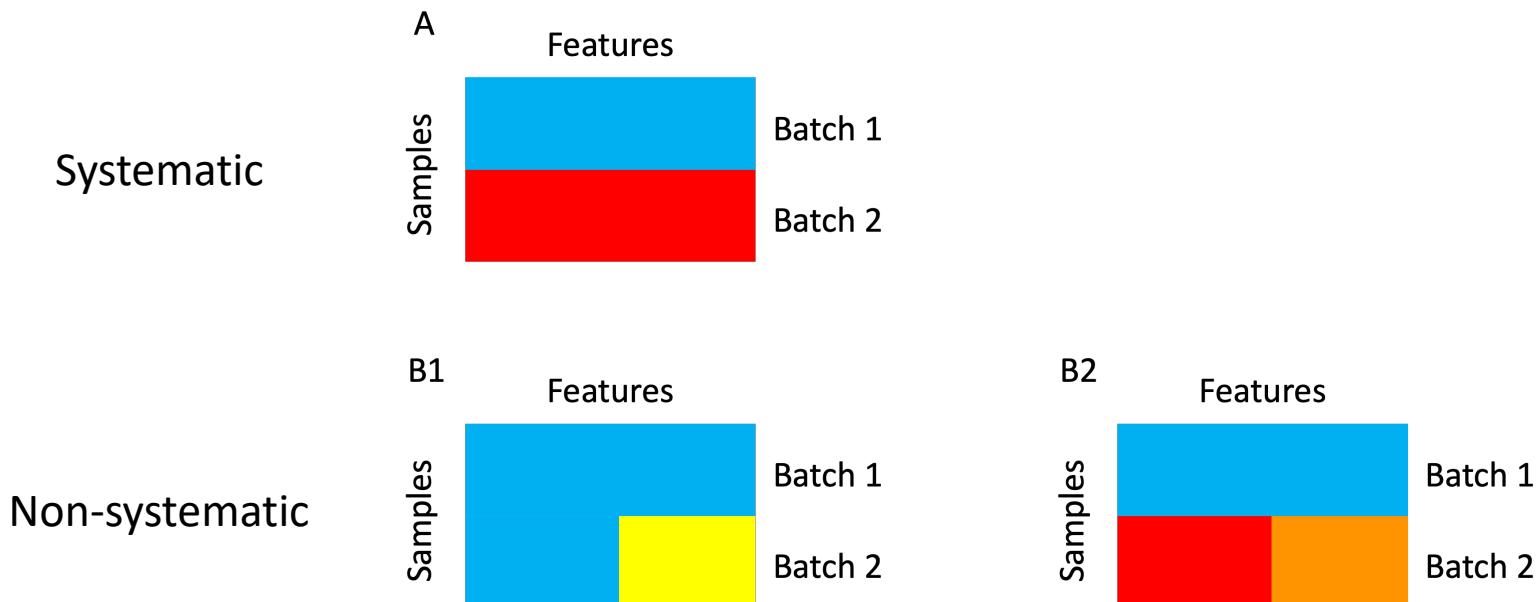
	Treat 1	Treat 2
Batch 1	0	20
Batch 2	20	0

Assumptions about batch effects

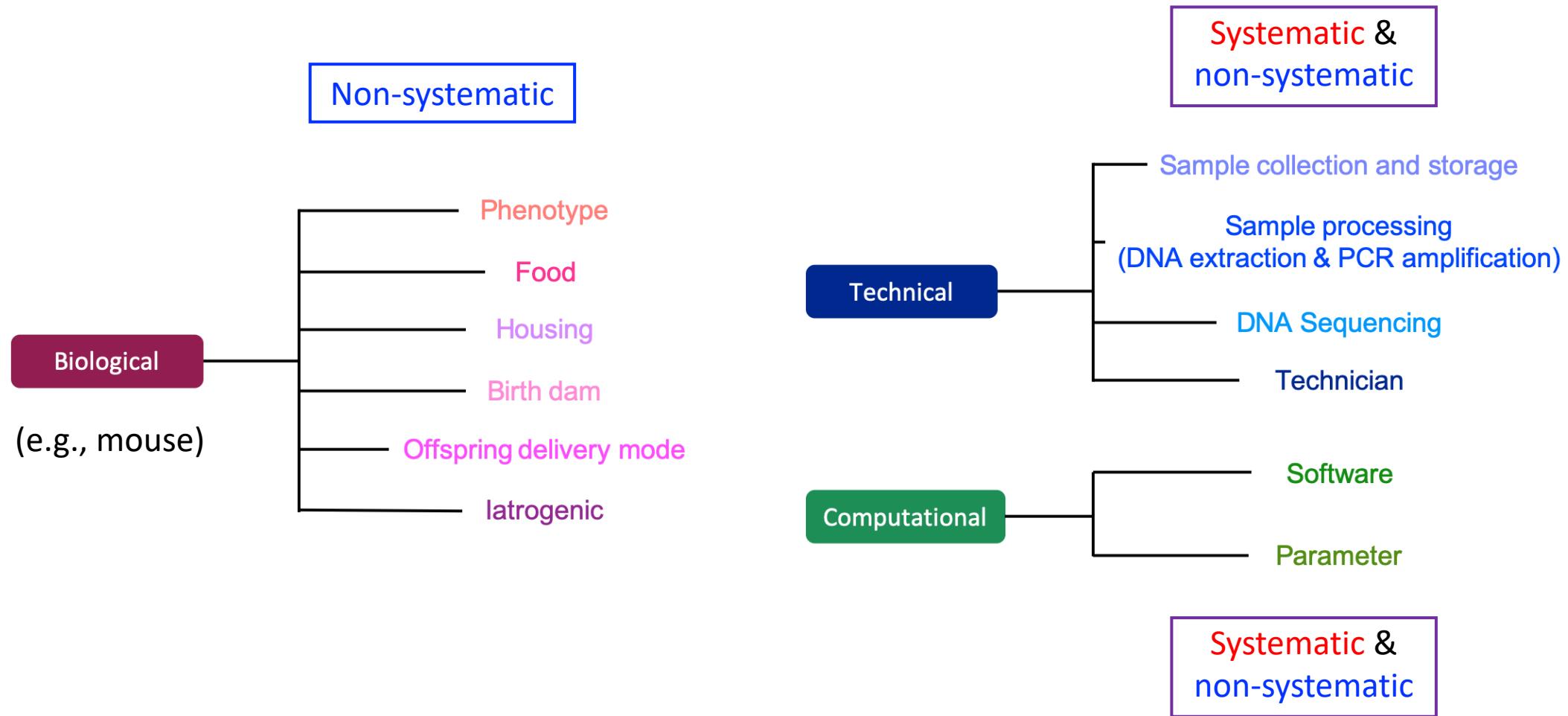
Batch effects have different scale of influence on variables:

- **Systematic** batch effects have a **homogeneous** influence on all variables, e.g., microbial growth rates follow the same distribution.
- **Non-systematic** batch effects have a **heterogeneous** influence on different variables.

Many methods assume **systematic** batch effects.



Assumptions about batch effects

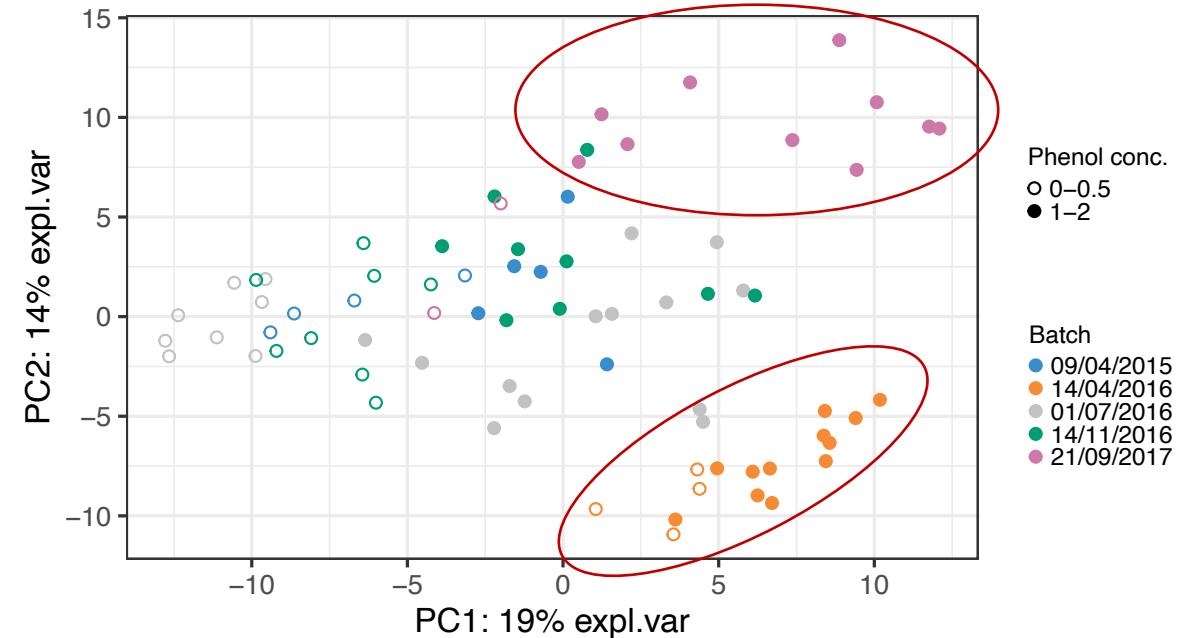


Case studies



Anaerobic Digestion (AD data): ★

- Bioreactor experiment: aimed at improving biowaste digestion
- 567 microbial variables & 75 samples
- Treatment effect: 2 levels of phenol concentrations
- (Technical) batch effect: samples processed on 5 different dates



Case studies



Sponge data:

- Investigating differences in microbial composition between specific sponge tissues
- 24 microbial variables & 32 samples
- Effect of interest: 2 different tissues
- (Technical) batch effect: sample processed on 2 different experimental gels
- Data characteristic: completely balanced design



	Tissue 1	Tissue 2
Batch 1	8	8
Batch 2	8	8

Case studies



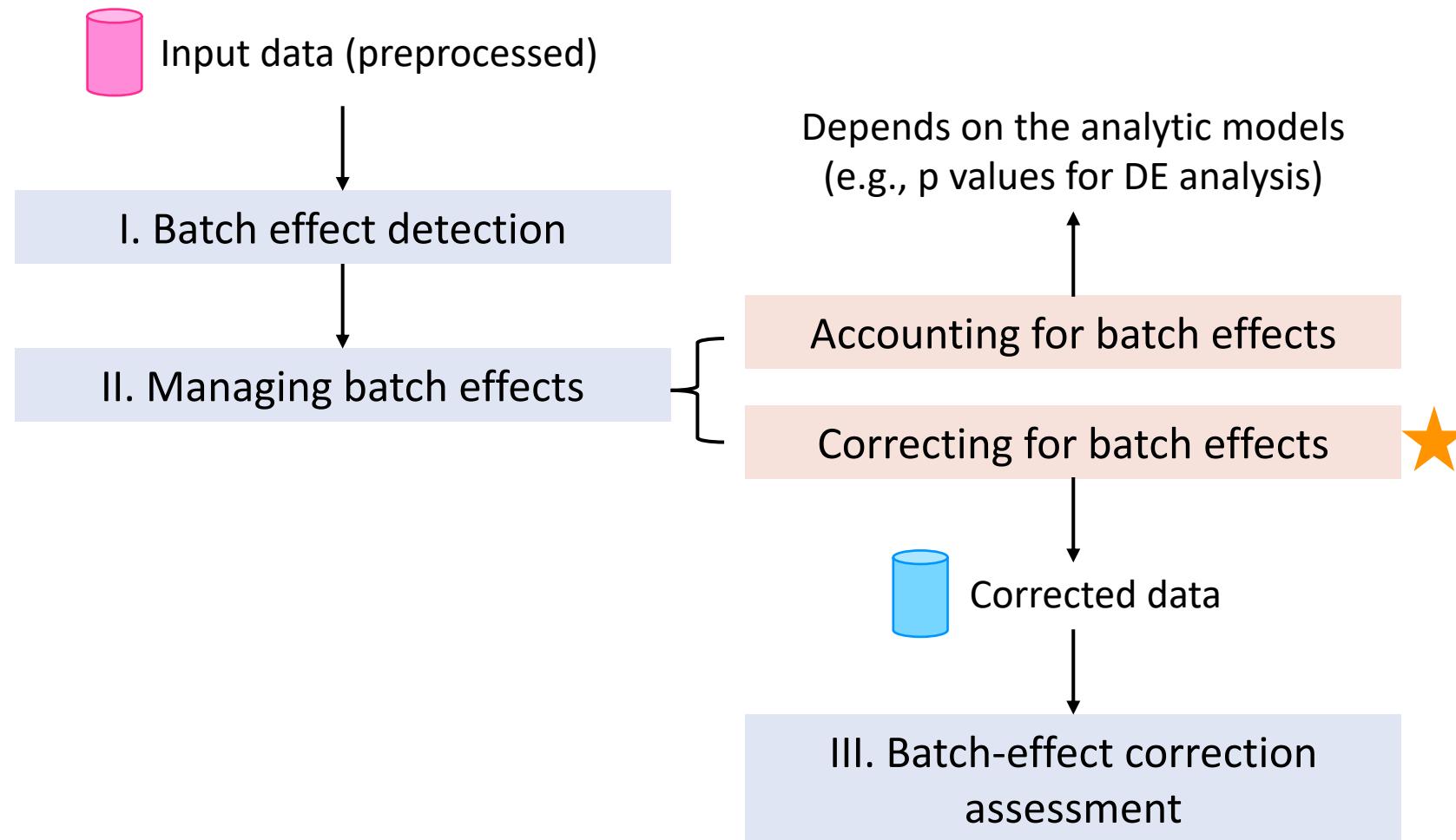
Mice models with Huntington's disease (HD data):

- Exploring differences in microbial composition between Huntington's disease and wild-type mice
- 368 microbial variables & 30 samples
- Effect of interest: 2 different genotypes
- (Biological) batch effect: samples from 10 different mouse cages



Cages\Genotypes	HD	WT
Cage A	2	0
Cage B	3	0
Cage C	2	0
Cage D	0	4
Cage E	0	4
Cage F	0	3
Cage G	3	0
Cage H	3	0
Cage I	2	0
Cage J	0	4

Workflow for batch effect management



I. Batch effect detection

Purpose: to detect batch effects and determine if the batch effect management is required.

A. Visual approaches: limited for very weak batch effects

- Principal component analysis (PCA)
- Boxplots and density plots
- Heatmap

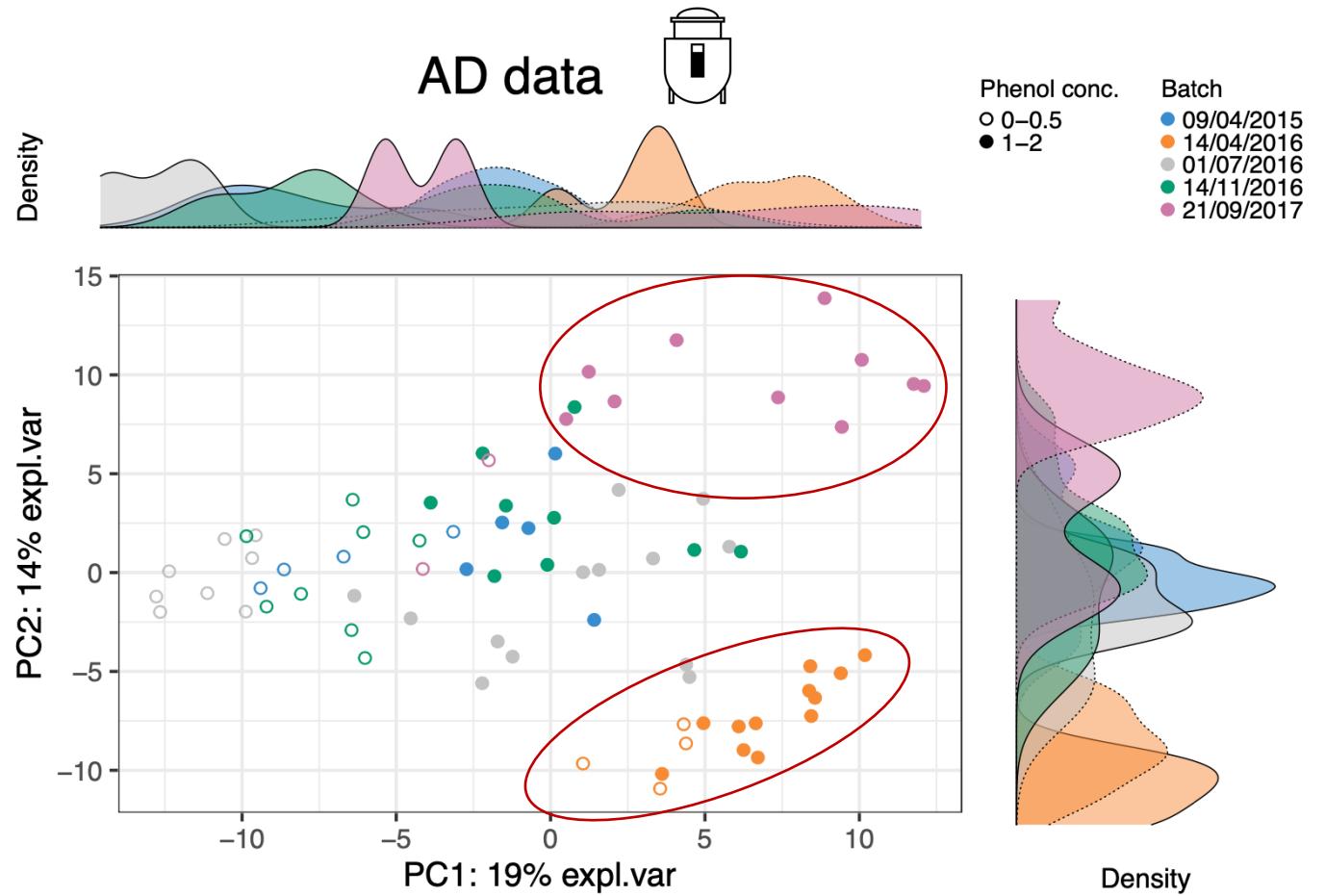
B. Quantitative methods: very sensitive to batch effects

- Partial redundancy analysis (pRDA)

I. Batch effect detection

PCA plots with density per component

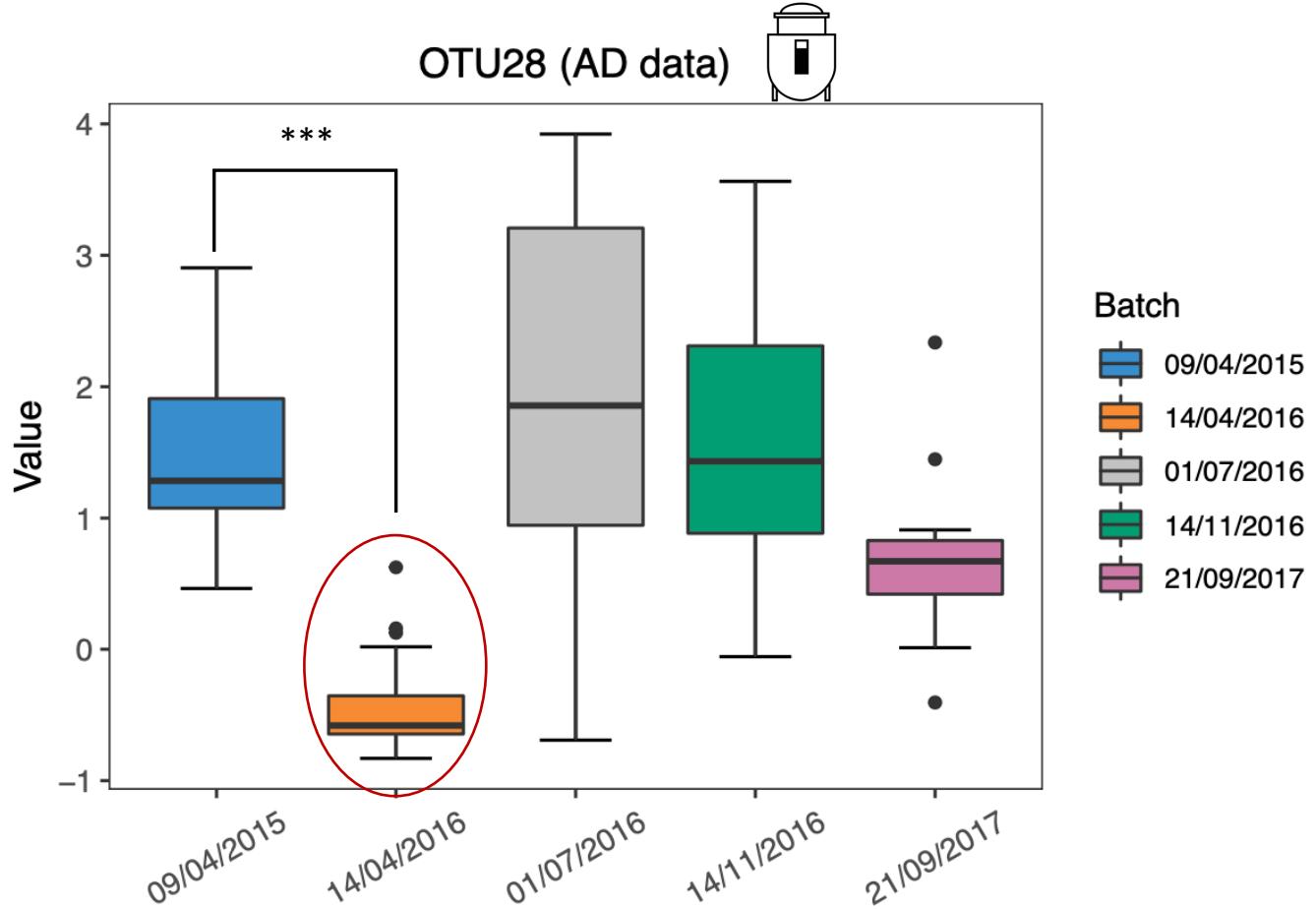
=> Multivariate: combination of all taxa



I. Batch effect detection

Boxplots

=> Univariate: single taxon

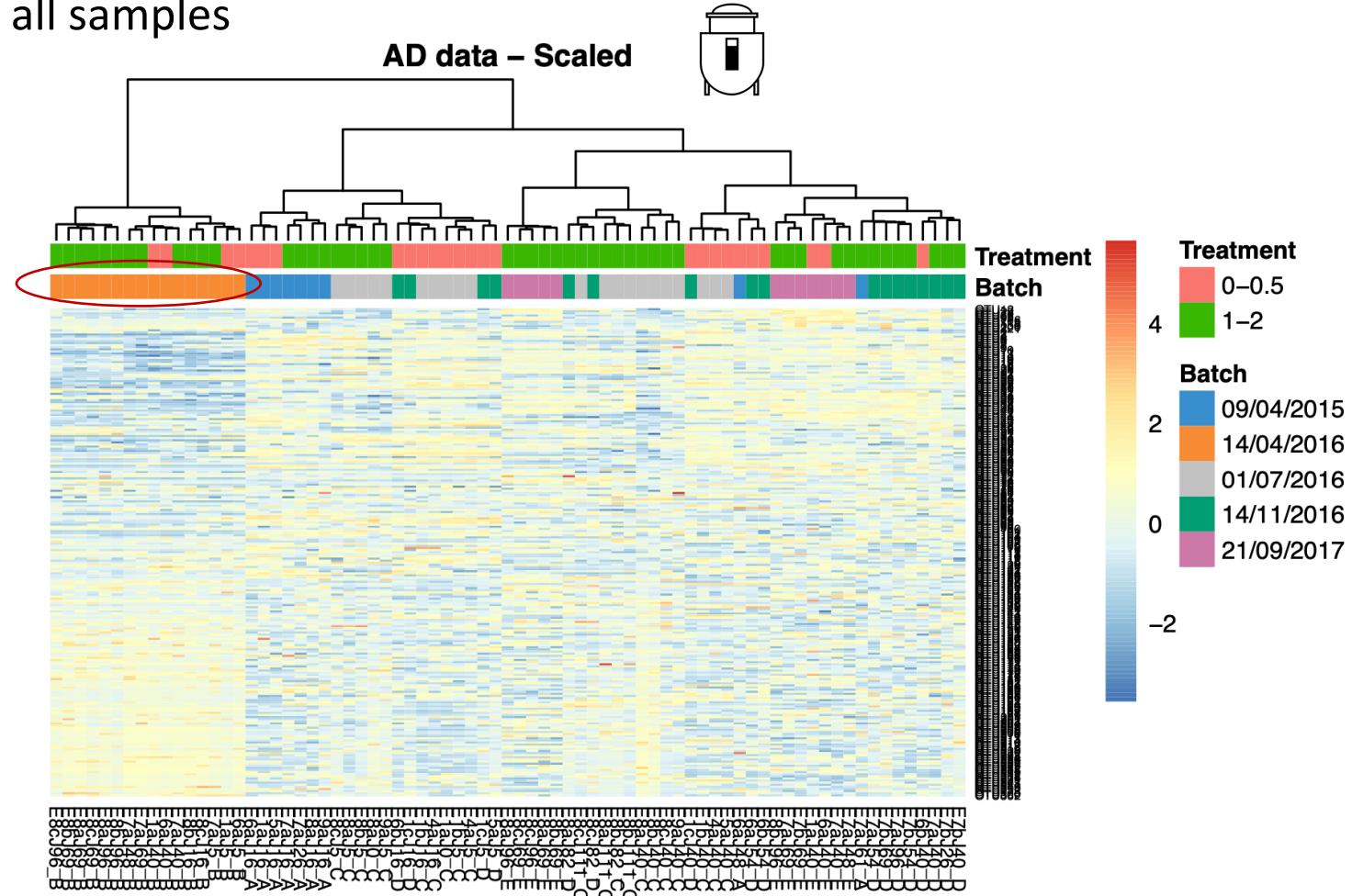


Test for the difference between batches (here P-value < 0.001, t-test).

I. Batch effect detection

Heatmap

=> All taxa and all samples



Samples from the
batch dated
"14/04/2016"
clustered together.

I. Batch effect detection

Purpose: to detect batch effects and determine if the batch effect management is required.

A. Visual approaches: limited for very weak batch effects

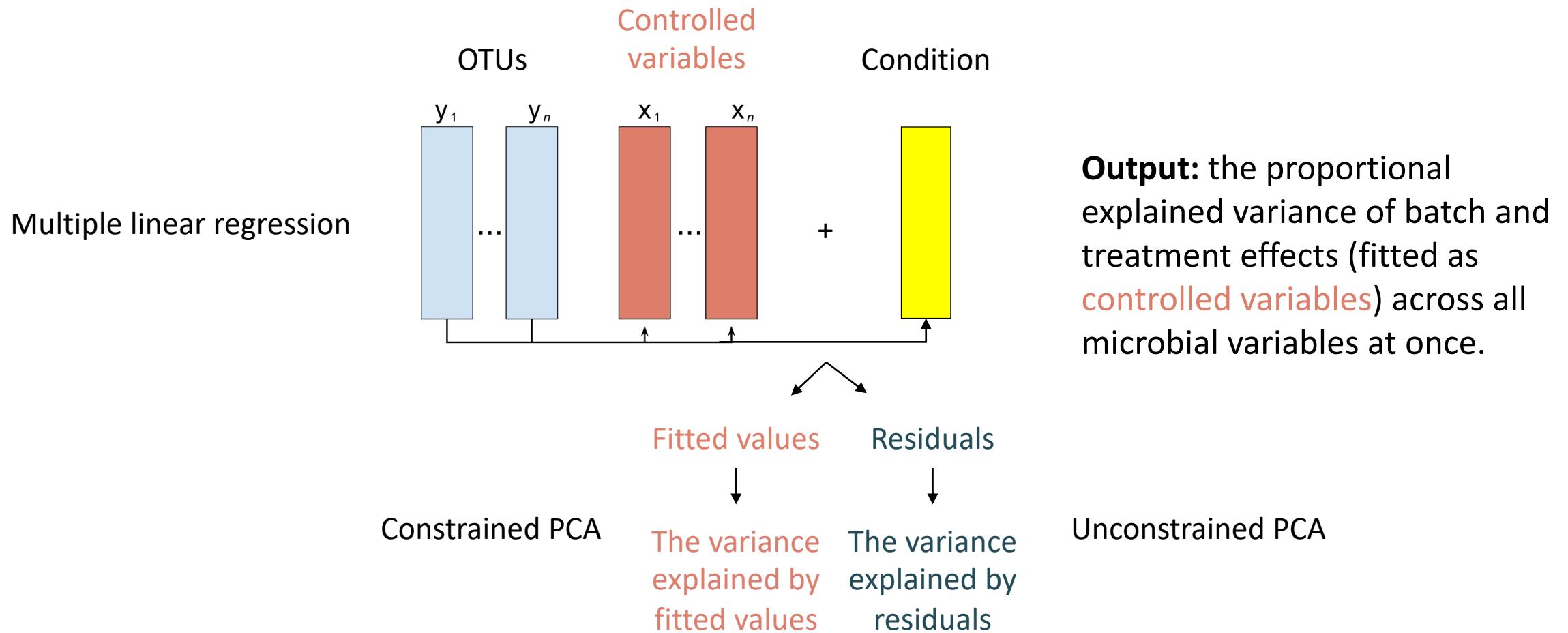
- Principal component analysis (PCA)
- Boxplots and density plots
- Heatmap

B. Quantitative methods: very sensitive to batch effects

- Partial redundancy analysis (pRDA)

I. Batch effect detection

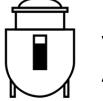
Partial redundancy analysis (pRDA): multivariate



Output: the proportional explained variance of batch and treatment effects (fitted as controlled variables) across all microbial variables at once.

I. Batch effect detection

pRDA: can indicate if batch x treatment design is balanced

Approx. balanced batch x treatment design (AD data )

Dates\Phenol conc.	0-0.5	1-2
09/04/2015	4	5
14/04/2016	4	12
01/07/2016	8	13
14/11/2016	8	9
21/09/2017	2	10

The intersection variance indicates how unbalanced the design is:

=> batch and treatment effects are correlated.

```
##                                     Df R.squared Adj.R.squared Testable
Treat only   = Treat | Batch     1    NA      0.08943682    TRUE
Intersection          0    NA      0.01296248    FALSE
Batch only    = Batch | Treat    4    NA      0.26604420    TRUE
## [d] = Residuals NA           NA      0.63155651    FALSE
```

I. Batch effect detection

pRDA:

Completely balanced batch x treatment design (**sponge data**)



	Tissue 1	Tissue 2
Batch 1	8	8
Batch 2	8	8

```
##                                     Df R.squared Adj.R.squared Testable
Treat only = Treat | Batch     1      NA    0.16572246    TRUE
Intersection                      0      NA   -0.01063501   FALSE
Batch only  = Batch | Treat    1      NA    0.16396277    TRUE
## [d] = Residuals   NA        NA    0.68094977   FALSE
```

No intersection variance:
=> batch and treatment effects are independent.

I. Batch effect detection

pRDA:

Nested batch x treatment design (HD data)



Cages\Genotypes	HD	WT	##	Df	R.squared	Adj.R.squared	Testable
Cage A	2	0	Treat only = Treat Batch	0	NA	-2.220446e-16	FALSE
Cage B	3	0	Intersection	0	NA	9.730583e-02	FALSE
Cage C	2	0	Batch only = Batch Treat	8	NA	1.608205e-01	TRUE
Cage D	0	4	## [d] = Residuals	NA	NA	7.418737e-01	FALSE
Cage E	0	4					
Cage F	0	3					
Cage G	3	0					
Cage H	3	0					
Cage I	2	0					
Cage J	0	4					

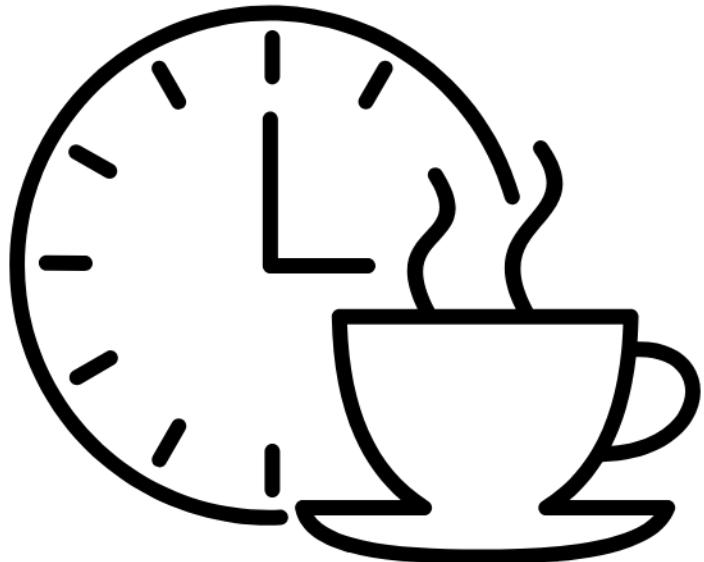
No treatment variance & considerable intersection variance:

=> batch and treatment effects are collinear.

Your turn!

Practice detecting batch effects in the AD data by following the steps in the "Batch effect detection" section. (30 mins)

Break



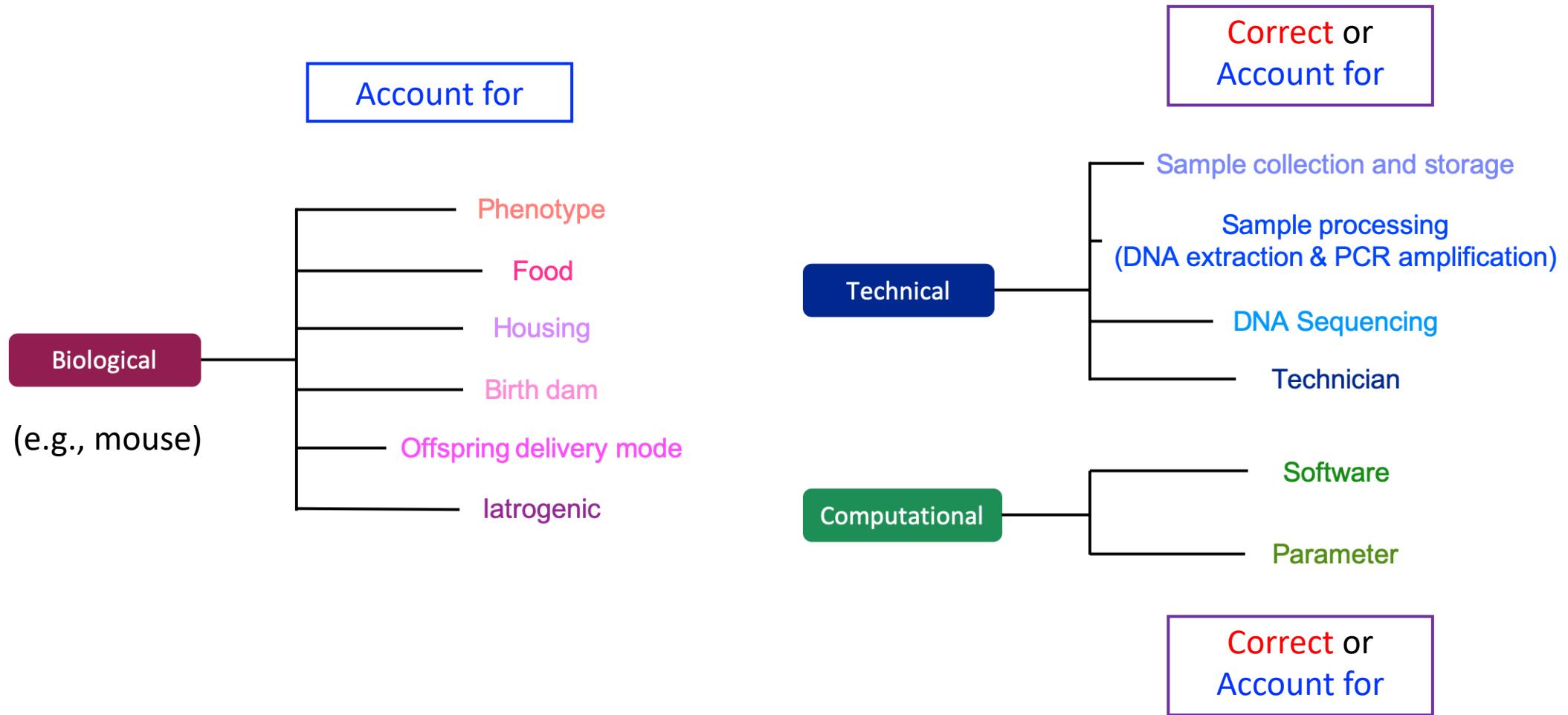
~ 15 mins

Created by Mukholifah
from Noun Project

II. Managing batch effects

- Accounting for batch effects:
 - include batch effects as covariates in statistical models
- Correcting for batch effects:
 - remove batch effects from the original data
- Univariate vs. multivariate:
 - Univariate: process each variable **individually**.
e.g., Differential expression analysis. in *DESeq2*, *edgeR*
 - Multivariate: process all variables **together**.
e.g., PCA, CCA, etc. in *phyloseq*
- **Multivariate** methods allow to consider microbial variables as multivariate, rather than independent.

III. Managing batch effects



III. Managing batch effects

Methods accounting for batch effects:

- Pros: can consider the [data characteristics](#) and [correlation](#) between batch and treatment effects.
- Cons:
 - Limited to specific analyses according to the model (e.g., [differential abundance analysis](#): taxa with p values)
 - Difficult to be assessed explicitly

e.g., Linear regression

III. Managing batch effects

Methods accounting for batch effects:

A. Designed for microbiome data (applied to count data):

- Cumulative-Sum Scaling normalisation + Zero-inflated Gaussian mixture model (CSS+ZIG):
 - Differential abundance analysis
 - Handle data characteristics, including uneven library sizes, compositional structure and undersampling zeroes

B. Adapted for microbiome data (after CLR transformation):

- ★ • Linear regression: handle nested batch x treatment design (**HD data** )
- Surrogate variable analysis (SVA): estimate unknown batch effects without extra information
- ★ • Remove unwanted variation in 4 steps (RUV4): estimate unknown batch effects but require negative control variables or sample replicates that capture the batch variation

III. Managing batch effects

Methods correcting for batch effects (main focus of this workshop):

- Applied to CLR transformed data
 - Pros: corrected data can be input in any downstream analyses
 - Dimension reduction; Visualisation; Clustering; Variable selection
 - Cons:
 - cannot account for specific data characteristics within models; these need to be addressed in advance, e.g., CLR transformation
 - cannot handle correlations between batch and treatment effects within models; require additional processing to consider these correlations
- e.g., ComBat

III. Managing batch effects

Methods correcting for batch effects:

- removeBatchEffect (rBE):
 - Linear regression (removeBatchEffect(), *limma*)
 - [Univariate](#)
- ★ • ComBat:
 - Empirical Bayesian method
 - Assumes all variables are affected by batch effects in a [systematic](#) manner
 - [Mixture of univariate and multivariate](#)
- Percentile normalization (PN):
 - Each case sample's feature values are converted into percentiles of the control distribution within each batch
 - Require sufficient control samples within each batch
 - [Univariate](#)

III. Managing batch effects

Methods correcting for batch effects:

- ★ • Remove Unwanted Variation-III (RUVIII):
 - Requires **negative control variables** and **sample replicates** that capture the batch variation
 - **Multivariate**

- ★ • PLSDA-batch:
 - **Non-parametric**: can handle non-Gaussian distributions
 - **No assumption of systematic** batch effects
 - Include two variants: sparse PLSDA-batch (avoid overfitting); weighted PLSDA-batch (for unbalanced batch x treatment design)
 - **Multivariate**

Your turn!

Practice accounting for and correcting batch effects in the AD data by following the steps in the "Managing batch effects" section. (40 mins)

IV. Assessing batch effect correction

➤ Methods that detect batch effects:

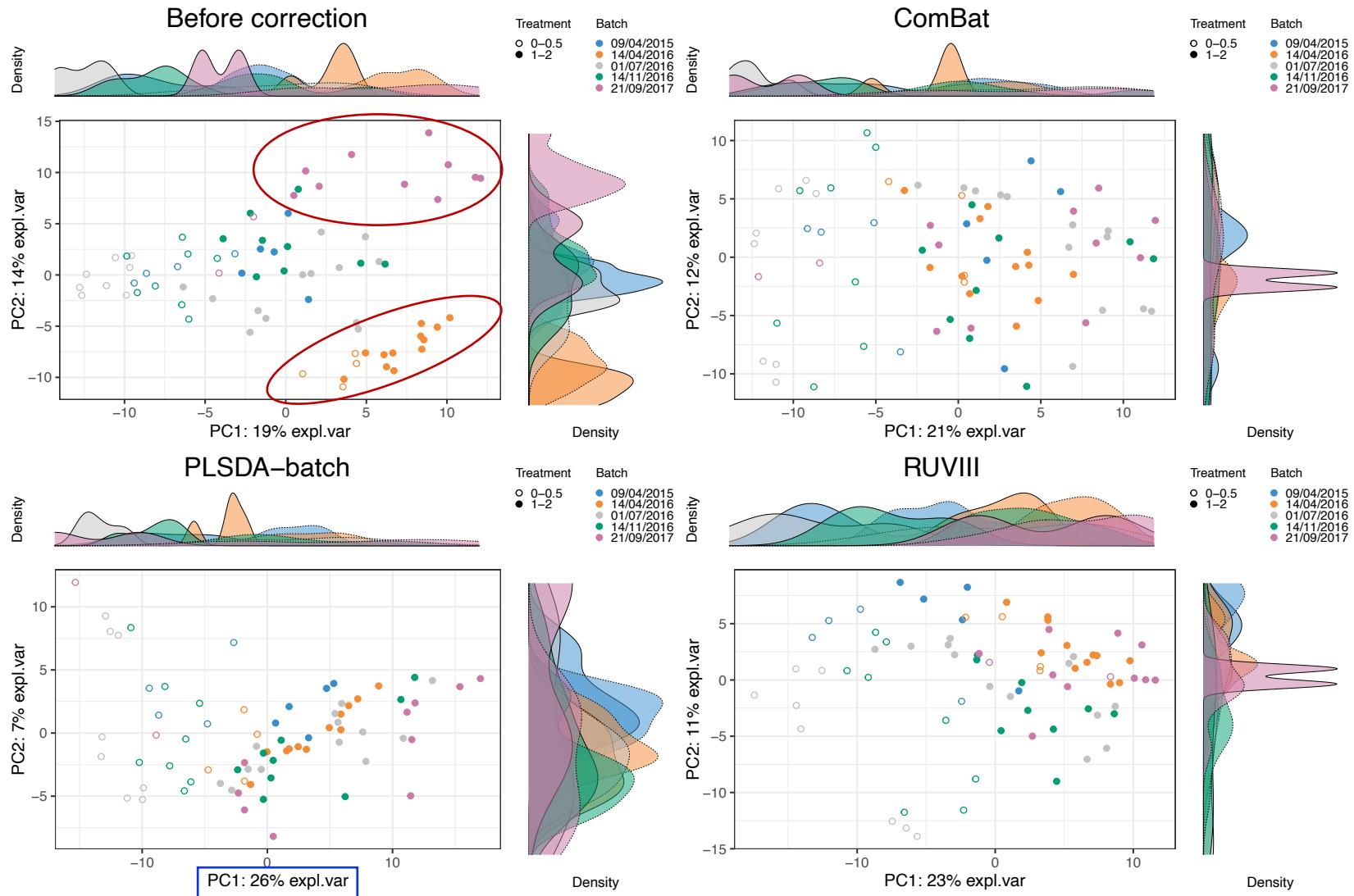
- Visualisation: PCA, boxplots, density plots, heatmap
- pRDA: proportion of explained variance across all variables

➤ Other methods:

- R^2 from one-way ANOVA: proportion of explained variance for each variable
- Alignment scores: [0,1], poor to excellent mixing samples among the different batches

IV. Assessing batch effect correction

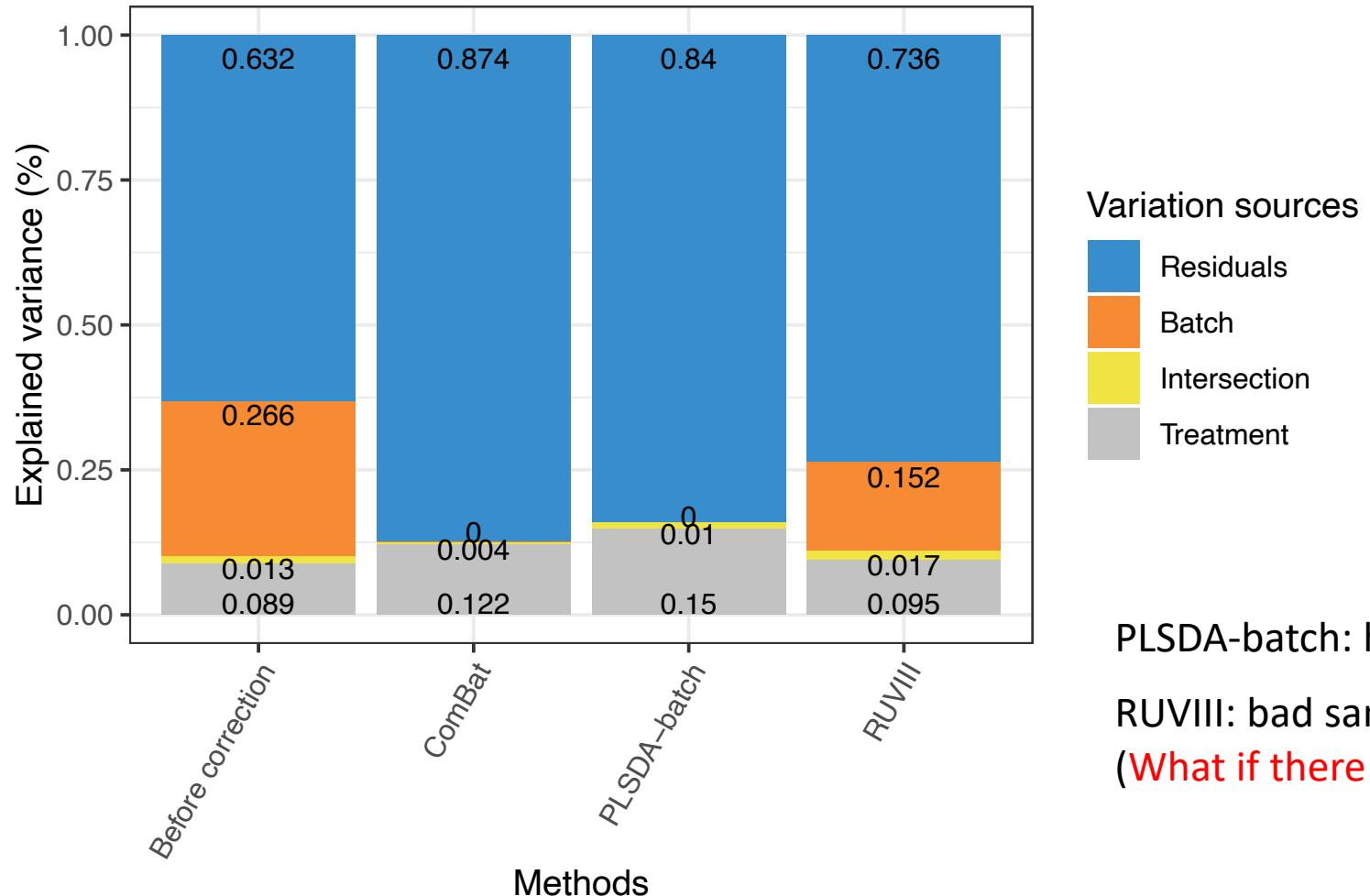
PCA:



Batch effect variation
is removed.

IV. Assessing batch effect correction

pRDA:



PLSDA-batch: higher treatment variance

RUVIII: bad sample replicates

(What if there were better replicates?)

IV. Assessing batch effect correction

➤ Methods that detect batch effects:

- Visualisation: PCA, boxplots, density plots, heatmap
- pRDA: proportion of explained variance across all variables

➤ Other methods:

- R^2 from one-way ANOVA: proportion of **explained variance** for **each variable**
- Alignment scores: [0,1], **poor** to **excellent** mixing samples among the different batches

IV. Assessing batch effect correction

Based on the sample dissimilarity matrix calculated from the PCA projection:

$$\text{Alignment scores} = 1 - \frac{\bar{x} - \frac{k}{n}}{k - \frac{n}{\bar{x}}},$$

k : the number of nearest neighbours

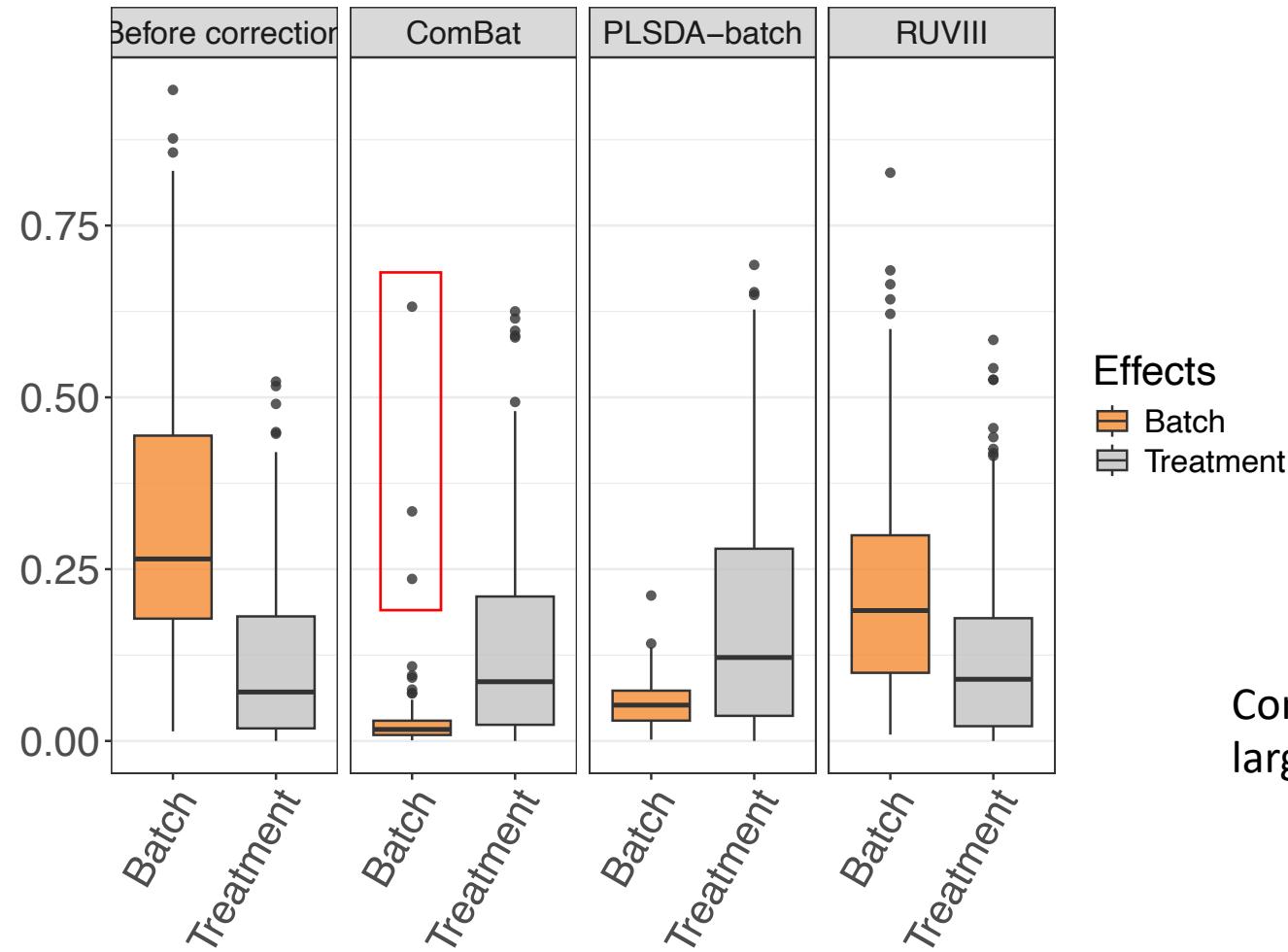
n : the sample size

x : the number of each sample's nearest neighbours that belong to the same batch

\bar{x} : the average of all x

IV. Assessing batch effect correction

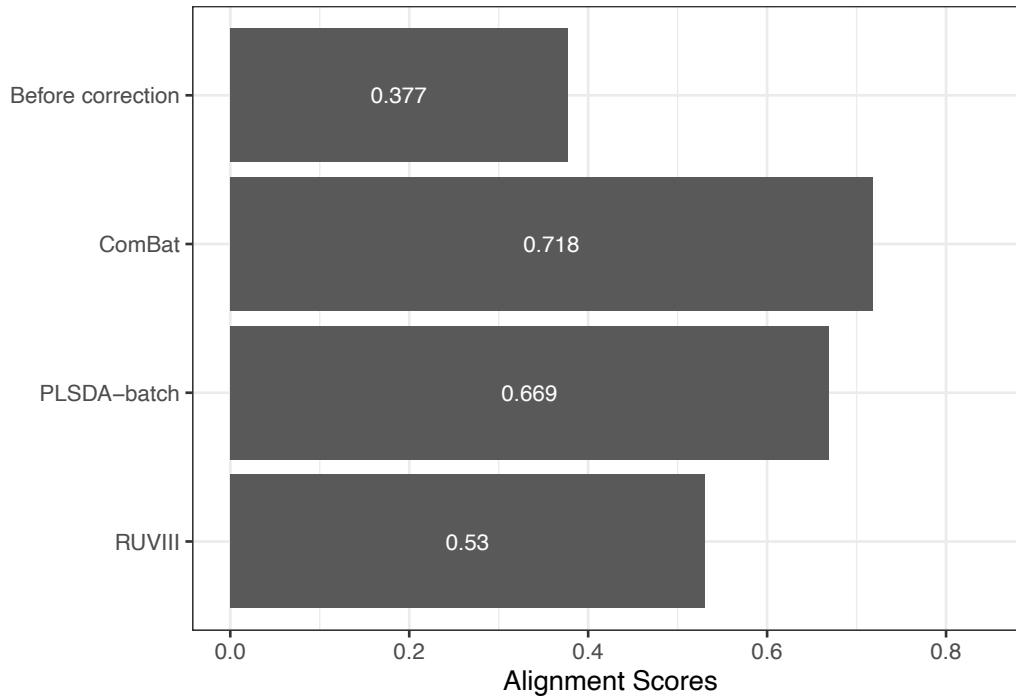
R^2 from one-way ANOVA:



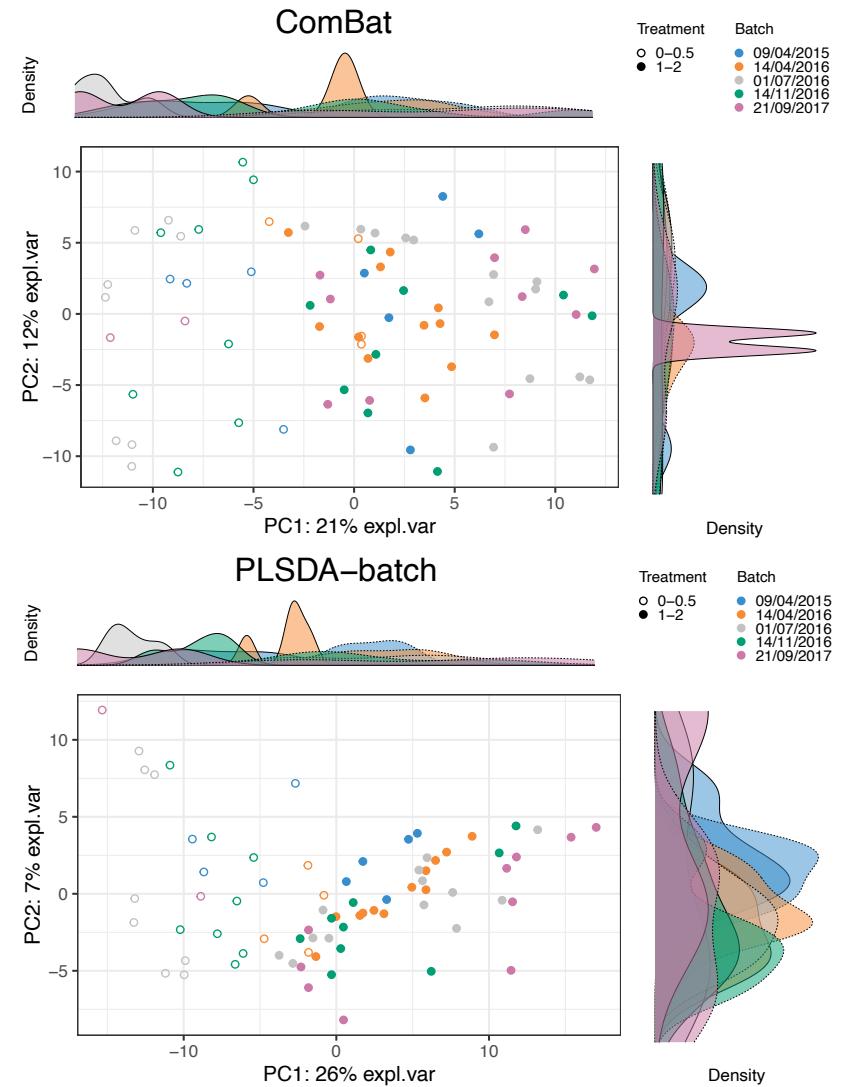
ComBat: a few variables with a large batch variance

IV. Assessing batch effect correction

Alignment scores:



ComBat vs. PLSDA-batch :
Better mixing of batches?
→ Greater variance in PCA projection



Your turn!

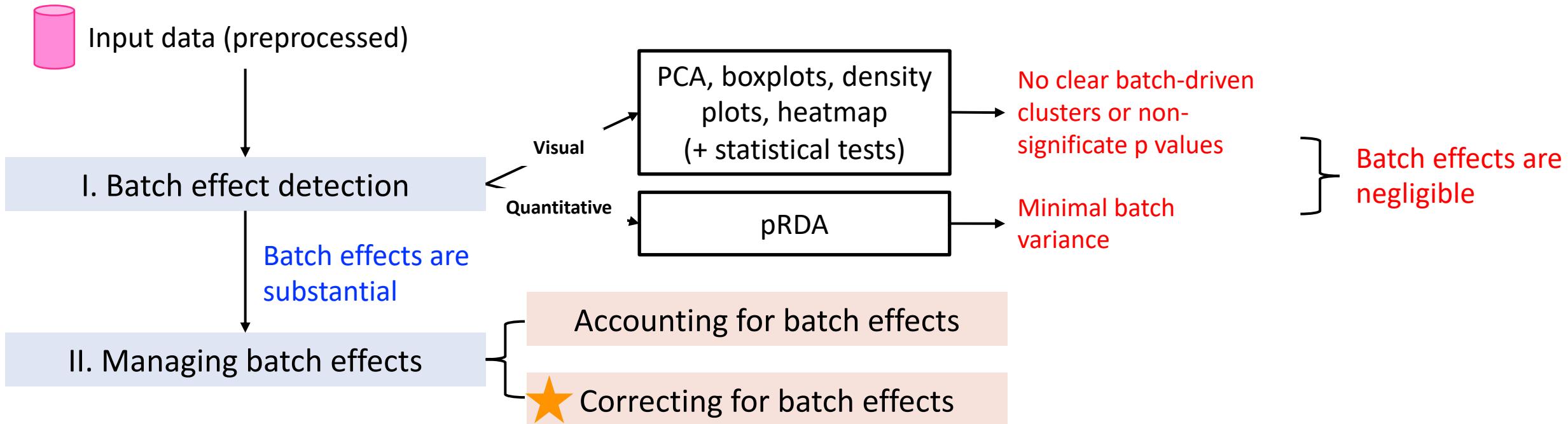
Practice using the AD data by following the steps in the "Assessing batch effect correction" section. (30 mins)

Conclusions

- Batch effects are ubiquitous, can arise from **biological, technical** and **computational** sources, and are sometimes unavoidable.

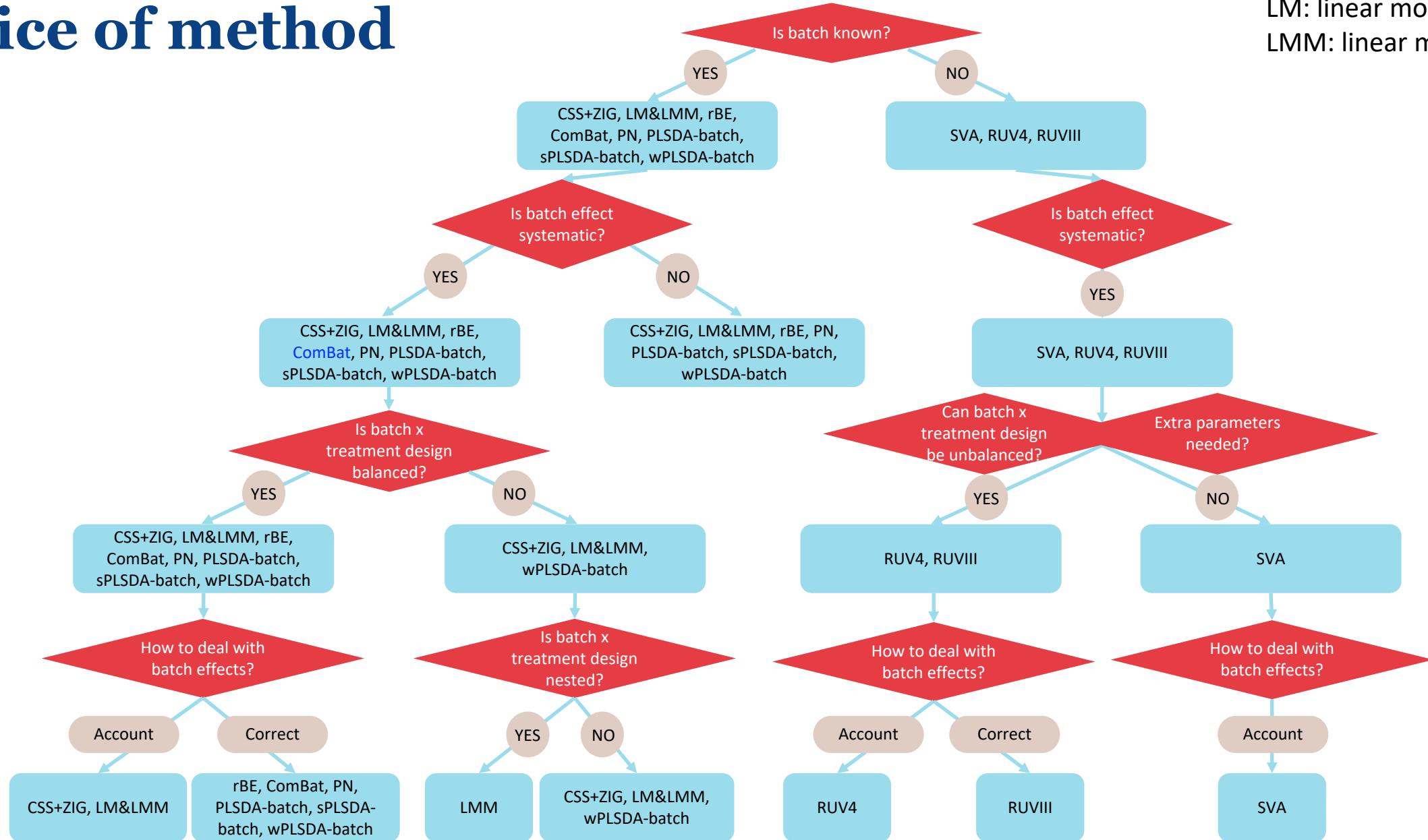
- Managing batch effects should consider corresponding data characteristics (**preprocessing beforehand or inclusion in the model**), batch sources (**account for or correct**), batch x treatment designs and the scale of influence (**method assumptions**).

Conclusions

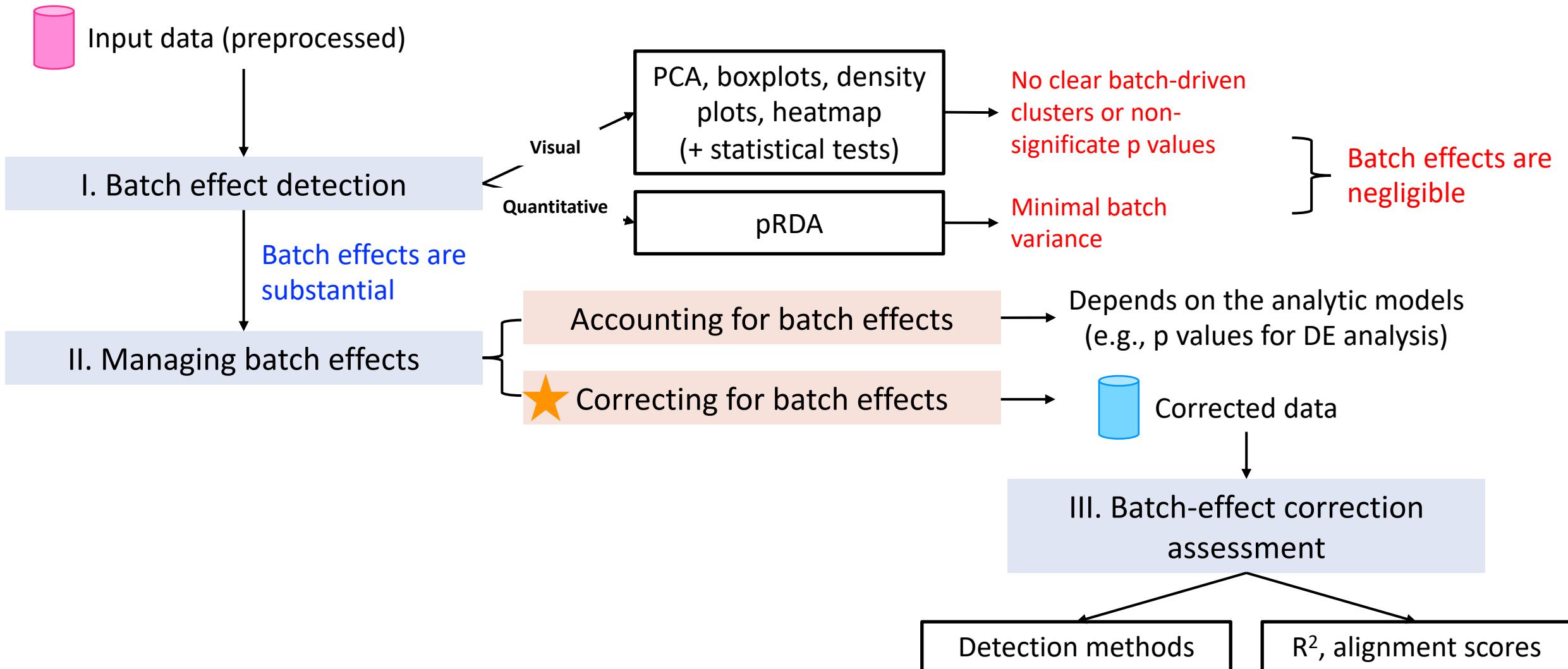


Choice of method

LM: linear model
LMM: linear mixed model



Conclusions



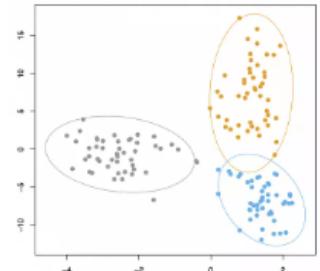
Next workshop

MIG Workshop: Multivariate analysis for omics data integration (bulk)

Training or Workshop

Tuesday 8 July 2025  [Add to my calendar](#)

[Book Now](#)



 Kenneth Myer Building (144), Education Room, Ground Floor

Contact email

 mig-ea@unimelb.edu.au

 **Date:**

Tuesday 8 July 2025

 **Time:**

9:30am - 12:30pm

 **Host:**

[Melbourne Integrative Genomics](#)

 **Location:**

Kenneth Myer Building (144), Education Room, Ground Floor

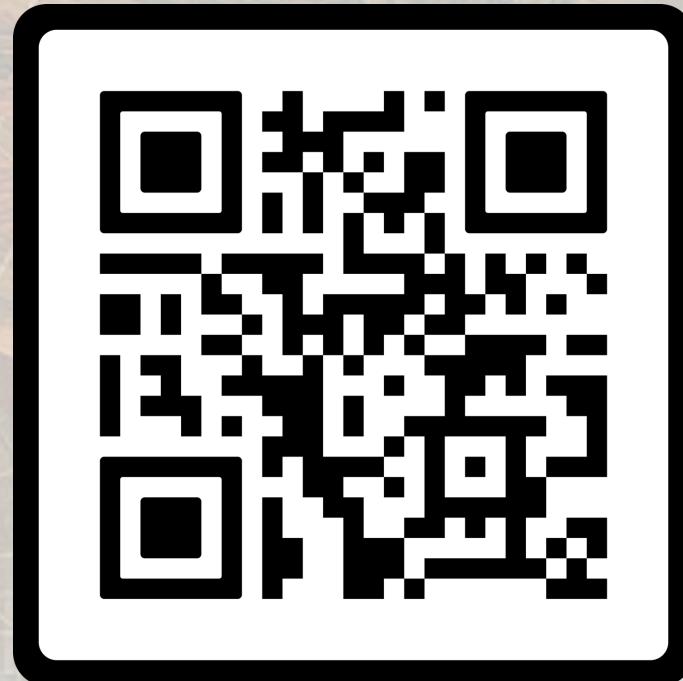
 **Cost:**

\$25

Lead instructors: Prof Kim-Anh Lê Cao (MIG)

Technological improvements have allowed for the collection of data from different molecular compartments (e.g. gene expression, protein abundance) resulting in multiple 'omics' data from the same set of biospecimens or individuals (e.g. transcriptomics, proteomics). This workshop will introduce multivariate analysis to explore and integrate omics data using the R package mixOmics. We will present a few methods implemented in the package and define statistical

Time to reflect and give feedback!



Please fill in the 3-question form before you leave!

It's really important for us to receive feedback so that we can continue delivering these workshops!

Appendix

Further information is available upon interest.

Data pre-processing for omics data

RNA-seq data:

Characteristics:

Count data (discrete, non-negative), zero inflation, overdispersion, uneven library sizes

Transformation:

- Trimmed Mean of M-values (TMM, *edgeR*)
- Median of Ratios (*DESeq2*)

Microbiome data:

Characteristics:

Count data, zero inflation (severe than RNA-seq data), overdispersion, uneven library sizes, compositional structure and multivariate nature

Transformation:

- Centered Log-Ratio (CLR, *mixOmics*)
- Cumulative Sum Scaling (CSS, *metagenomeSeq*)

Data pre-processing for omics data

Metabolomic and proteomic data :

Characteristics:

Continuous and right-skewed data, high intra-group (replicate) variability, missing values

Transformation:

- Imputation of missing values
- Log transformation to reduce skewness
- Median or quantile normalisation to match distributions across samples
- Normalisation to internal standards / housekeeping variables to control for sort of technical variation (systematic)