- Data Sciences for Molecular Phenotyping and Precision Medicine team
- CEA, INRAE, Paris Saclay University, MetaboHUB, 91191 Gif-sur-Yvette, France
- https://scidophenia.github.io

etienne.thevenot@cea.fr

DE LA RECHERCHE À L’INDUSTRIE

ProMetIS: Proteomics and metabolomics data integration

Alyssa Imbert and Etienne Thévenot (*ProMetIS* consortium)

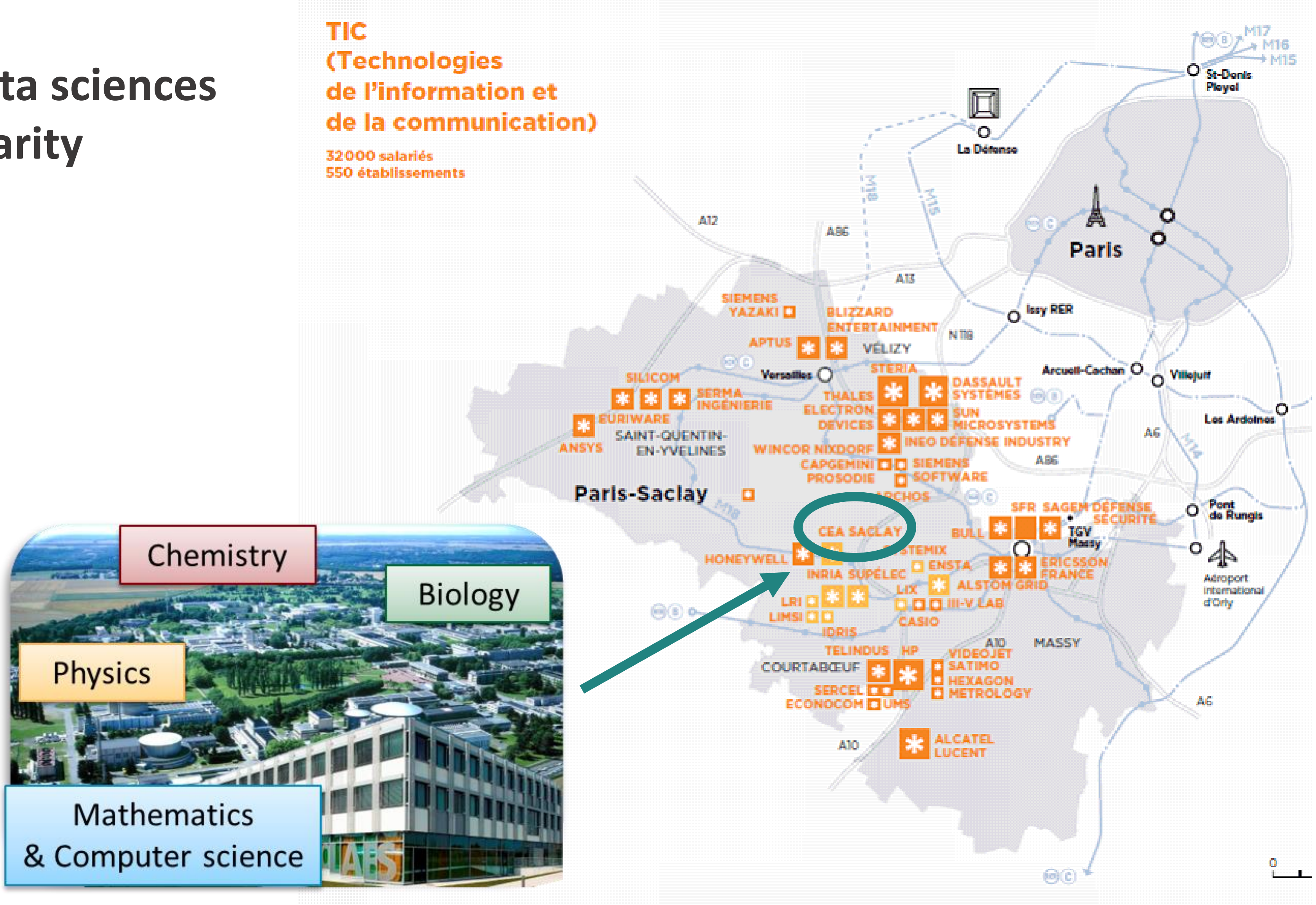
with the help from Camilo Broc and Olivier Sand

DU Bioinformatique intégrative (DUBii)

Who we are

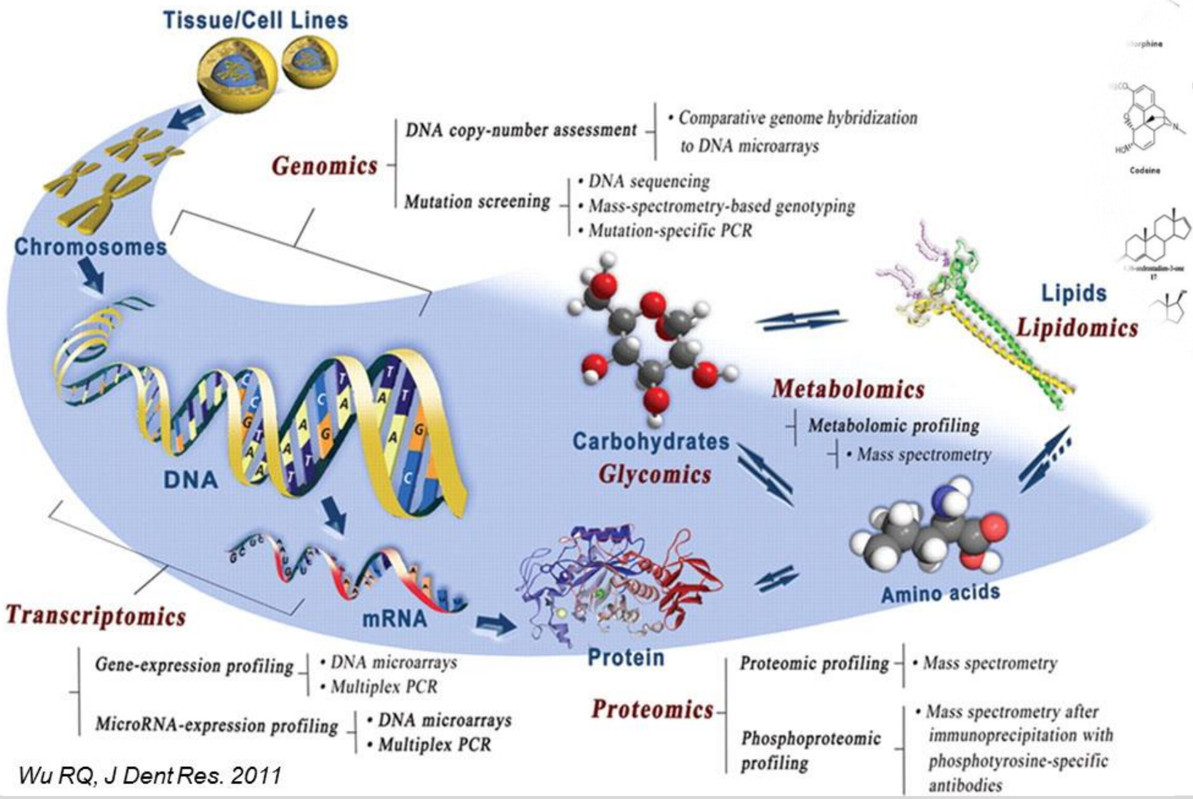
Data sciences at Paris-Saclay and CEA

- Cluster for data sciences
- Interdisciplinarity

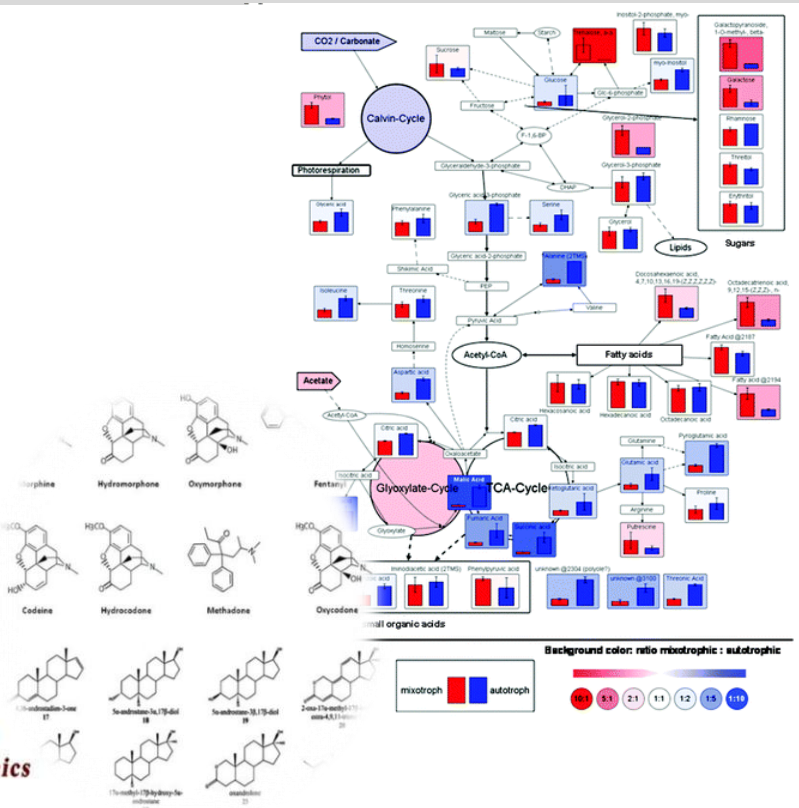


Metabolomics

- omics science
- dedicated to small molecules (< 1kDa)
- involved in metabolic chemical reactions

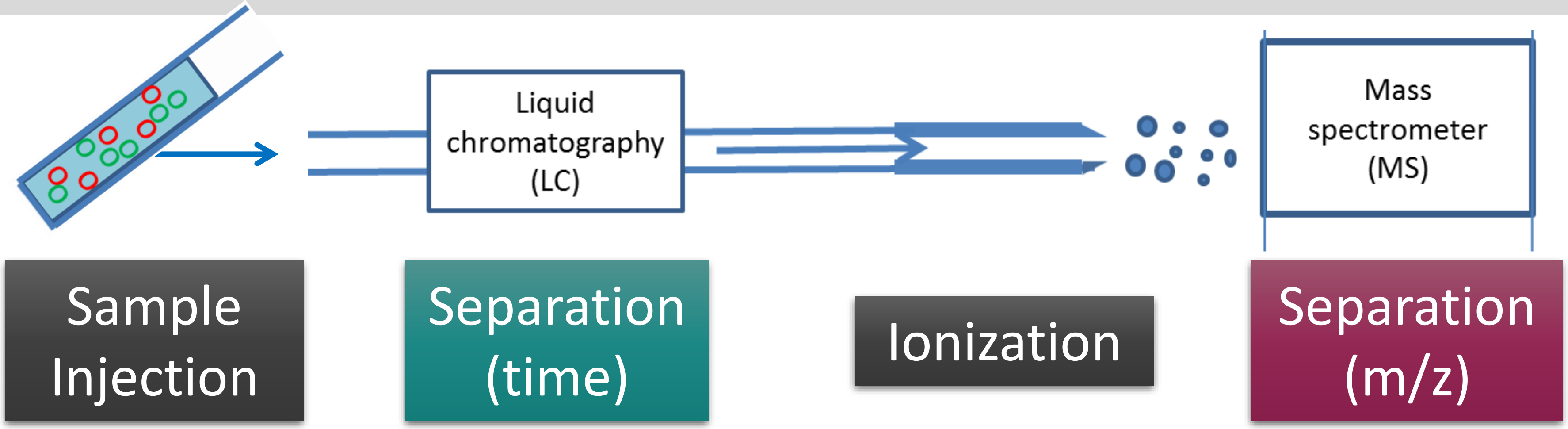


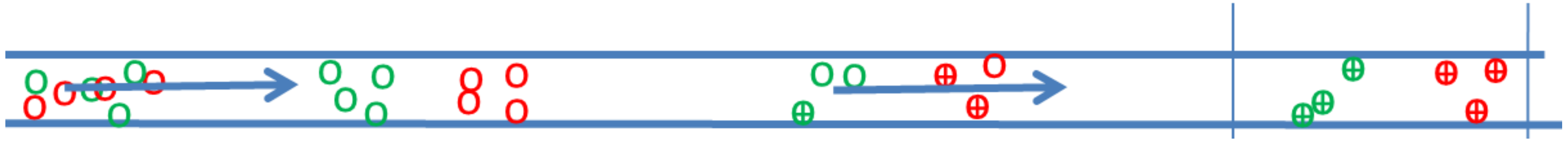
Wu RQ, J Dent Res. 2011

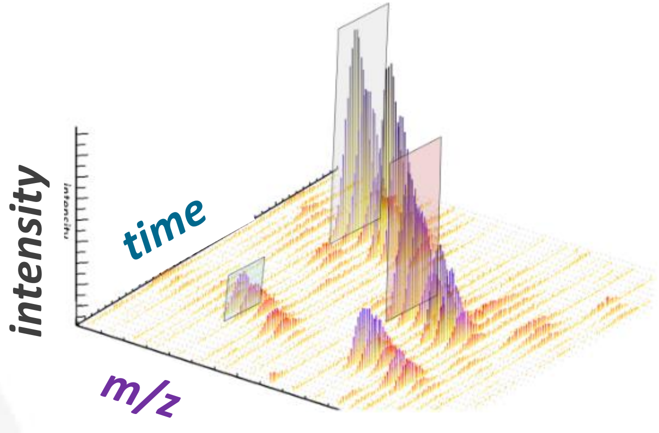


Wishart, 2019. Metabolomics for investigating physiological and pathophysiological processes. Physiological Reviews.

Mass spectrometry







The Data Science team

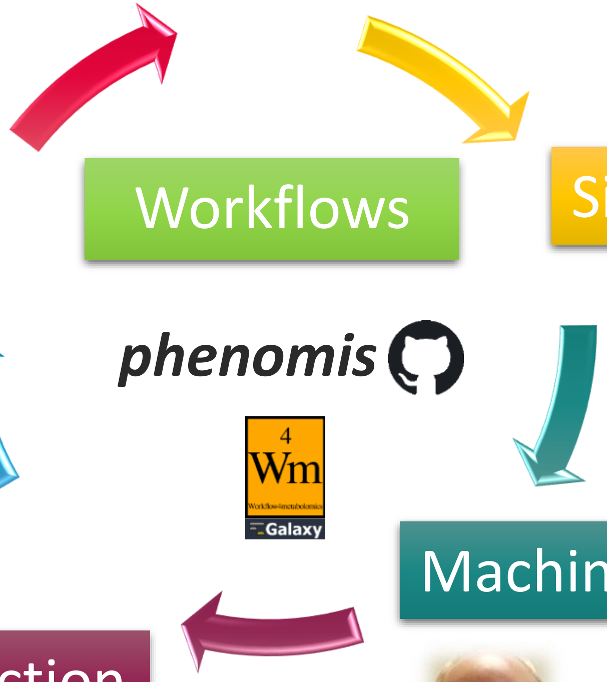
biodb

Pierrick Roger

Sylvain Dechaumet

mineMS2

Alexis Delabrière



Annotation

Workflows

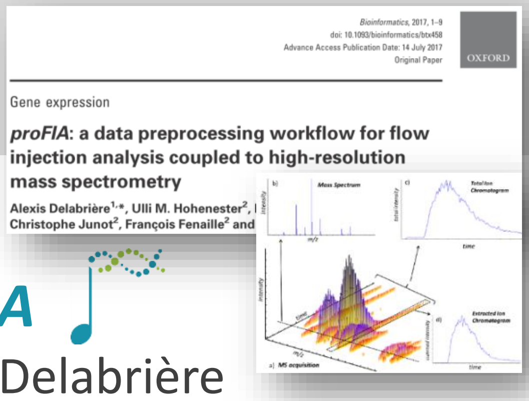
Signal processing

Data integration

phenomis

Machine learning

Feature selection



Bioinformatics, 2017, 1–9
doi: 10.1093/bioinformatics/btx458
Advance Access Publication Date: 14 July 2017
Original Paper

OXFORD

Gene expression

proFIA: a data preprocessing workflow for flow injection analysis coupled to high-resolution mass spectrometry

Alexis Delabrière[1,*], Ulli M. Hohenester[2], Christophe Junot[2], François Fenaille[2] and

proFIA

Alexis Delabrière

ptairMS

Camille Roquencourt

Krystyna Biletska

ProMetIS

Alyssa Imbert

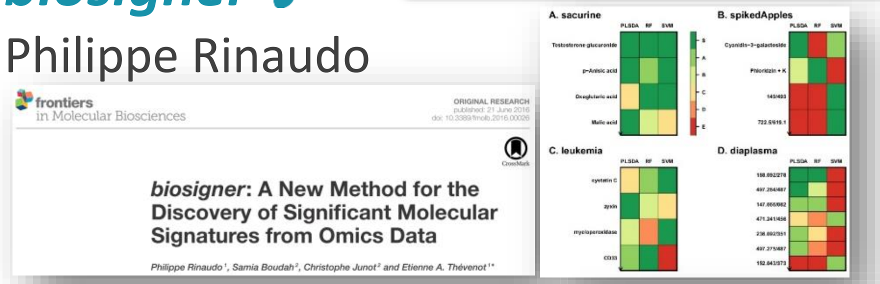
Camilo Broc

ropls

Etienne Thévenot

biosigner

Philippe Rinaudo



frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 21 June 2016
doi: 10.3389/fmolb.2016.00026

CrossMark

biosigner: A New Method for the Discovery of Significant Molecular Signatures from Omics Data

Philippe Rinaudo[1], Samia Boudah[2], Christophe Junot[2] and Etienne A. Thévenot[1*]

Eric Venot

Journal of proteome research

Article

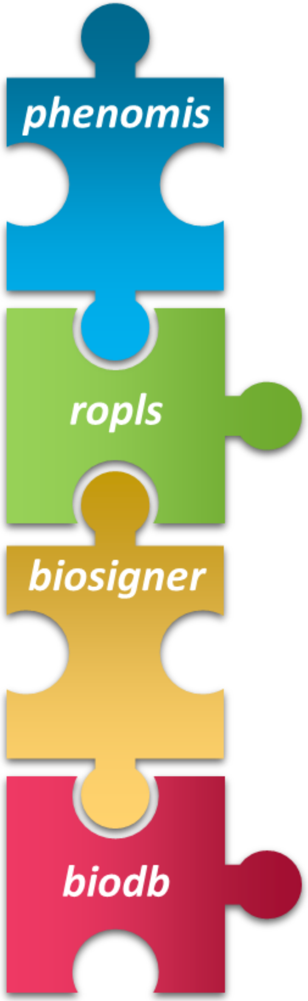
pubs.acs.org/jpr

Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses

Etienne A. Thévenot,[*,†,¶] Aurélie Roux,[‡,¶] Ying Xu,[‡] Eric Ezan,[‡] and Christophe Junot[*,‡]

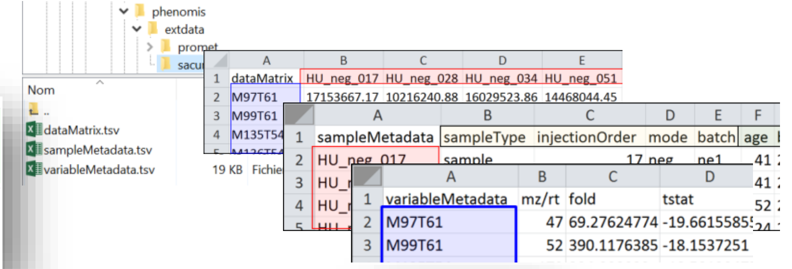
Natacha Lenuzza

Omics data analysis workflow



Importing

```
eSet <- phenomis::reading(dirC)
```



	A	B	C	D	E
1	dataMatrix	HU_neg_017	HU_neg_028	HU_neg_034	HU_neg_051
2	M97T61	17153667.17	10216240.88	16029523.86	14468044.45
3	M99T61				
4	M135T54				

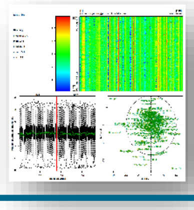
	A	B	C	D	E	F
1	sampleMetadata	sampleType	injectionOrder	mode	batch	age
2	HU_neg_017	sample	17	neg	ne1	41
3	HU_					41
4	HU_					52

	A	B	C	D
1	variableMetadata	mz/rt	fold	tstat
2	M97T61	47	69.27624774	-19.6615585
3	M99T61	52	390.1176385	-18.1537251

Post-processing

Quality control

```
eSet <- phenomis::inspecting(eSet)
```



Signal drift and batch effect correction

```
eSet <- phenomis::correcting(eSet)
```

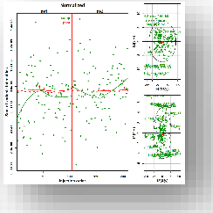
Transformation

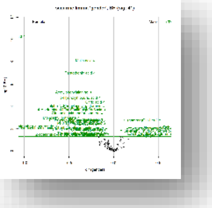
```
eSet <- phenomis::transforming(eSet, methodC = ‘log2’)
```

Statistics

Univariate hypothesis testing

```
eSet <- phenomis::hypotesting(eSet,  testC = ‘limma’,
                              factorNamesVc = ‘gender’, adjustC = ‘BH’)
```



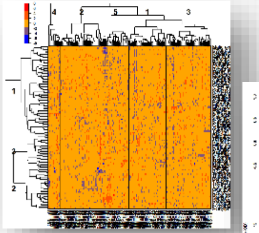


PCA

```
setPca <- ropls::opls(eSet)
eSet <- ropls::getEset(setPca)
```

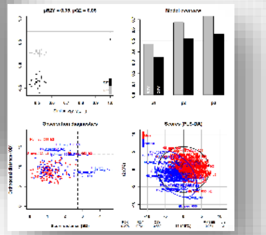
Clustering

```
eSet <- phenomis::clustering(eSet, clustersVi = c(2, 2))
```



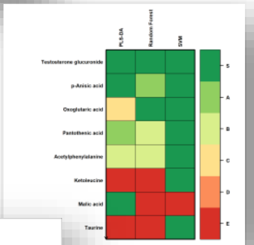
(O)PLS(-DA)

```
setPlsda <- ropls::opls(eSet, ‘gender’)
eSet <- ropls::getEset(setPlsda)
```



Feature selection

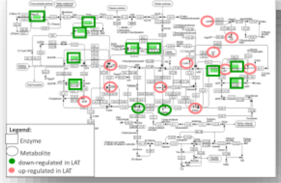
```
setBiosign <- biosigner::biosign(eSet, ‘gender’, seedI = 123)
eSet <- biosigner::getEset(setBiosign)
```



Annotation

Chemical annotation

```
eSet <- phenomis::annotating(eSet, databaseC = c(‘chebi’, ‘local.ms’))
```



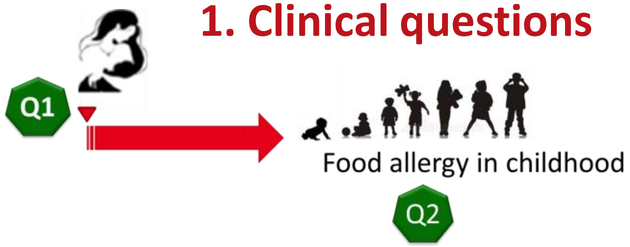
Exporting

```
phenomis::writing(eSet, dirC = getwd())
```

FusionOmics: Computational integration of omics data for high-throughput phenotyping and personalized medicine

Camilo Broc, Karine Patient, Christophe Junot, François Fenaille and Etienne Thévenot

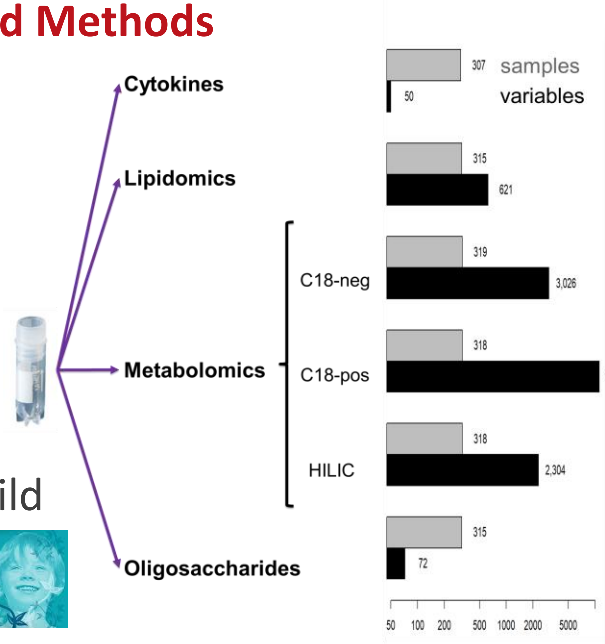
1. Clinical questions



Q1: Characterize **early breast milk**

Q2: Predict the development of **food allergy**

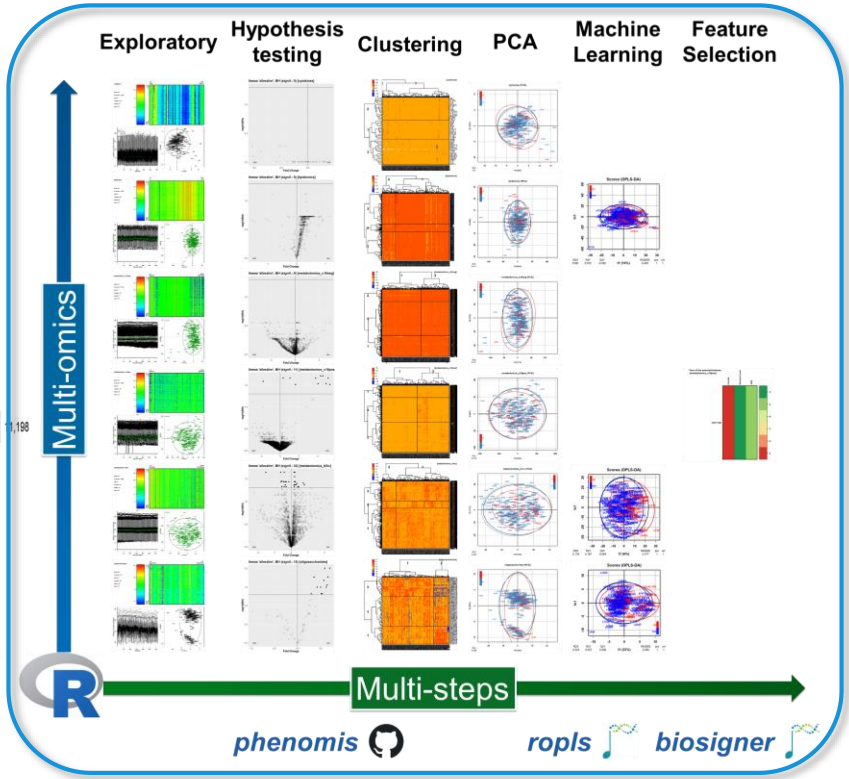
3. Materials and Methods



EDEN cohort
300 mother/child

3. Preliminary results

- Development of a **multi-omics statistical framework** (incl. differential analysis, classification and feature selection)
- Next: **data integration** (multi-blocs, correlation analysis)



Statistician Post-doctorate

Camilo Broc

4. Leverage effects

- ProMetIS (PIA: France Genomique, MetaboHUB, ProFI, IFB),
- MICROB-PREDICT (H2020)
- Master 2 « Systems Immunology » (Sorbonne)

Biology
(Immuno-Allergie Alimentaire)
Analytical Chemistry
(Métabolisme des Médicaments)
Mathematics
(Sciences des Données et de la Décision)

5. Multidisciplinary consortium

CEA (DRF/DRT) - INRA

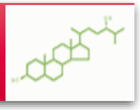
Introduction: proteomics and metabolomics integration

Proteomics and metabolomics

Proteomics

- large-scale study of proteins
- post-translational modifications

Metabolomics



- small molecule substrates, intermediates, and products of metabolism
- peptides, carbohydrates, lipids, nucleosides
- “functional readout of the physiological state”

Complementarity between proteomics and metabolomics

Proteins – Metabolites

- **Interactions**
  - Building blocks of proteins
  - Substrates, cofactors, products of enzymatic reactions
  - Allosteric regulators (enzymes, receptors, transcription factors)
  - Post-translational modifications by covalent link to metabolites

Piazza *et al.* (2018). A map of protein-metabolite interactions reveals principles of chemical communication. *Cell*, **172**:358–372.

Increase the biological understanding

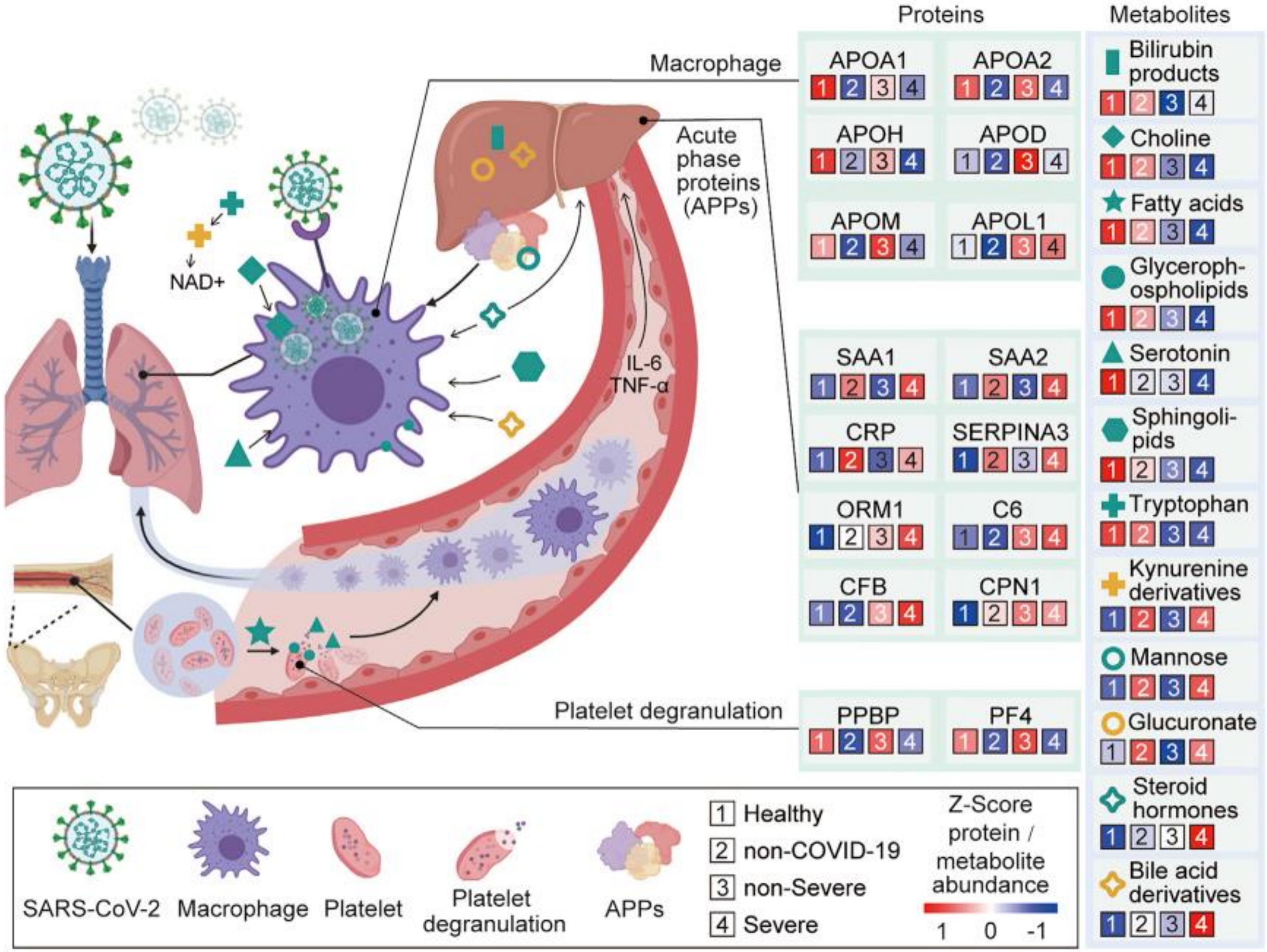
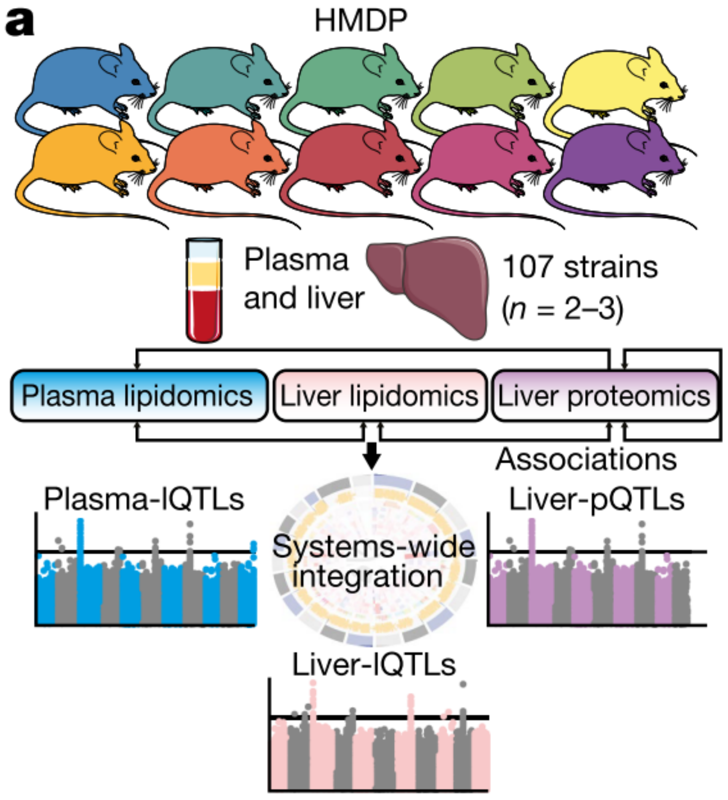


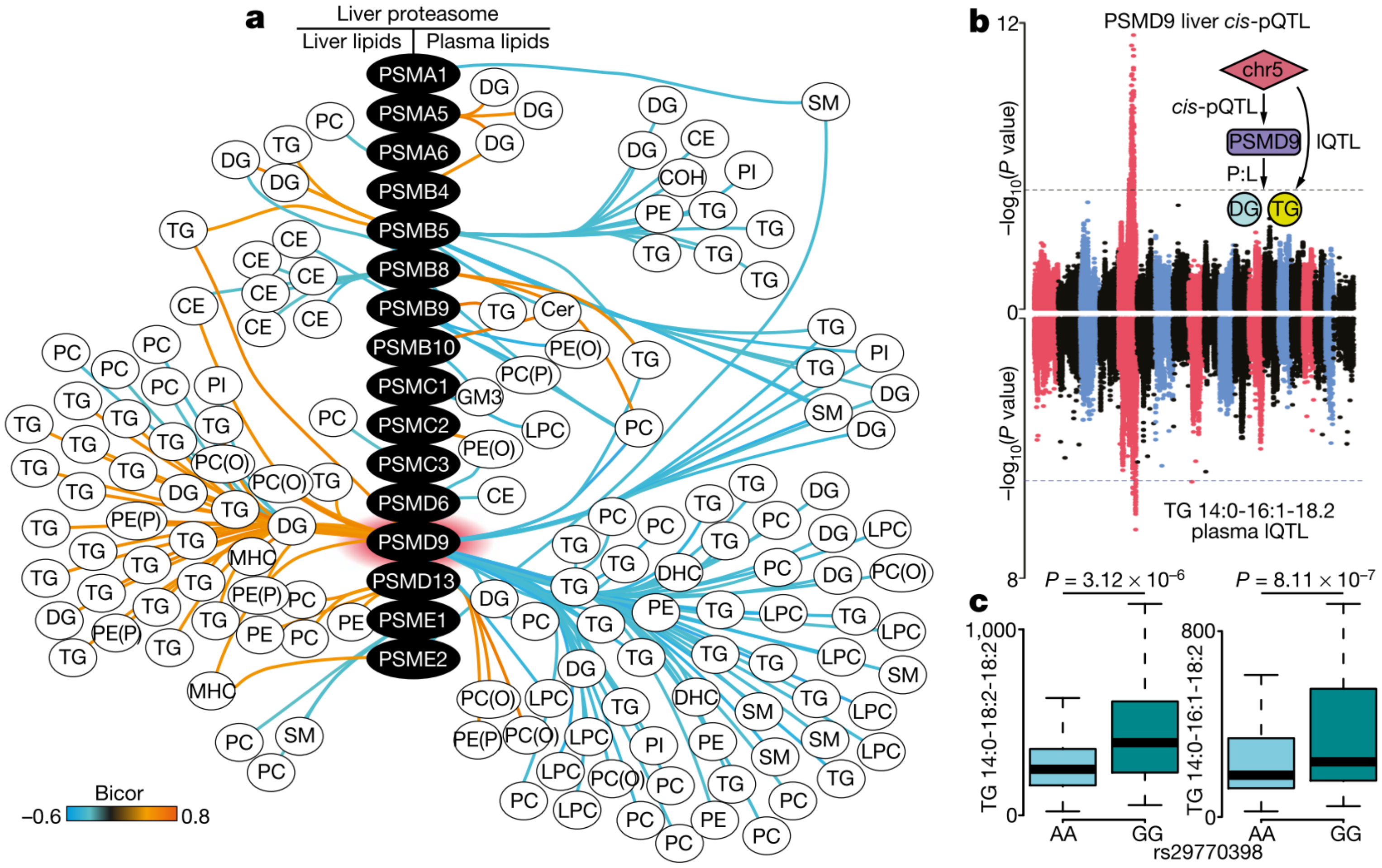
Figure 5. Key Proteins and Metabolites Characterized in Severe COVID-19 Patients in a Working Model

SARS-CoV-2 may target alveolar macrophages via ACE2 receptor, leading to an increase of secretion of cytokines including IL-6 and TNF-α, which subsequently induce the elevation of various APPs such as SAP, CRP, SAA1, SAA2, and C6, which are significantly upregulated in the severe group. Proteins involved in macrophage, lipid metabolism, and platelet degranulation were indicated with their corresponding expression levels in four patient groups.

Shen *et al.* (2020). Proteomic and metabolomic characterization of COVID-19 patient sera. *Cell*, **9**:59–72.

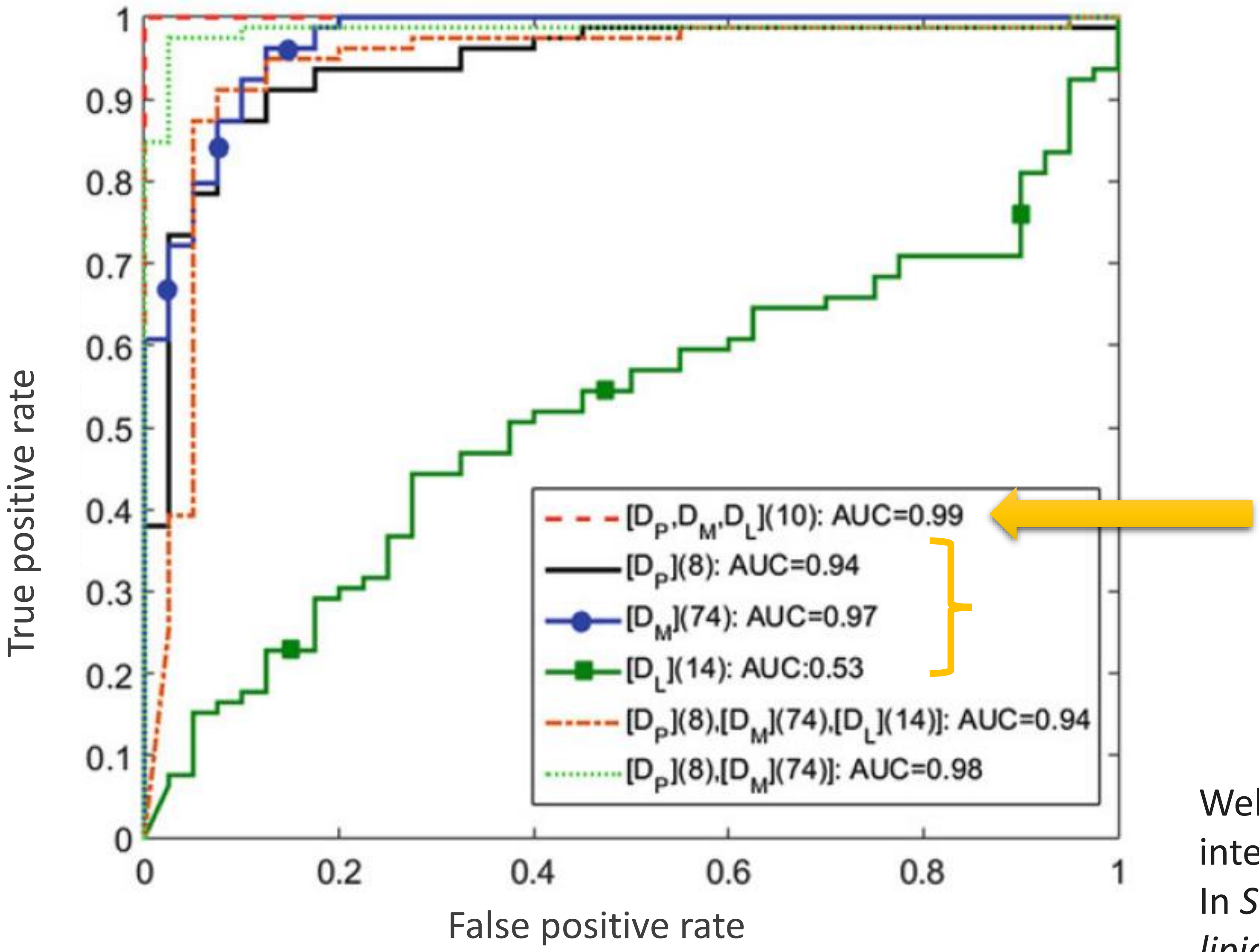
Elucidate gene function





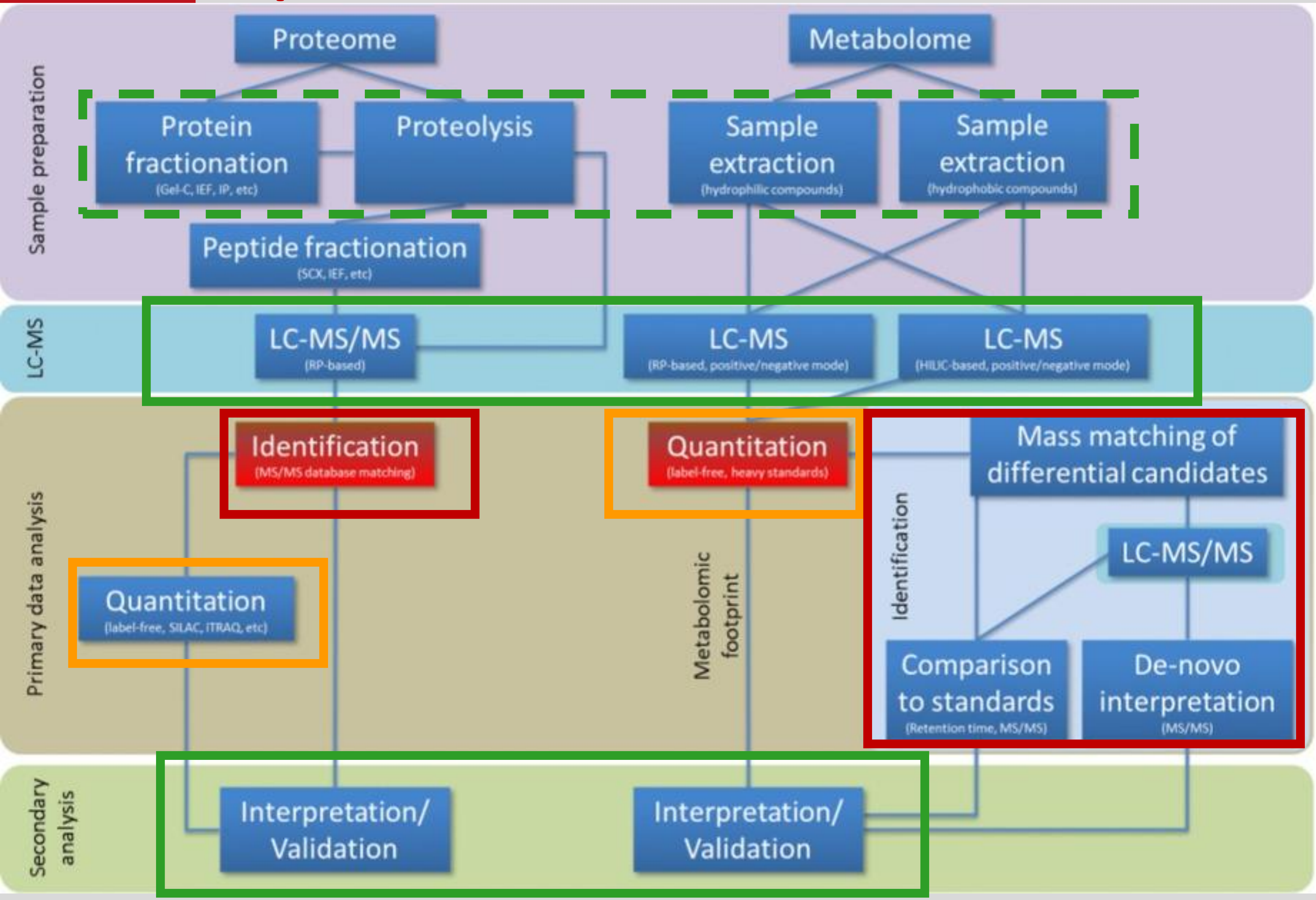
Parker *et al.* (2019). An integrative systems genetic analysis of mammalian lipid metabolism. *Nature*, **567**:187–193.

Increase the predictive performance



Webb-Robertson *et al.* (2016). Bayesian posterior integration for classification of mass spectrometry data. In *Statistical analysis of proteomics, metabolomics, and lipidomics data using mass spectrometry* (pp. 203–211).

Similarities in data quantification and statistical analysis
Specificities in data annotation

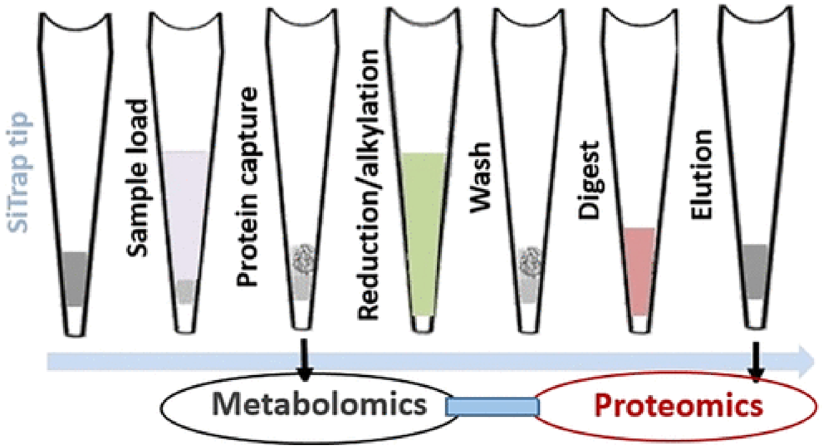


Fischer *et al.* (2013). Two birds with one stone: doing metabolomics with your proteomics kit. *Proteomics*, **13**:3371-3386.

Common sample preparation studies

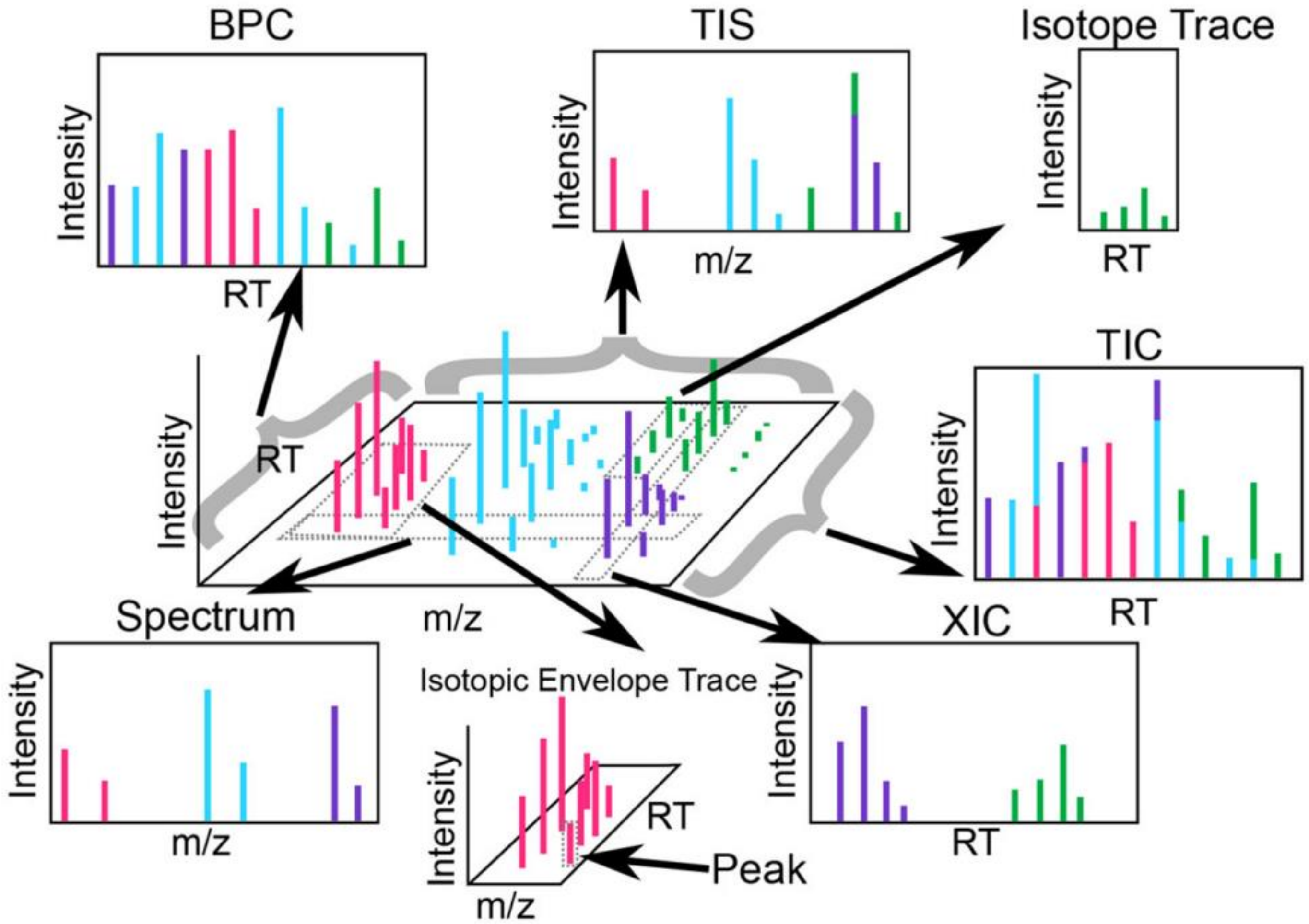
Fischer *et al.* (2013). Two birds with one stone: doing metabolomics with your proteomics kit. *PROTEOMICS*, **13**:3371-3386.

Blum *et al.* (2018). Single-platform ‘multi-omic’ profiling: unified mass spectrometry and computational workflows for integrative proteomics–metabolomics analysis. *Molecular Omics*, **14**:307–319.



Zougman *et al.* (2019). Detergent-free simultaneous sample preparation method for proteomics and metabolomics. *Journal of Proteome Research*, **19**:2838–2844.

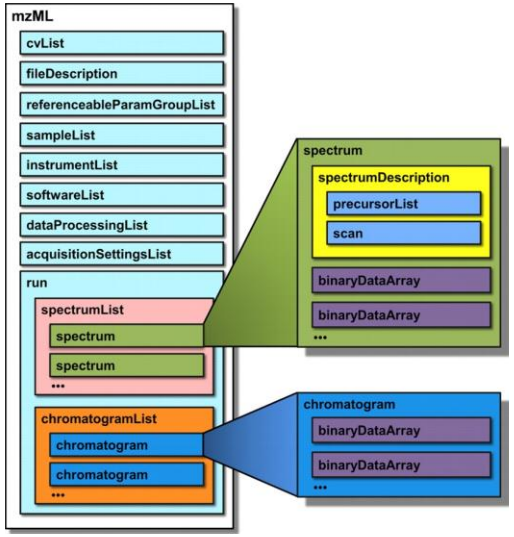
Common nomenclature



Smith *et al.* (2014). Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view. *BMC Bioinformatics*. **15**.

Common formats

- **Storage**
  - raw data: **mzML**

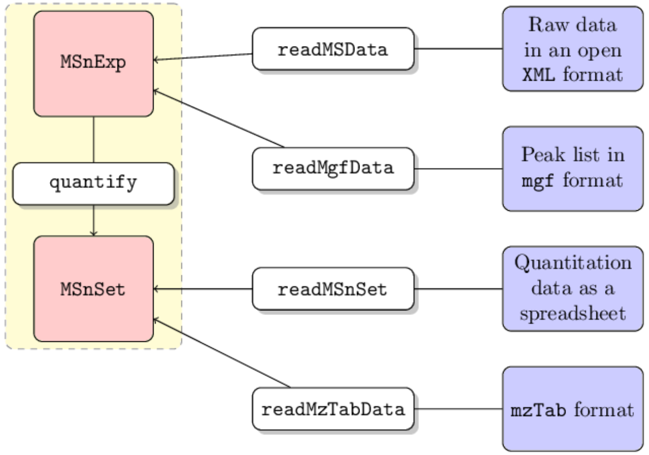


Martens *et al*. (2010). mzML - a community standard for mass spectrometry data. *Molecular & Cellular Proteomics*. **10**.

  - processed data: **mzTab**
    - quantification
    - identification

Griss *et al*. (2014). The mzTab data exchange format: Communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Molecular & Cellular Proteomics*. **13**:2765-2775.

- **Computation**
  - R object: **MSnbase**

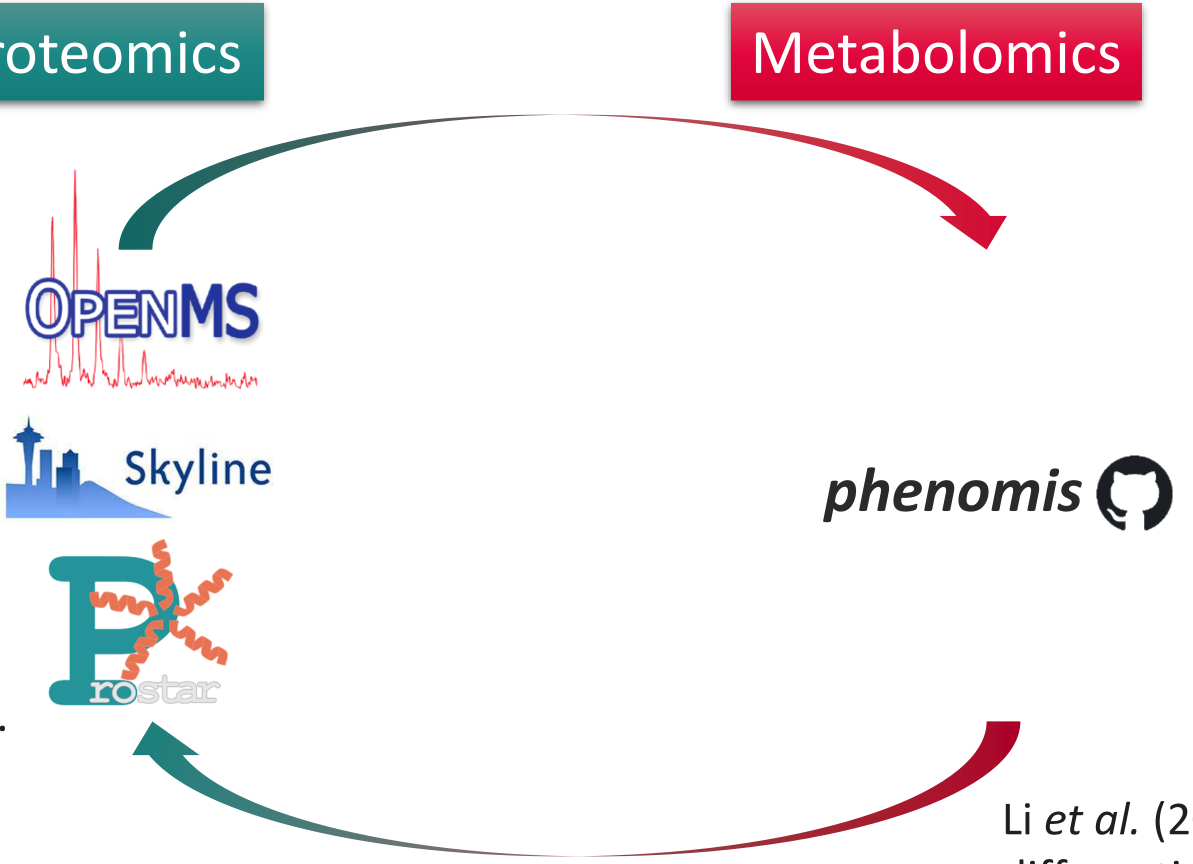


Gatto *et al*. (2020). MSnbase, efficient and elegant R-based processing and visualization of raw mass spectrometry data. *Journal of Proteome Research*.

Toward common software

Rurik *et al.* (2020). Metabolomics data processing using OpenMS. *Methods in Molecular Biology,* **2104**.

Adams *et al.* (2020). Skyline for small molecules: A unifying software package for quantitative metabolomics. *Journal of Proteome Research*, **19**:1447-1458.



SDA

WGCNA

ProMetIS

Li *et al.* (2019). SDA: a semi-parametric differential abundance analysis method for metabolomics and proteomics data. *BMC Bioinformatics*. **20**.

Pei *et al.* (2017). WGCNA application to proteomic and metabolomic data analysis. *Methods in Enzymology*. 135-158.

Data integration: questions

- Which blocks are the most important for the stratification/prediction?
- Which features?
- What is the specific/shared information from each block?
- How are the features from different blocks correlated?
- Which biological pathways/networks are significantly involved?

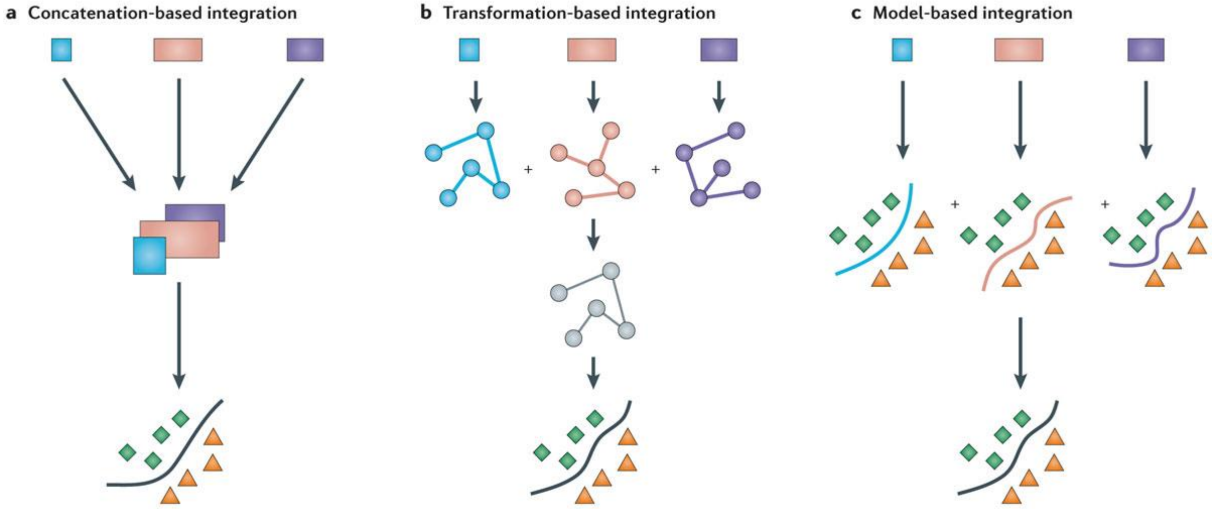
Data integration: challenges

- Normalization of each block
- Confounding effects (for each block)
- Overfitting (limited number of samples)

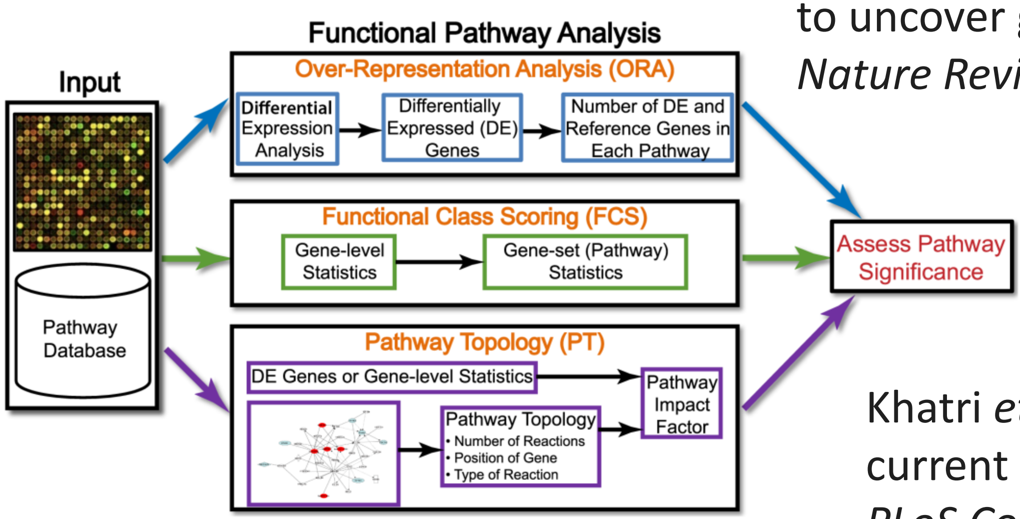
  ⇒ validation (statistical, biological)
- Feature selection
- Limited annotation of metabolites
- Redundancy/specificity/ambiguity of chemical/biological identifiers in the databases
- Partial coverage of the proteome and metabolome

Data integration: approaches

- **Biostatistics**
  - Fusion
    - low (concatenation of blocks)
    - middle (feature selection/latent variables from each block + model on top)
    - high (one model for each block + vote)
  - Correlation networks
- **Bioinformatics**
  - Mapping
  - Enrichment
    - Molecule set
    - Topology-based


a Concatenation-based integration
b Transformation-based integration
c Model-based integration
+
+
+
+

Ritchie *et al.* (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, **16**:85–97.

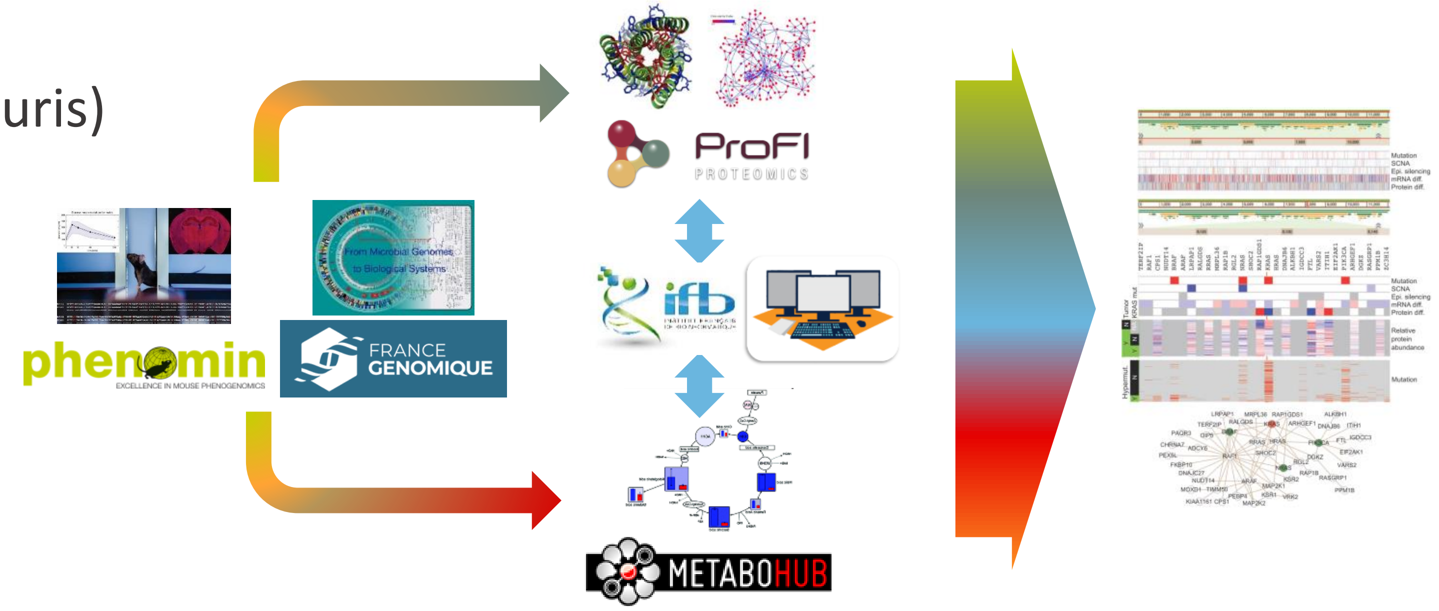

Functional Pathway Analysis
Input
Pathway Database
Over-Representation Analysis (ORA)
Differential Expression Analysis
Differentially Expressed (DE) Genes
Number of DE and Reference Genes in Each Pathway
Functional Class Scoring (FCS)
Gene-level Statistics
Gene-set (Pathway) Statistics
Pathway Topology (PT)
DE Genes or Gene-level Statistics
Pathway Topology
• Number of Reactions
• Position of Gene
• Type of Reaction
Pathway Impact Factor
Assess Pathway Significance

Khatri *et al.* (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology*, **8**: e1002375.

ProMetIS: deep phenotyping of mouse models by proteomics and metabolomics

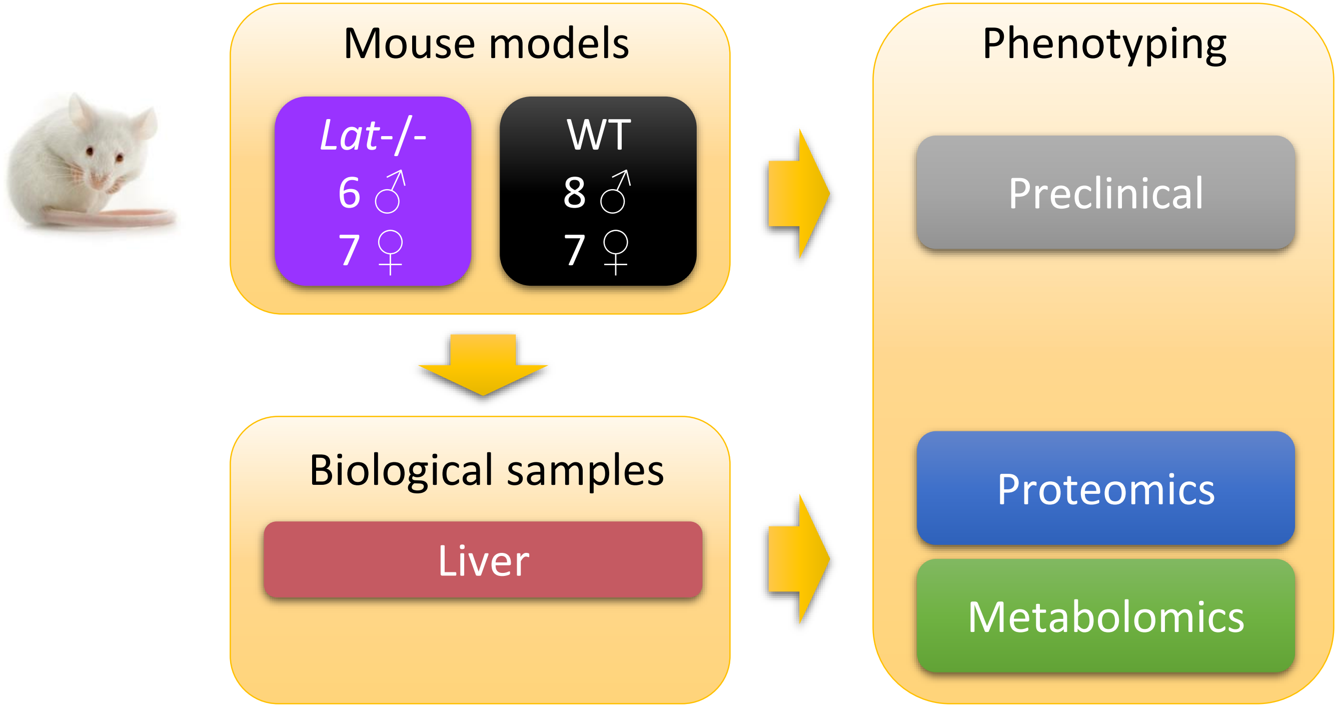
The ProMetIS project

- **Objective: high-throughput integration of proteomics and metabolomics data**
  - innovative methods
  - high-quality datasets
  - software tools
  - workflows
- **Case study: molecular phenotyping of mouse models from the IMPC consortium**
- **Partner infrastructures**
  - France Génomique
  - PHENOMIN (Institut Clinique de la Souris)
  - ProFI proteomics
  - MetaboHUB
  - Institut Français de Bioinformatique
- **Post-doctorate**
  - Alyssa Imbert



phenomin
EXCELLENCE IN MOUSE PHENOGENOMICS
From Microbial Genomes to Biological Systems
FRANCE GENOMIQUE
ProFI
PROTEOMICS
ifb
METABOHUB

Biological question: characterization of LAT knock-out mice



Mouse models

Lat-/-
6 ♂
7 ♀

WT
8 ♂
7 ♀

Biological samples

Liver

Phenotyping

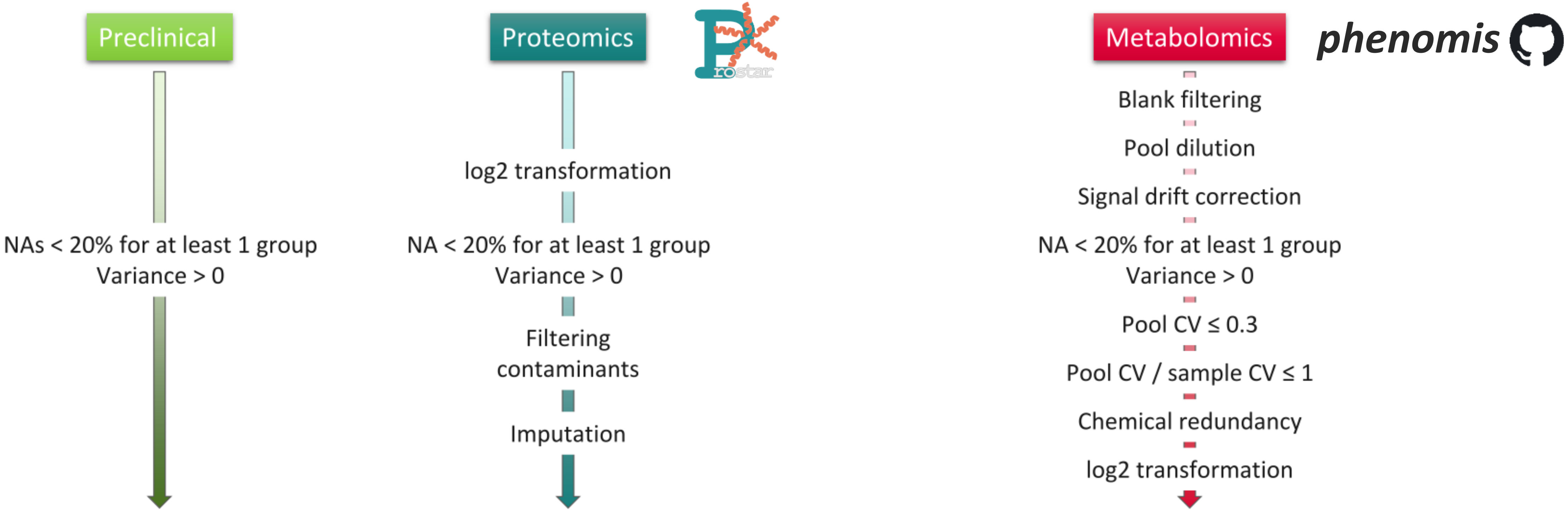
Preclinical

Proteomics

Metabolomics

- **LAT (linker for activation of T cells; OMIM: 602354) involved in:**
  - T-cell receptor (TCR) signaling — Roncagalli et al. (2010). LAT signaling pathology: an ”autoimmune” condition without T cell self-reactivity. *Trends in Immunology*, **31**:253–259.
  - Neurodevelopmental diseases — Loviglio *et al.* (2017). The immune signaling adaptor LAT contributes to the neuroanatomical phenotype of 16p11.2 BP2-BP3 CNVs. *The American Journal of Human Genetics*, **101**:564–577.

Datasets: preclinical, proteomics and metabolomics



Preclinical

NAs < 20% for at least 1 group
Variance > 0

Proteomics

log2 transformation

NA < 20% for at least 1 group
Variance > 0

Filtering contaminants

Imputation

Metabolomics

Blank filtering

Pool dilution

Signal drift correction

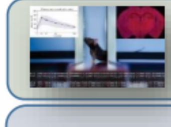
NA < 20% for at least 1 group
Variance > 0

Pool CV ≤ 0.3

Pool CV / sample CV ≤ 1

Chemical redundancy

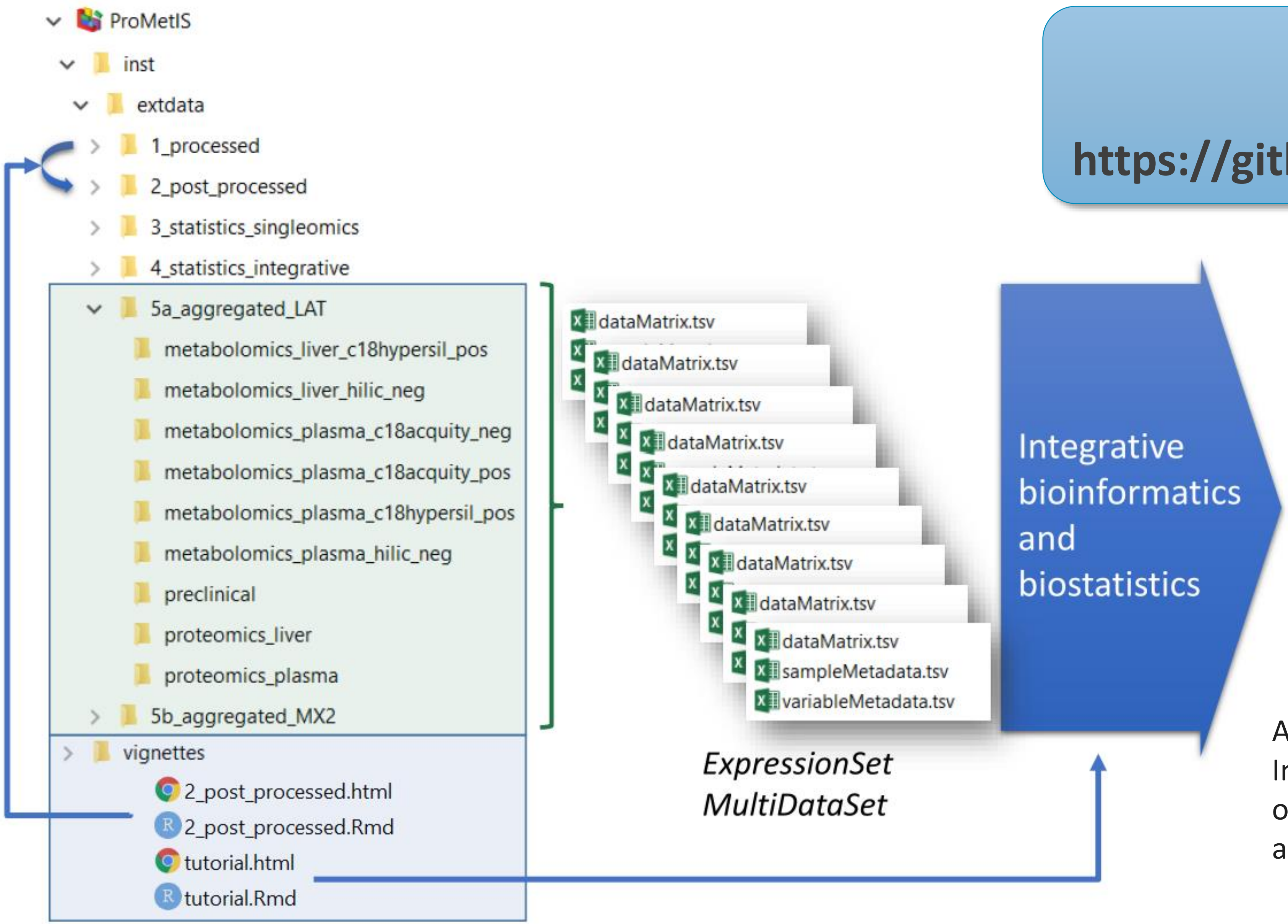
log2 transformation

	preclinical	236			
			liver_proteomics	2,187	liver_metabo_c18hypersil_pos	5,665 [138]
					liver_metabo_hilic_neg	2,866 [199]
			plasma_proteomics	446	plasma_metabo_c18hypersil_pos	4,788 [113]
					plasma_metabo_hilic_neg	3,131 [191]
					plasma_metabo_c18acquity_pos	6,104 [78]
					plasma_metabo_c18acquity_neg	1,584 [49]

Datasets & code availability: the *ProMetIS* package

ProMetIS

https://github.com/IFB-ElixirFr/ProMetIS

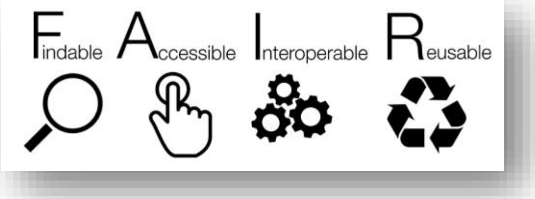


```
ProMetIS
  inst
    extdata
      1_processed
      2_post_processed
      3_statistics_singleomics
      4_statistics_integrative
      5a_aggregated_LAT
        metabolomics_liver_c18hypersil_pos
        metabolomics_liver_hilic_neg
        metabolomics_plasma_c18acquity_neg
        metabolomics_plasma_c18acquity_pos
        metabolomics_plasma_c18hypersil_pos
        metabolomics_plasma_hilic_neg
        preclinical
        proteomics_liver
        proteomics_plasma
      5b_aggregated_MX2
  vignettes
    2_post_processed.html
    2_post_processed.Rmd
    tutorial.html
    tutorial.Rmd
```

dataMatrix.tsv
dataMatrix.tsv
dataMatrix.tsv
dataMatrix.tsv
dataMatrix.tsv
dataMatrix.tsv
dataMatrix.tsv
dataMatrix.tsv
dataMatrix.tsv
sampleMetadata.tsv
variableMetadata.tsv

ExpressionSet
MultiDataSet

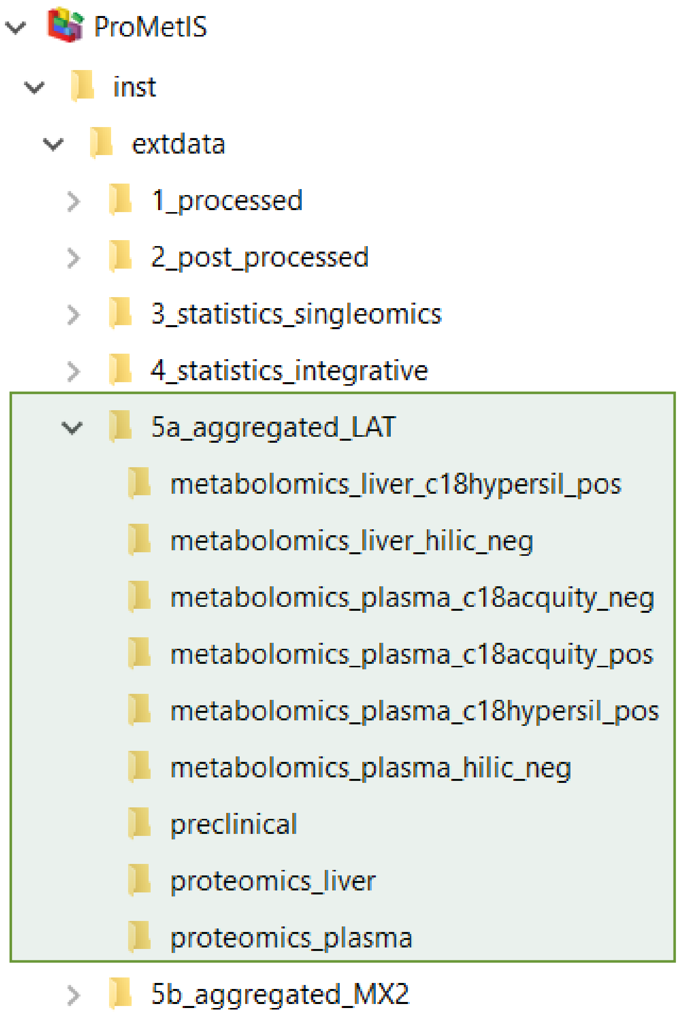
Integrative bioinformatics and biostatistics



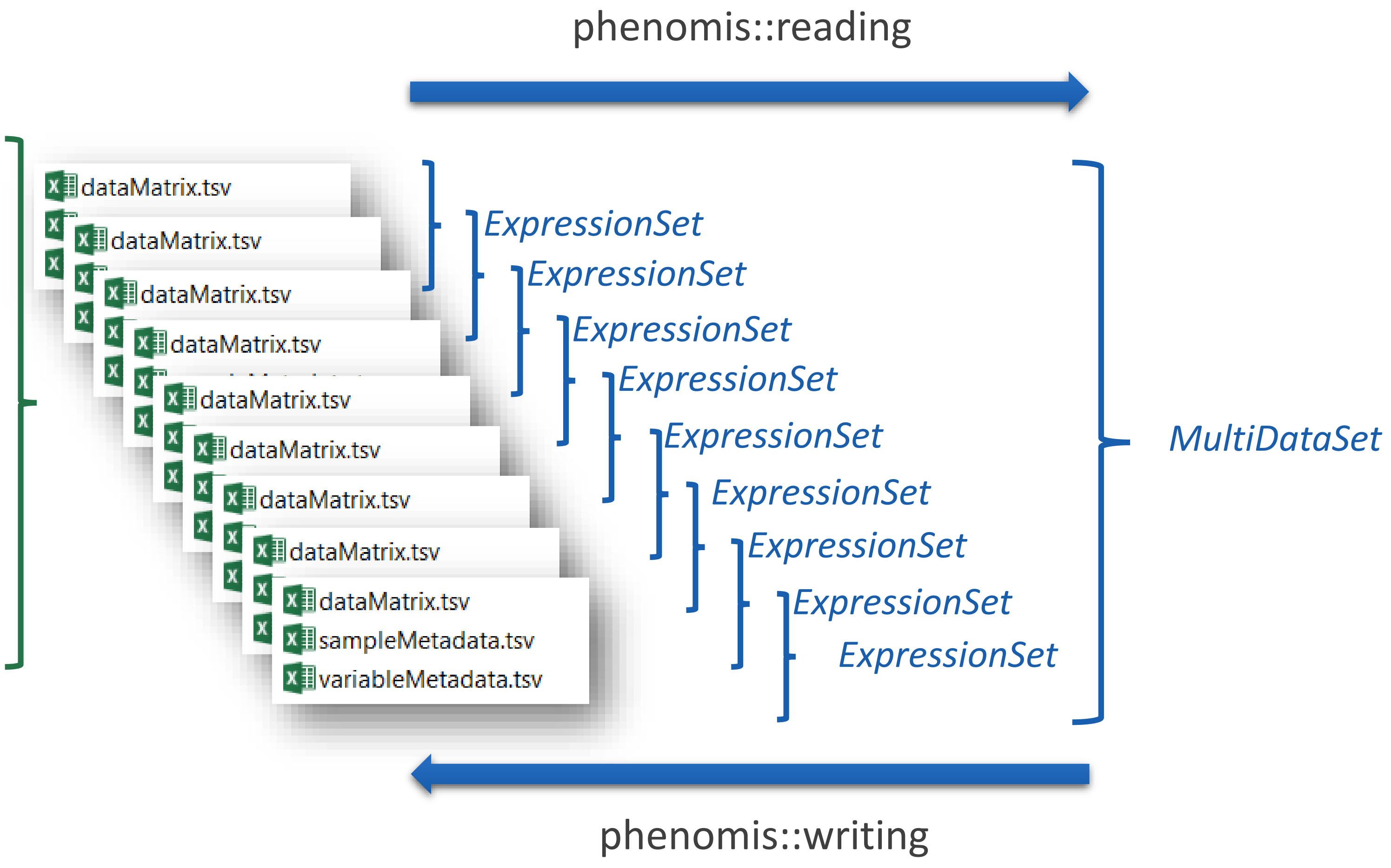
About to be submitted to *Scientific Data*: Imbert *et al*. ProMetIS: deep phenotyping of mouse models by combined proteomics and metabolomics analysis. *submitted*.

Hands-on

Reading the data



- ProMetIS
  - inst
    - extdata
      - 1_processed
      - 2_post_processed
      - 3_statistics_singleomics
      - 4_statistics_integrative
      - 5a_aggregated_LAT
        - metabolomics_liver_c18hypersil_pos
        - metabolomics_liver_hilic_neg
        - metabolomics_plasma_c18acquity_neg
        - metabolomics_plasma_c18acquity_pos
        - metabolomics_plasma_c18hypersil_pos
        - metabolomics_plasma_hilic_neg
        - preclinical
        - proteomics_liver
        - proteomics_plasma
      - 5b_aggregated_MX2



phenomis::reading

dataMatrix.tsv
dataMatrix.tsv
dataMatrix.tsv
dataMatrix.tsv
dataMatrix.tsv
dataMatrix.tsv
dataMatrix.tsv
dataMatrix.tsv
dataMatrix.tsv
sampleMetadata.tsv
variableMetadata.tsv

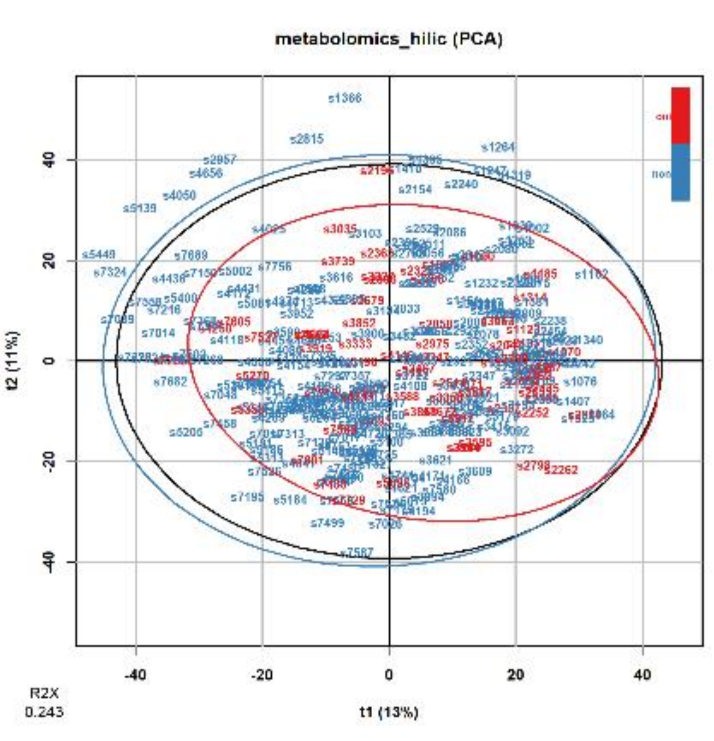
ExpressionSet
ExpressionSet
ExpressionSet
ExpressionSet
ExpressionSet
ExpressionSet
ExpressionSet
ExpressionSet
ExpressionSet

MultiDataSet

phenomis::writing

Statistical workflow

- **Loading the datasets**
  - restricting to the liver tissue and to the annotated metabolomics features only
- **Single-omics analysis**
  - exploratory (PCA)
  - how much information about the LAT knock-out is provided by each dataset
    - univariate hypothesis testing
    - multivariate PLS-DA
- **Multi-omics analysis**
  - unsupervised (MCIA)
    - are the two genotypes separated?
    - what about the difference between genders?
  - supervised (multi-block PLS-DA)
    - which dataset(s) most contribute to the discrimination?
    - which features most contribute to the discrimination within each dataset?*
    - what is the correlation between those features



Principal Component Analysis (PCA)

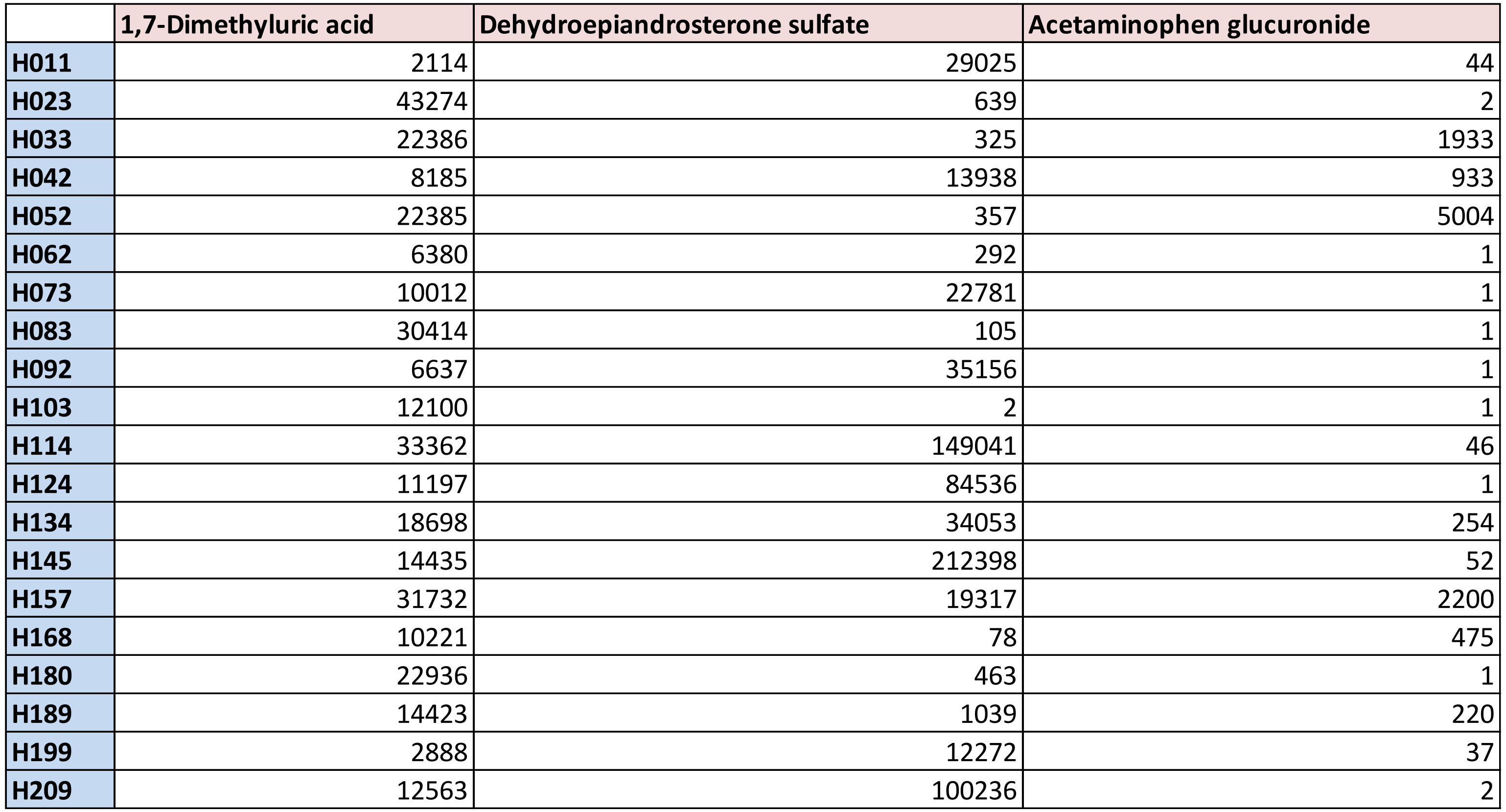
Objectives

- **Visualize the data**
  - by selecting a few components which capture most of the spread (variance) of the cloud of samples
- **Detect outliers**
  - which may bias the computation of the component
- **Detect clusters of samples**
  - which may suggest an internal structuration of the data

Unsupervised analysis

p = 110 (quantitative) variables

n = 183 samples



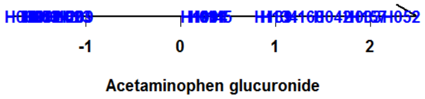
	1,7-Dimethyluric acid	Dehydroepiandrosterone sulfate	Acetaminophen glucuronide
H011	2114	29025	44
H023	43274	639	2
H033	22386	325	1933
H042	8185	13938	933
H052	22385	357	5004
H062	6380	292	1
H073	10012	22781	1
H083	30414	105	1
H092	6637	35156	1
H103	12100	2	1
H114	33362	149041	46
H124	11197	84536	1
H134	18698	34053	254
H145	14435	212398	52
H157	31732	19317	2200
H168	10221	78	475
H180	22936	463	1
H189	14423	1039	220
H199	2888	12272	37
H209	12563	100236	2

...

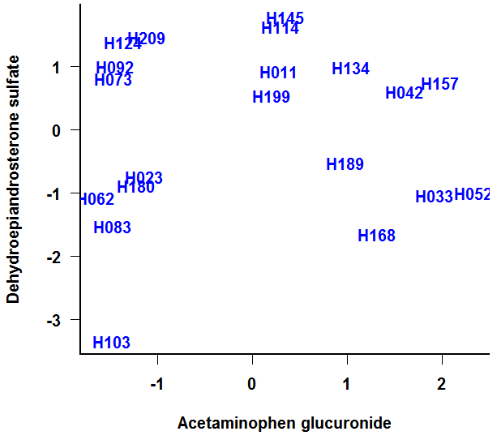
$\mathbf{X}$

How to visualize multivariate observations?

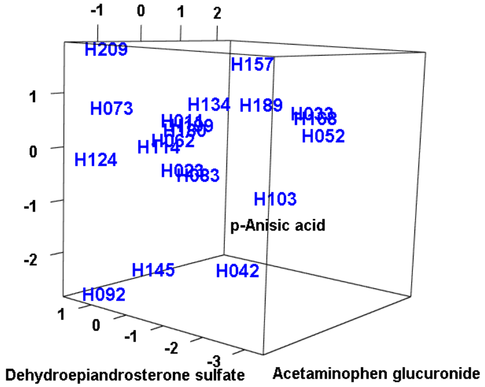
1 variable



2 variables



3 variables



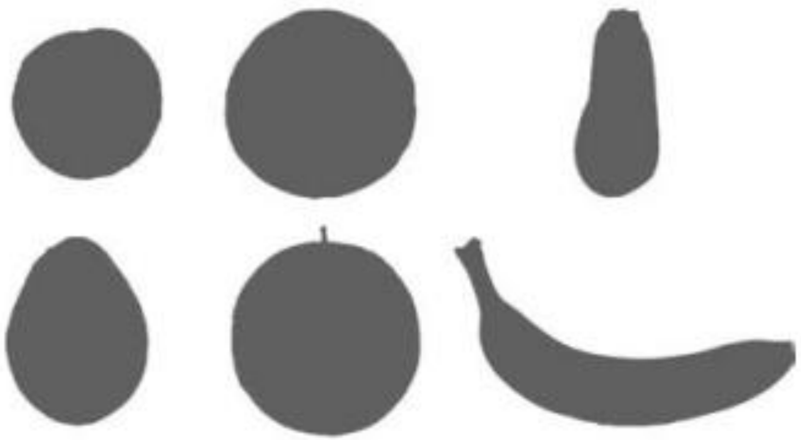
p variables

?

=> Dimension reduction

Projection

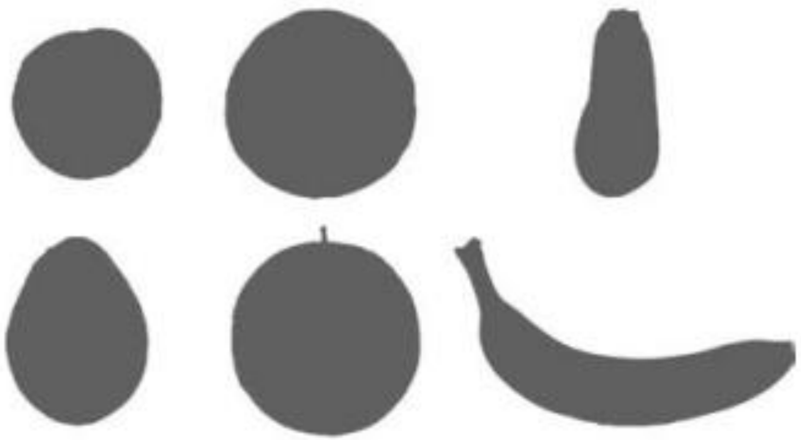
- **Projected distances as high as possible**

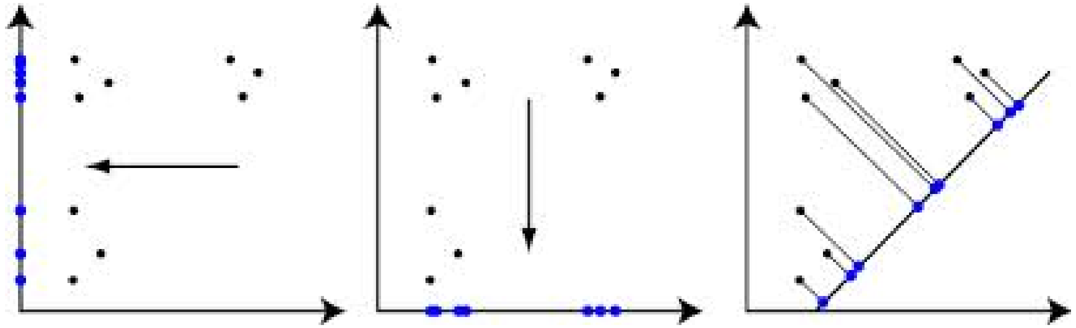


Husson and Pages (2011). Exploratory multivariate analysis by example using R. Chapman & Hall/CRC

Projection on latent variables

- **Projected distances as high as possible**
- **Define new variables as linear combination of original ones**

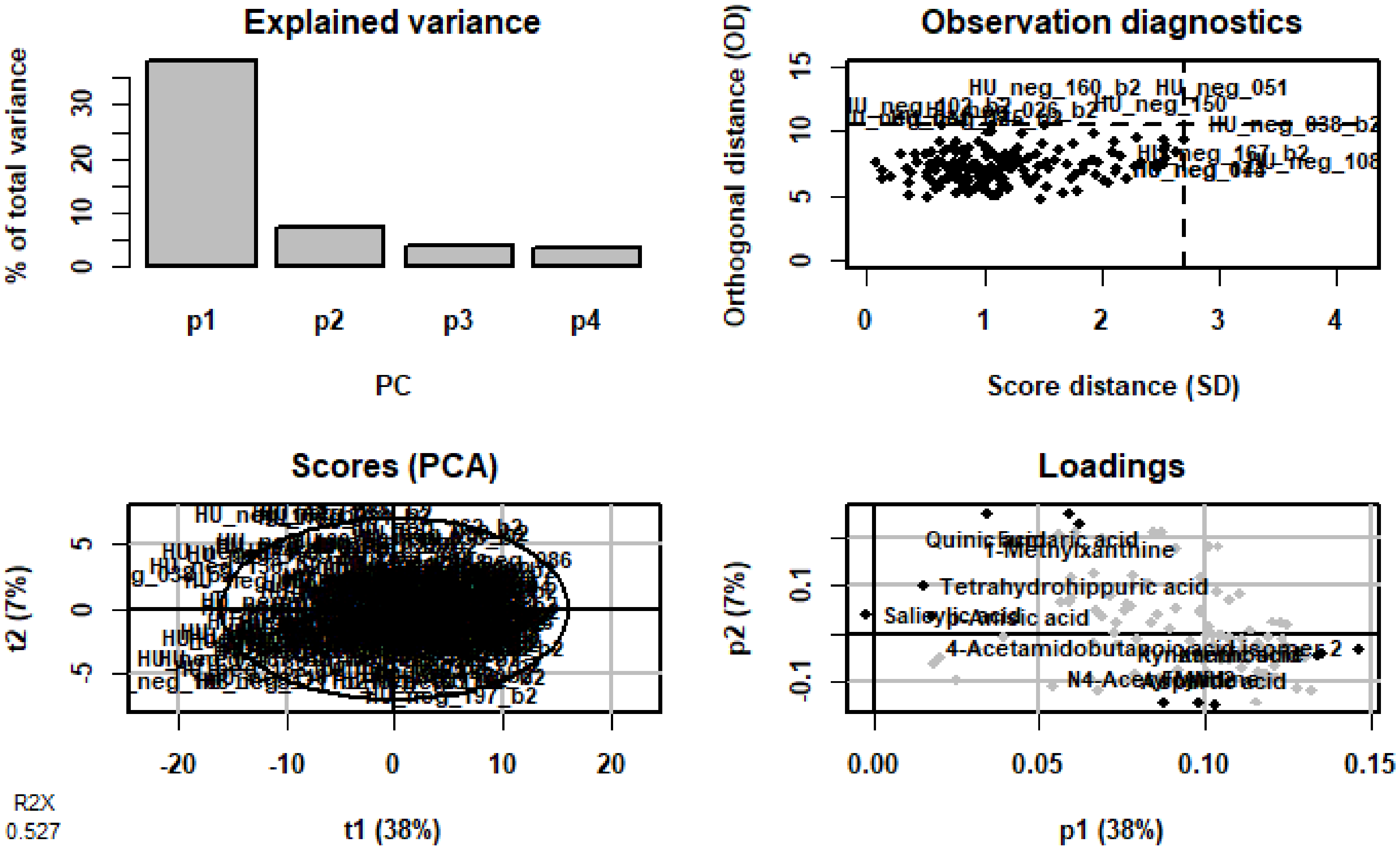




Husson and Pages (2011). Exploratory multivariate analysis by example using R. *Chapman & Hall/CRC*

Graphical results

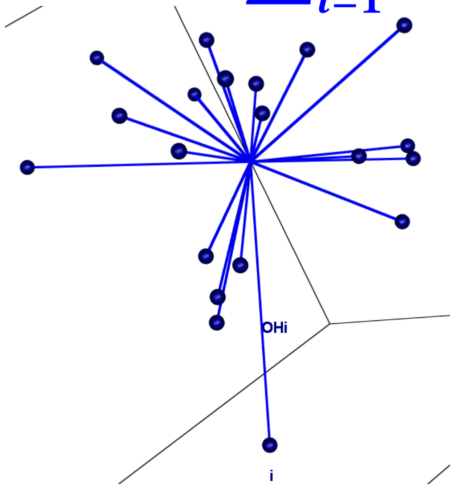
- scree plot, outliers, and the loading and score plots



Diagnostics R2X: How much of the original inertia is still reflected by the model?

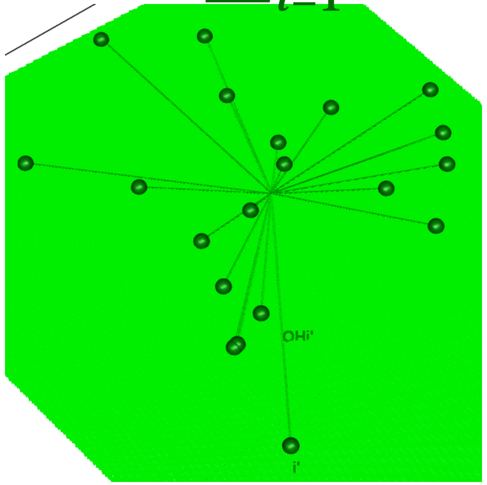
Total

$$TSS = \sum_{i=1}^{n} OH_i^2$$



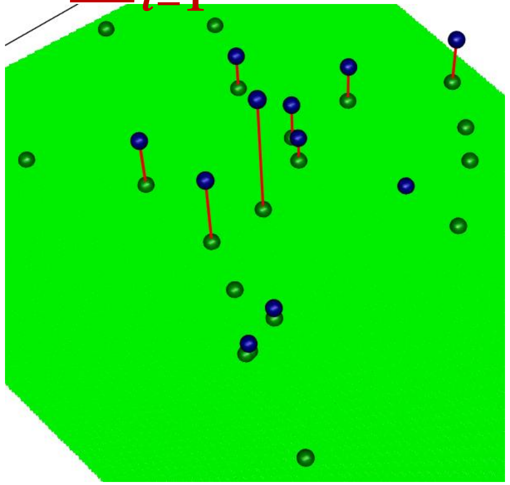
Explained

$$ESS = \sum_{i=1}^{n} OH'^2_i$$



Residual

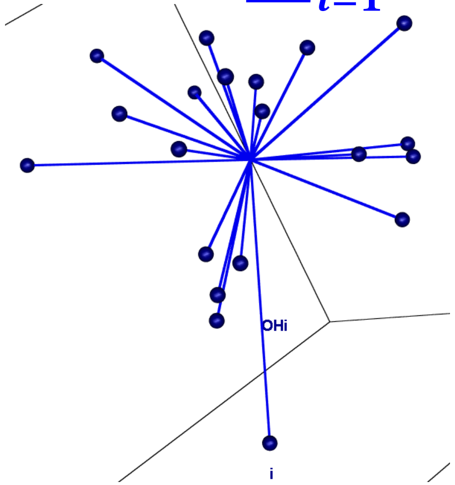
$$RSS = \sum_{i=1}^{n} HH'^2_i = TSS - ESS$$



Diagnostics R2X: How much of the original inertia is still reflected by the model?

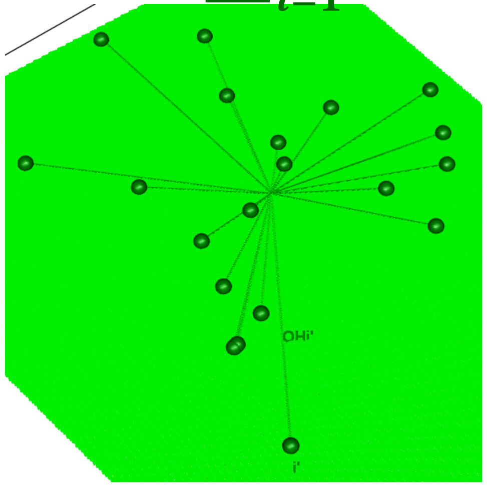
Total

$$TSS = \sum_{i=1}^{n} OH_i^2$$



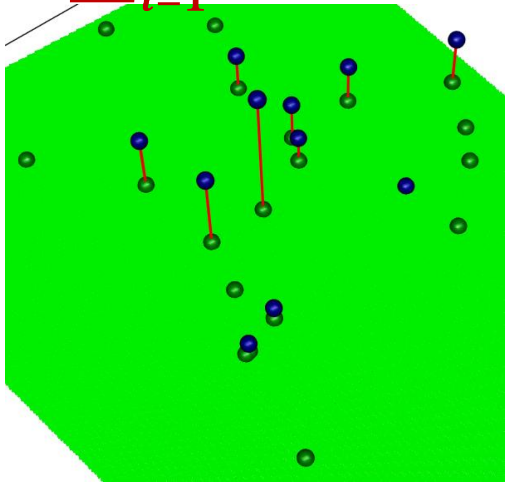
Explained

$$ESS = \sum_{i=1}^{n} OH'^2_i$$



Residual

$$RSS = \sum_{i=1}^{n} HH'^2_i = TSS - ESS$$



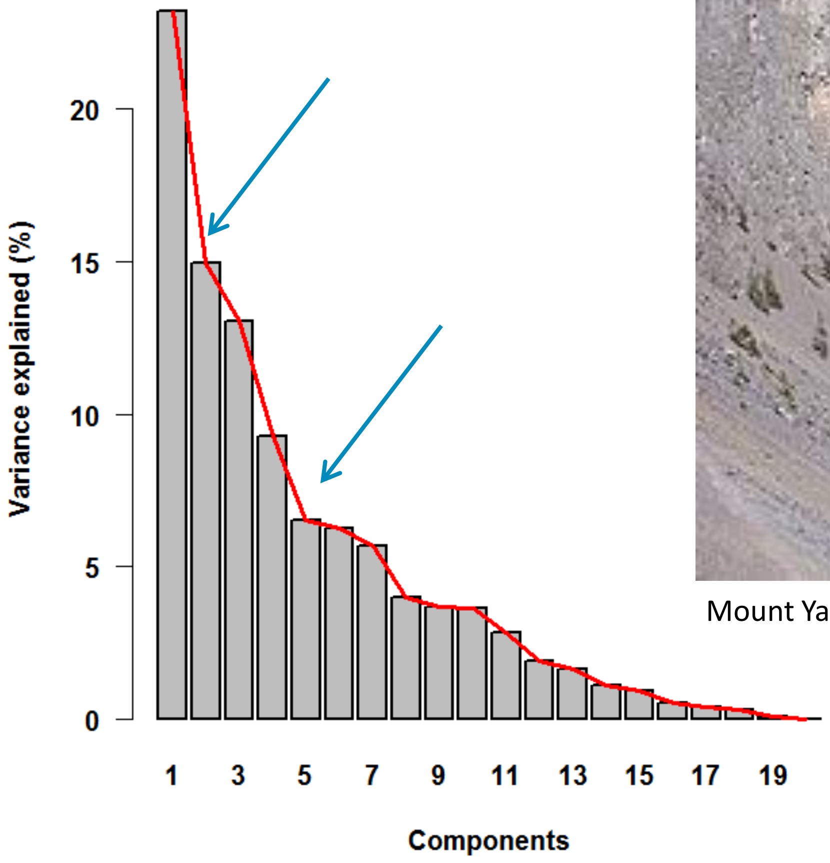
$$R2X = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- R2X increases with the number of components in the model
- For a given number of components, the higher the R2X, the more inertia is captured by the model (projection)

$$0 \leq R2X \leq 1$$

Scree plot

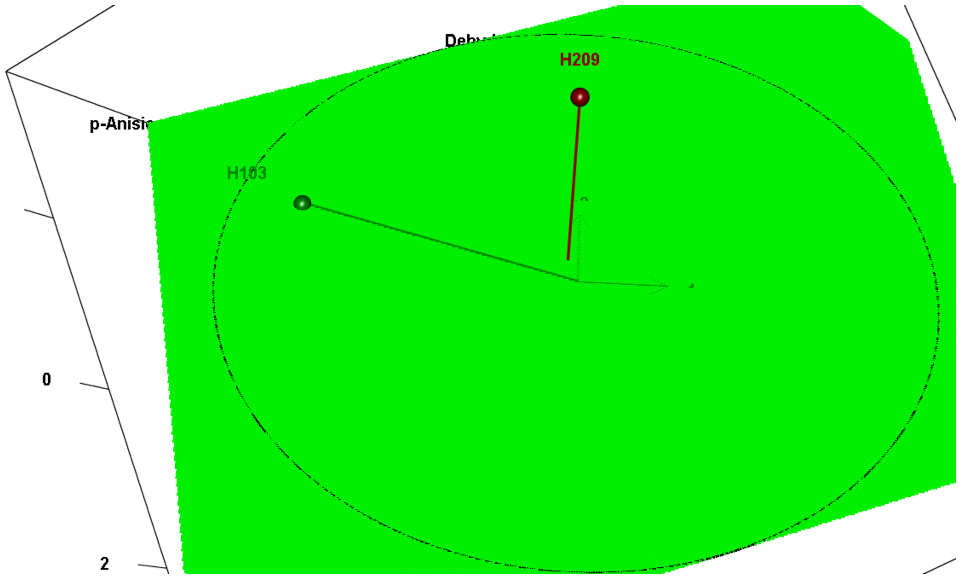
- Check that the first components capture most of the variance

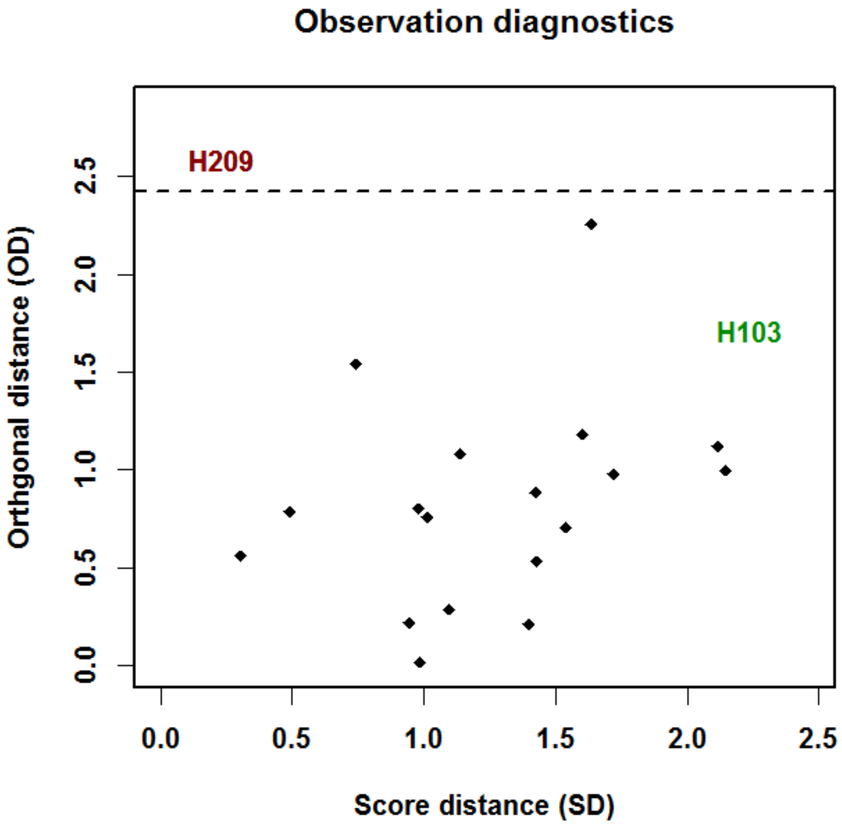


Mount Yamnuska, Alberta. Wikipedia

Observation diagnostics

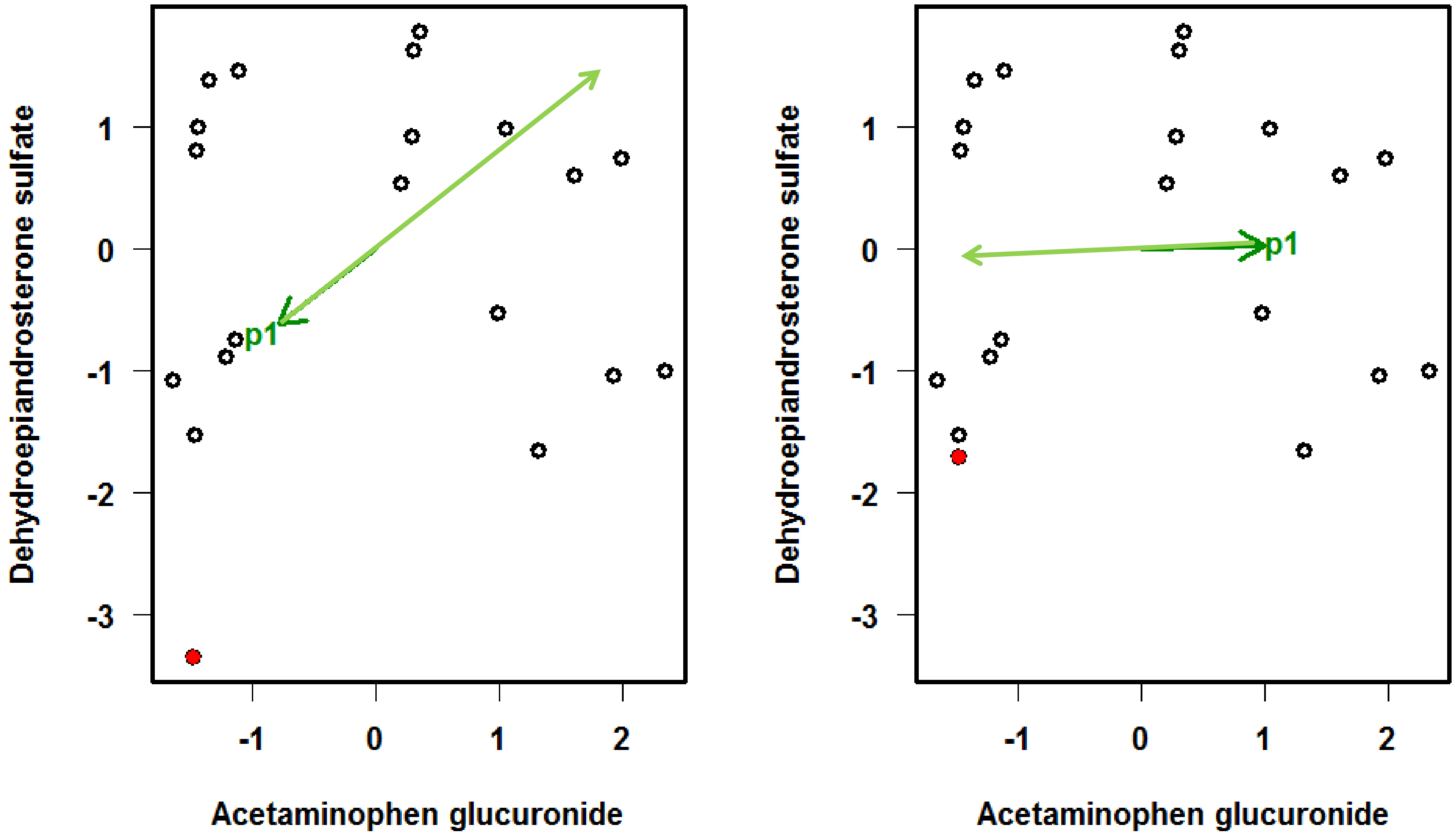
- **Samples which may bias the PCA computation and/or may not be faithfully visualized by the score plot**





Hubert M., Rousseeuw P. and Vanden Branden K. (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics,* **47:**64-79. DOI:

Sensitivity to outliers



Scaling

- **Variables are usually centered:**

$$x'_j = x_j - \bar{x}_j$$

- **In addition, variables may be**
  - unit-variance scaled (default in *ropls*):

$$x''_j = \frac{x'_j}{\sigma_j}$$

  - pareto scaled:

$$x''_j = \frac{x'_j}{\sqrt{\sigma_j}}$$

R code (*ropls* package)

Loading

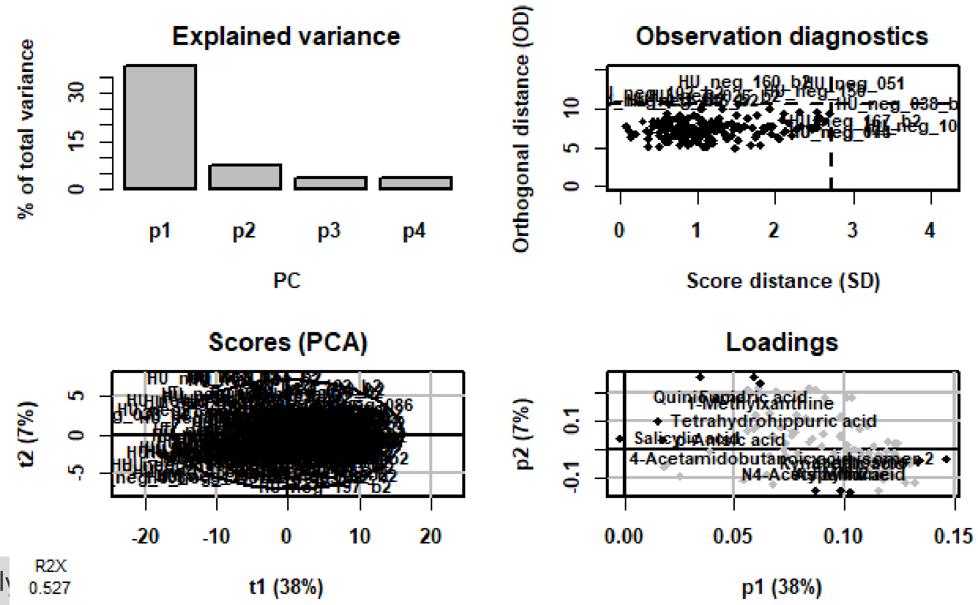
```
sacurine_dir.c <-
"C:/Users/et207099/Documents/sources/training/inst/extdata/sacurine_annotated_postprocessed"
sacurine.eset <- phenomis::reading(sacurine_dir.c)
```

Inspecting

```
sacurine.eset <- phenomis::inspecting(sacurine.eset)
```

Computing the PCA

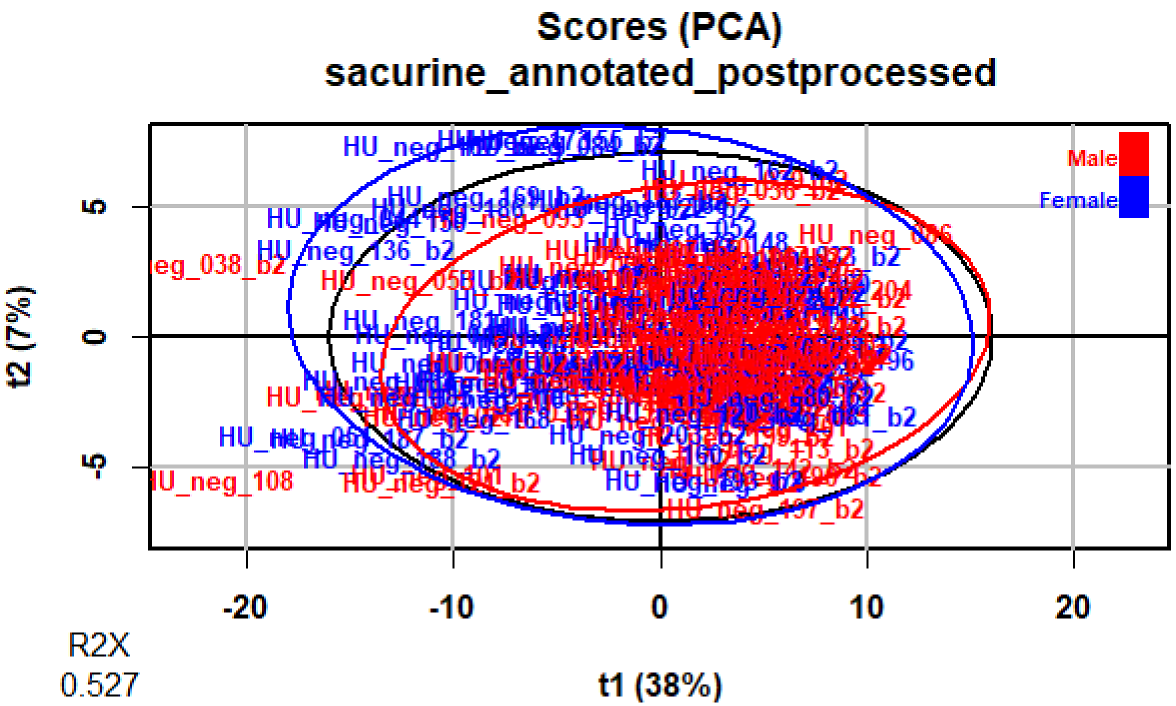
```
sacurine.pca <- ropls::opls(sacurine.eset)
```



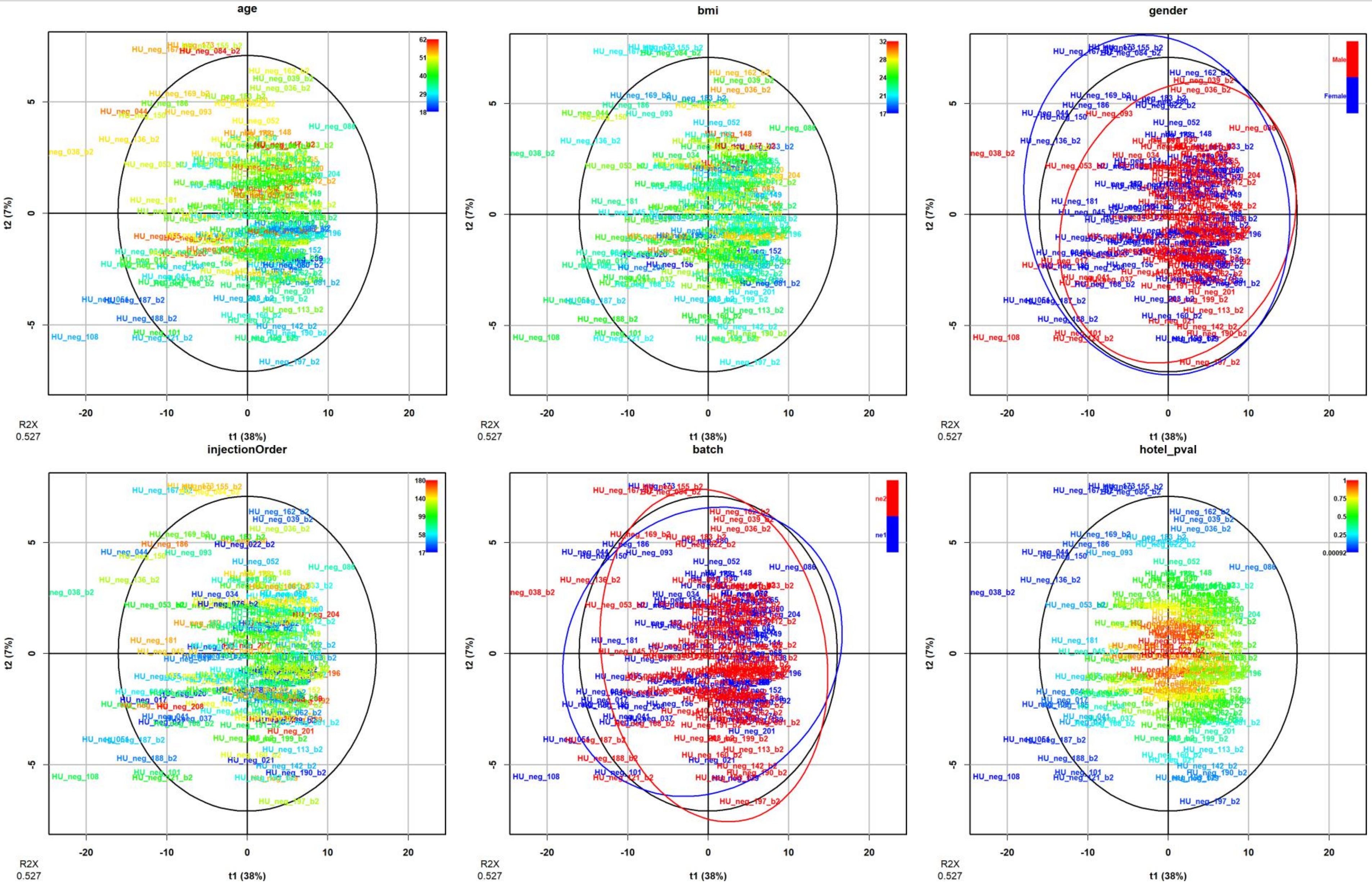
Coloring the score plot according to the metadata

- **Coloring the score plot according to ‘gender’ (column of the sampleMetadata)**

```
ropls::plot(sacurine.pca,
            typeVc = "x-score",
            parAsColFcVn = Biobase::pData(sacurine.eset)[, "gender"]))
```



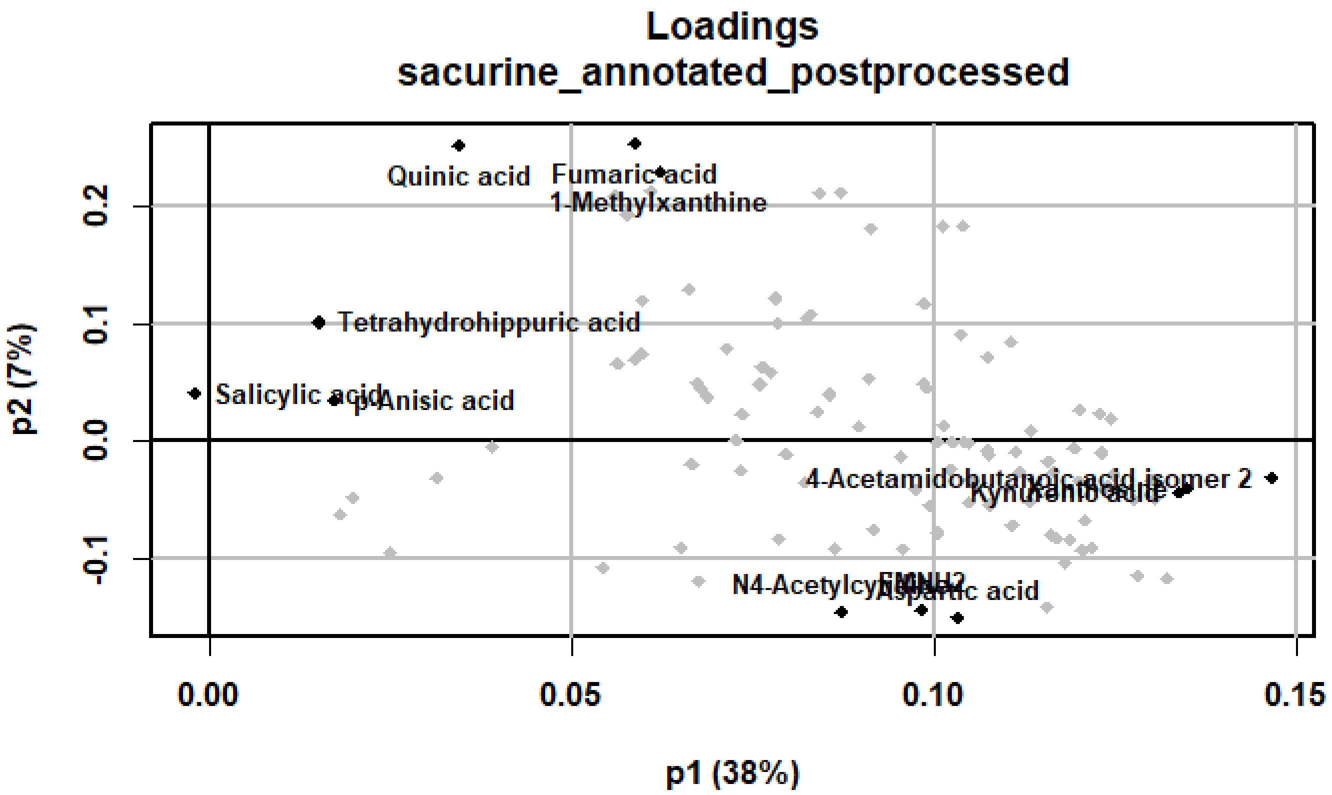
Coloring the score plot according to the metadata



Loadings

- Studying the variables which most contribute to the new components

```
ropls::plot(sacurine.pca,
            typeVc = "x-loading")
```



Getting back the ExpressionSet object

- **Getting back the ExpressionSet object**

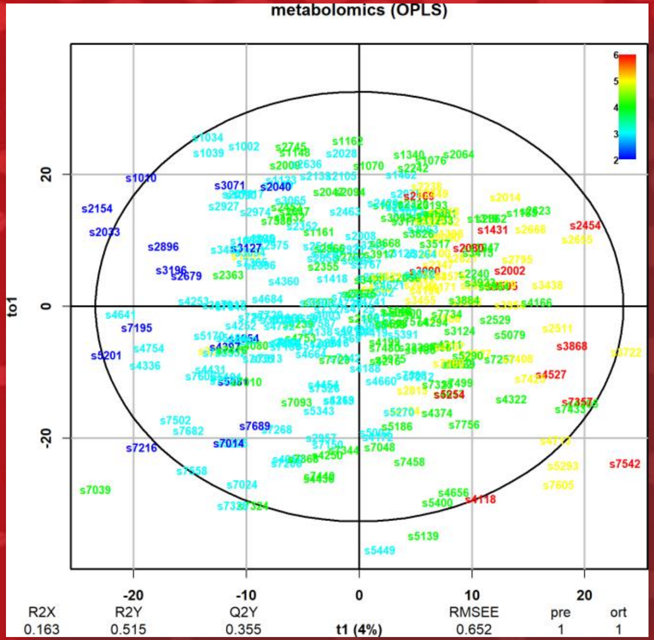
```
sacurine.eset <- ropls::getEset(sacurine.pca)
```

- **The scores and loadings values have been added to the sampleMetadata and variableMetadata:**

```
head(Biobase::pData(sacurine.eset)[, c("PCA_xscor-p1", "PCA_xscor-p2")])
head(Biobase::fData(sacurine.eset)[, c("PCA_xload-p1", "PCA_xload-p2")])
```

References

- Husson F., Le S. and Pages J. (2011). Exploratory multivariate analysis by example using *R. Chapman & Hall/CRC*
- Baccini A. (2010). Statistique Descriptive Multidimensionnelle (pour les nuls). *Institut de Mathématiques de Toulouse, Université Paul Sabatier*.
- Ringner M. (2008). What is principal component analysis? *Nature Biotechnology*, 26:303-304.
- Wehrens, R. (2011). Chemometrics with R. *Springer*. https://doi.org/10.1007/978-3-642-17841-2



Partial Least Squares (PLS)

PLS(-DA) modelling

- **Powerful regression method when**

$$n_{samples} < p_{variables}$$

- **Complementary to univariate hypothesis testing (where variables are tested independantly)**

- **Risk of overfitting: i.e., building a model whose (apparently) good performances result from chance only**

The data: the data matrix X (metabolomics measurements)

p = 109 variables (quantitatives)

n = 183 samples

	(2-methoxyethoxy)propanoic acid isomer	(gamma)Glu-Leu/Ile	1-Methyluric acid	1-Methylxanthine	1,3-Dimethyluric acid	…	Threonic acid/Erythronic acid	Tryptophan	Valerylglycine isomer 1	Valerylglycine isomer 2	Xanthosine
HU_011	3.02	3.89	3.87	3.72	3.54	…	4.31	4.01	4.02	3.89	4.08
HU_014	3.81	4.28	3.84	3.78	3.93	…	4.47	4.42	3.88	4.18	4.20
HU_015	3.52	4.20	4.10	4.29	3.96	…	4.12	4.44	4.19	4.25	4.12
HU_017	2.56	4.32	4.54	4.43	4.23	…	4.56	4.54	4.15	4.29	4.25
HU_018	3.78	4.63	4.18	4.12	4.01	…	4.45	4.22	4.10	4.14	4.36
…	…	…	…	…	…	…	…	…	…	…	…
HU_205	3.86	4.54	4.24	4.19	4.38	…	4.36	4.12	4.16	4.22	4.41
HU_206	1.32	4.34	4.62	4.61	4.82	…	4.27	4.04	3.93	4.28	4.39
HU_207	4.19	4.28	4.48	4.46	4.45	…	4.64	4.00	4.74	4.65	4.26
HU_208	3.75	4.52	4.36	4.36	4.23	…	4.70	4.69	4.44	4.63	4.49
HU_209	4.21	4.68	4.19	4.21	4.15	…	4.52	4.50	4.47	4.47	4.22

$\mathbf{X}$

The data: the response to be explained y (e.g. body mass index

1 response

	IMC
HU_011	19.75
HU_014	22.64
HU_015	22.72
HU_017	23.03
HU_018	20.96
...	…
HU_205	28.37
HU_206	22.15
HU_207	19.47
HU_208	18.61
HU_209	21.48

$\mathbf{y}$

The goal: to learn a predictive model f ...

p = 109 variables (quantitatives)

1 response

n = 183 sample

	(2-methoxyethoxy)propanoic acid isomer	(gamma)Glu-Leu/Ile	1-Methyluric acid	1-Methylxanthine	1,3-Dimethyluric acid	...	Threonic acid/Erythronic acid	Tryptophan	Valerylglycine isomer 1	Valerylglycine isomer 2	Xanthosine
HU_011	3.02	3.89	3.87	3.72	3.54	...	4.31	4.01	4.02	3.89	4.08
HU_014	3.81	4.28	3.84	3.78	3.93	...	4.47	4.42	3.88	4.18	4.20
HU_015	3.52	4.20	4.10	4.29	3.96	...	4.12	4.44	4.19	4.25	4.12
HU_017	2.56	4.32	4.54	4.43	4.23	...	4.56	4.54	4.15	4.29	4.25
HU_018	3.78	4.63	4.18	4.12	4.01	...	4.45	4.22	4.10	4.14	4.36
...	...	...	...	...	...	...	...	...	...	...	...
HU_205	3.86	4.54	4.24	4.19	4.38	...	4.36	4.12	4.16	4.22	4.41
HU_206	1.32	4.34	4.62	4.61	4.82	...	4.27	4.04	3.93	4.28	4.39
HU_207	4.19	4.28	4.48	4.46	4.45	...	4.64	4.00	4.74	4.65	4.26
HU_208	3.75	4.52	4.36	4.36	4.23	...	4.70	4.69	4.44	4.63	4.49
HU_209	4.21	4.68	4.19	4.21	4.15	...	4.52	4.50	4.47	4.47	4.22

	IMC
HU_011	19.75
HU_014	22.64
HU_015	22.72
HU_017	23.03
HU_018	20.96
...	...
HU_205	28.37
HU_206	22.15
HU_207	19.47
HU_208	18.61
HU_209	21.48

$$f(\mathbf{X}) = \mathbf{y}$$

... which can be used to predict the y' values for new samples X'

p = 109 variables (quantitatives)

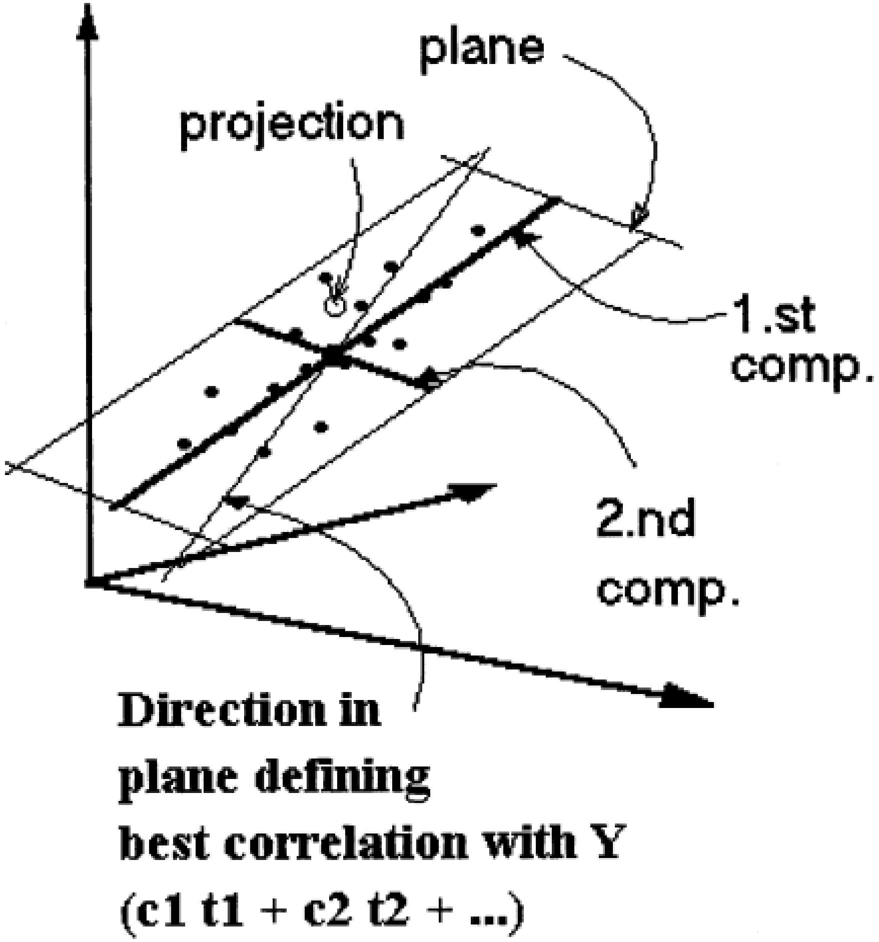
n' samples

	(2-methoxyethoxy)propanoic acid isomer	(gamma)Glu-Leu/Ile	1-Methyluric acid	1-Methylxanthine	1,3-Dimethyluric acid	...	Threonic acid/Erythronic acid	Tryptophan	Valerylglycine isomer 1	Valerylglycine isomer 2	Xanthosine
new 1	3.00	4.47	4.54	4.54	4.62	...	4.46	4.30	4.44	4.41	4.54
new 2	3.48	4.20	3.73	3.31	3.44	...	4.57	4.17	4.15	4.16	4.26
new 3	4.03	2.55	4.27	4.23	4.34	...	4.26	3.58	4.07	3.96	4.15

	IMC
new 1	?
new 2	?
new 3	?

$$f(X') = y'$$

PLS regression: building latent variables with maximal covariance with the response



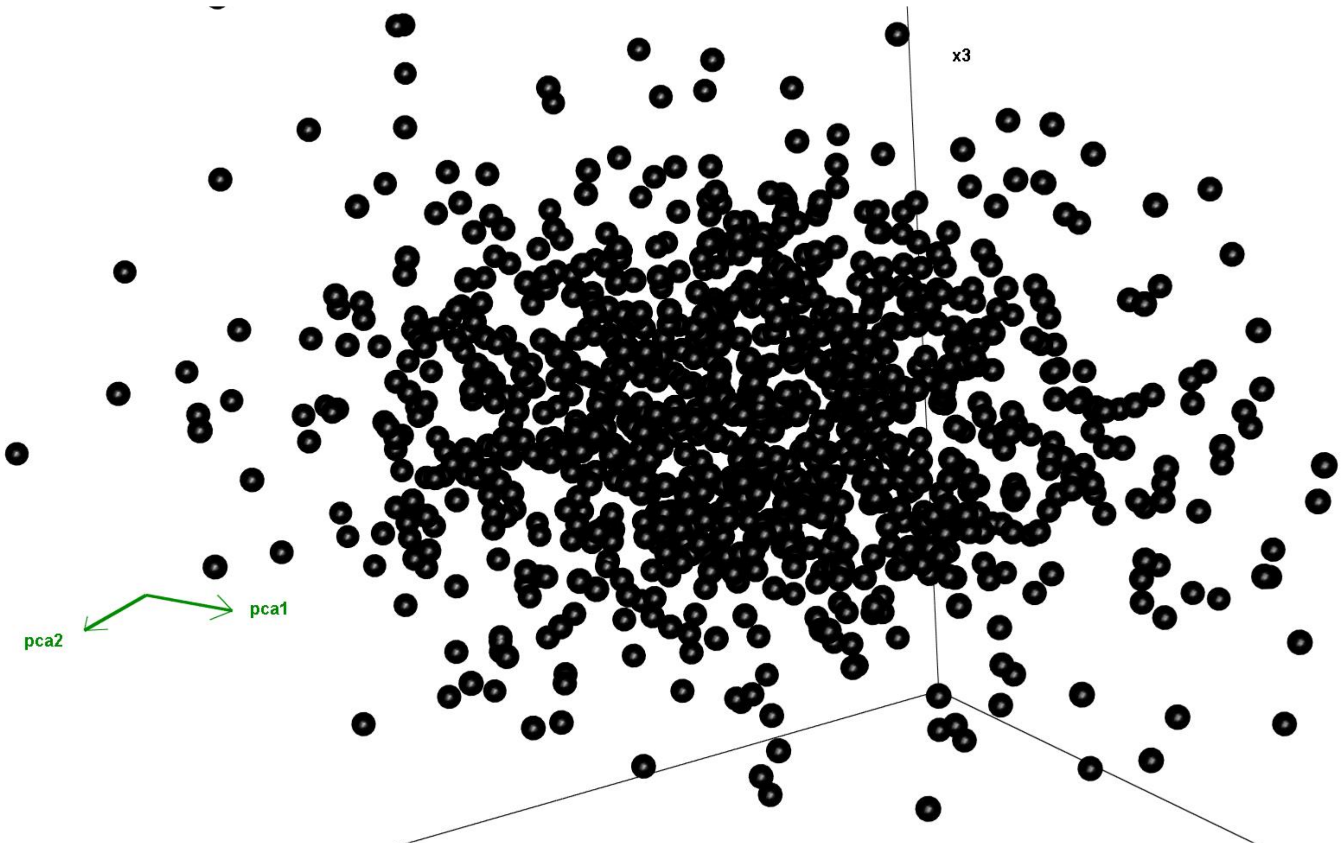
$$cov(x,y) = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{\mathbf{1}}{(\boldsymbol{n}-\mathbf{1})} \boldsymbol{cor}(\boldsymbol{x},\boldsymbol{y}) \, \|\boldsymbol{x}\| \|\boldsymbol{y}\|$$

Wold et al. (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58:109-130.

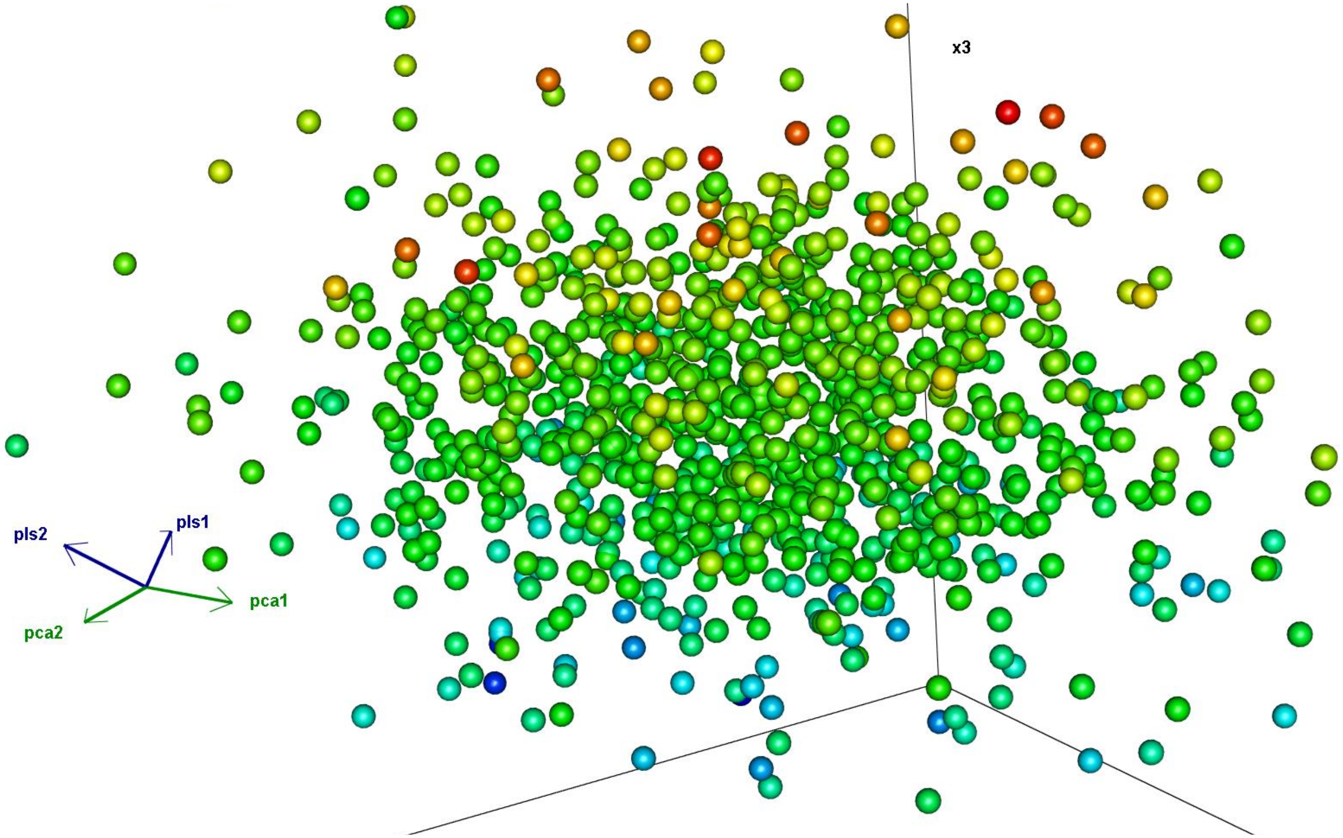
PLS vs PCA

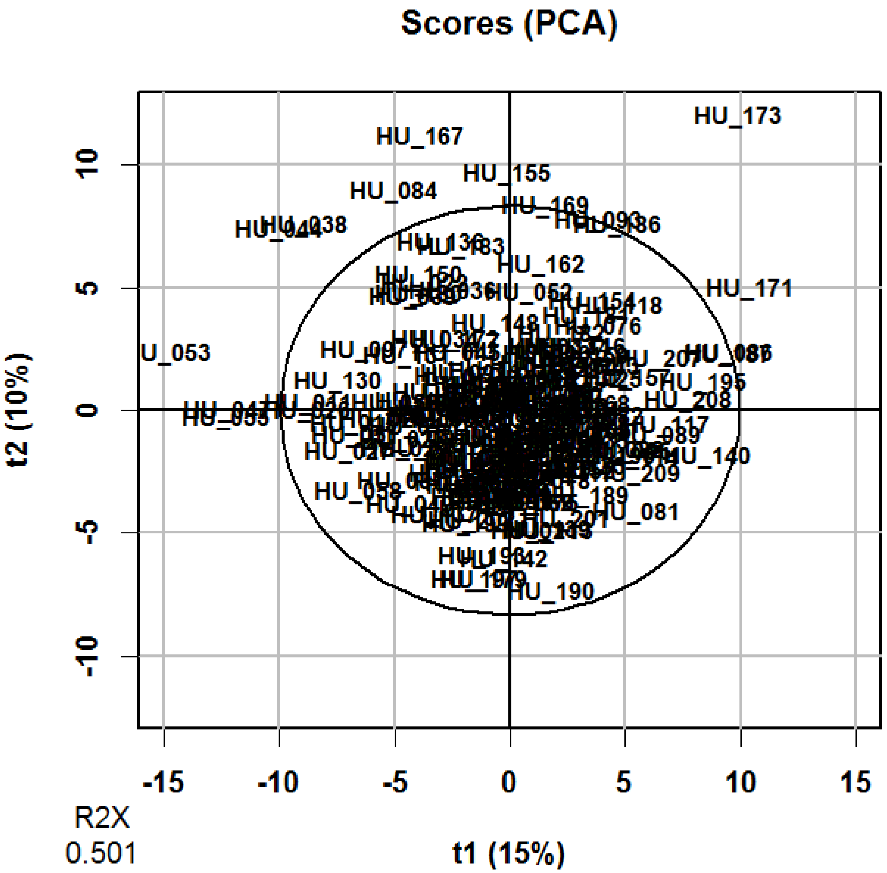
- **PCA finds the directions of maximum variance**

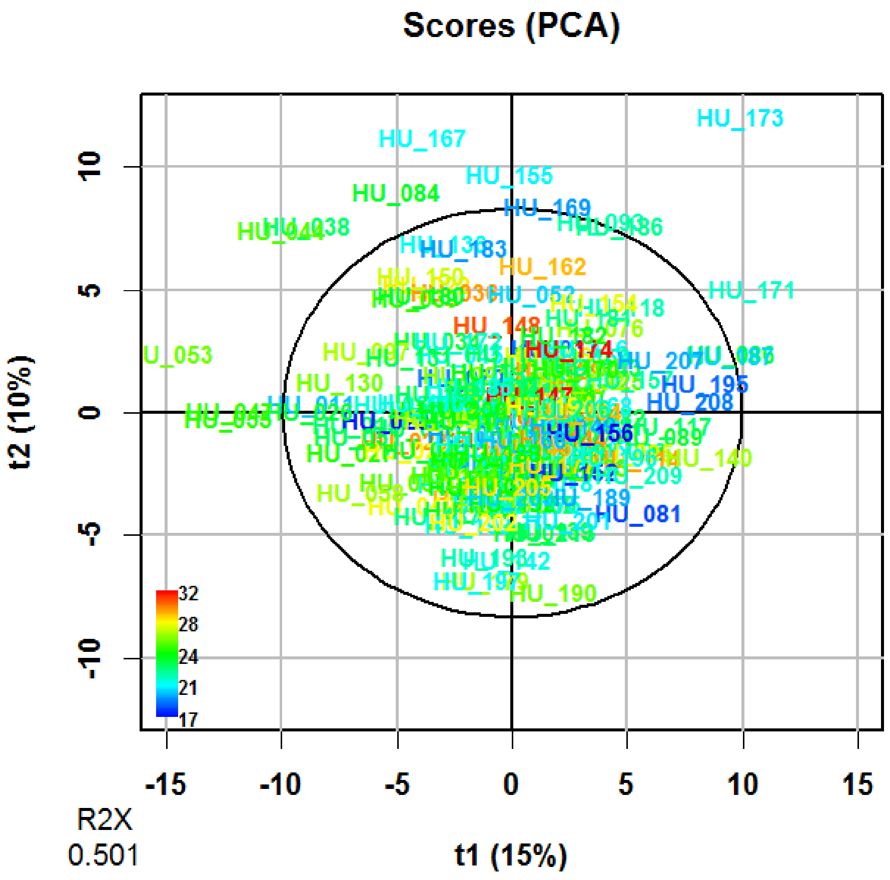


PLS vs PCA

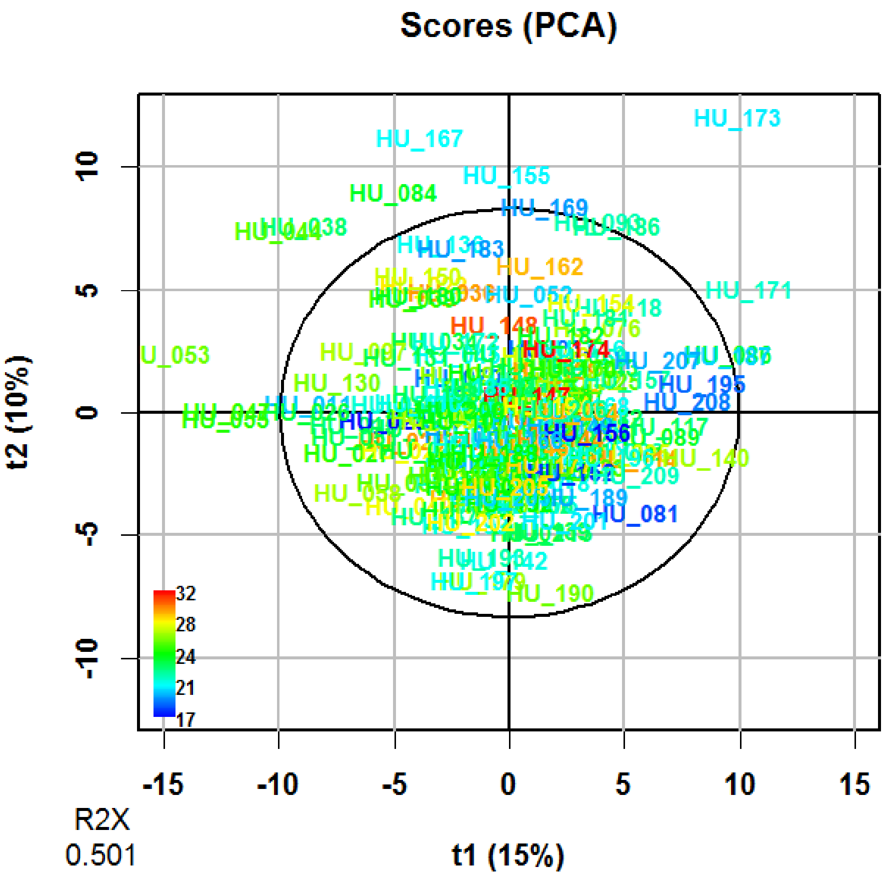
- **PLS includes the labels into the model**



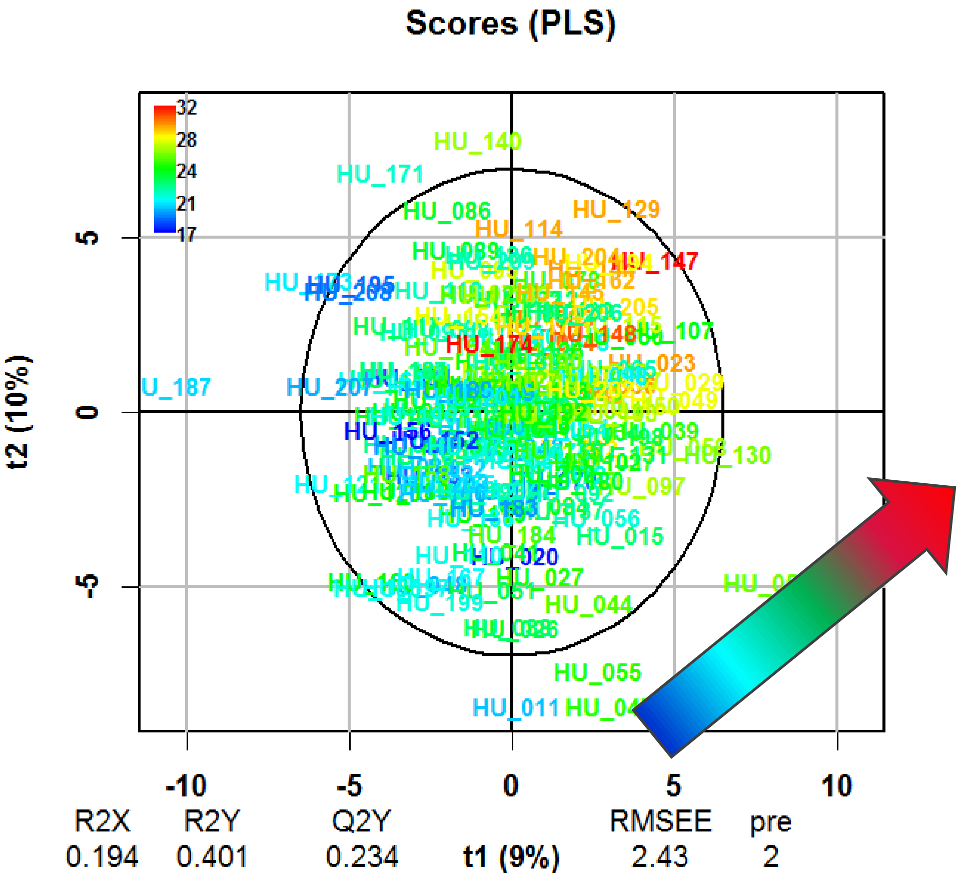

Scores (PCA)
HU_173
HU_167
HU_155
HU_084
HU_169
HU_171
HU_162
HU_053
HU_130
HU_208
HU_195
HU_140
HU_209
HU_189
HU_081
HU_190
t2 (10%)
10
5
0
-5
-10
-15
-10
-5
0
5
10
15
t1 (15%)
R2X
0.501


Scores (PCA)
HU_173
HU_167
HU_155
HU_084
HU_169
HU_038
HU_044
HU_183
HU_162
HU_150
HU_171
HU_052
HU_148
HU_174
HU_207
HU_195
HU_208
U_053
HU_130
HU_055
HU_156
HU_140
HU_209
HU_189
HU_081
HU_142
HU_190
32
28
24
21
17
t2 (10%)
10
5
0
-5
-10
-15
-10
-5
0
5
10
15
R2X
0.501
t1 (15%)

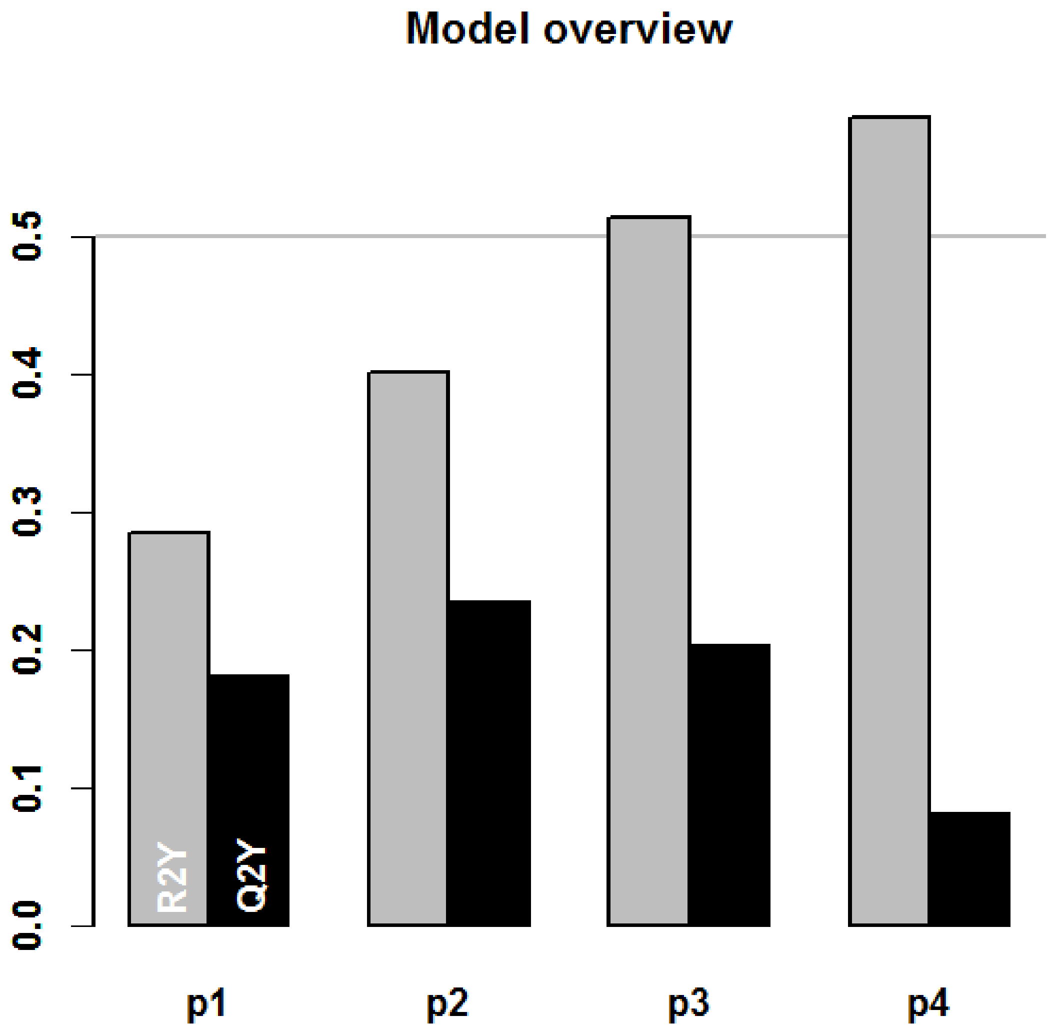
Scores (PCA)



Scores (PLS)

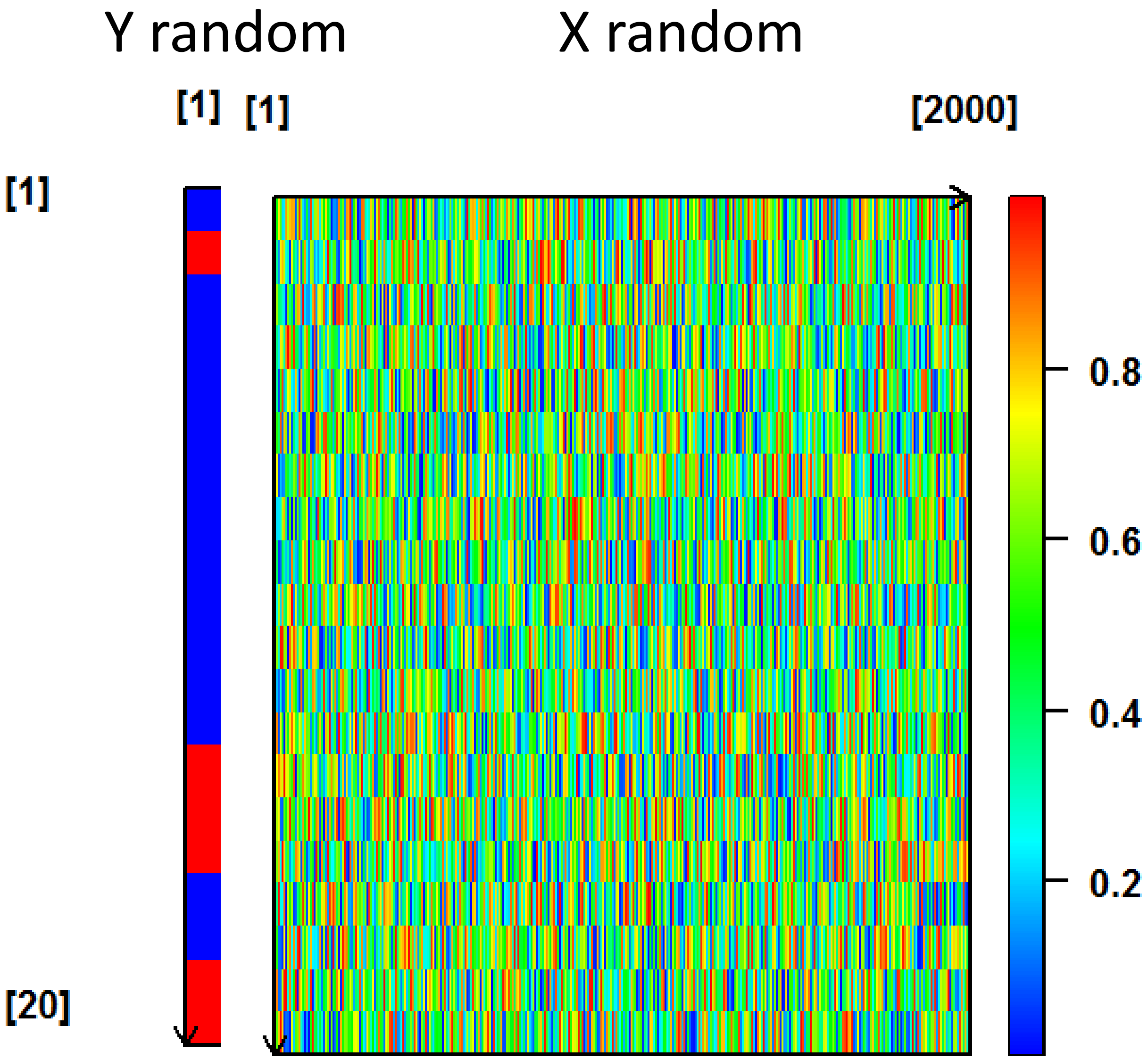


R2Y vs Q2Y



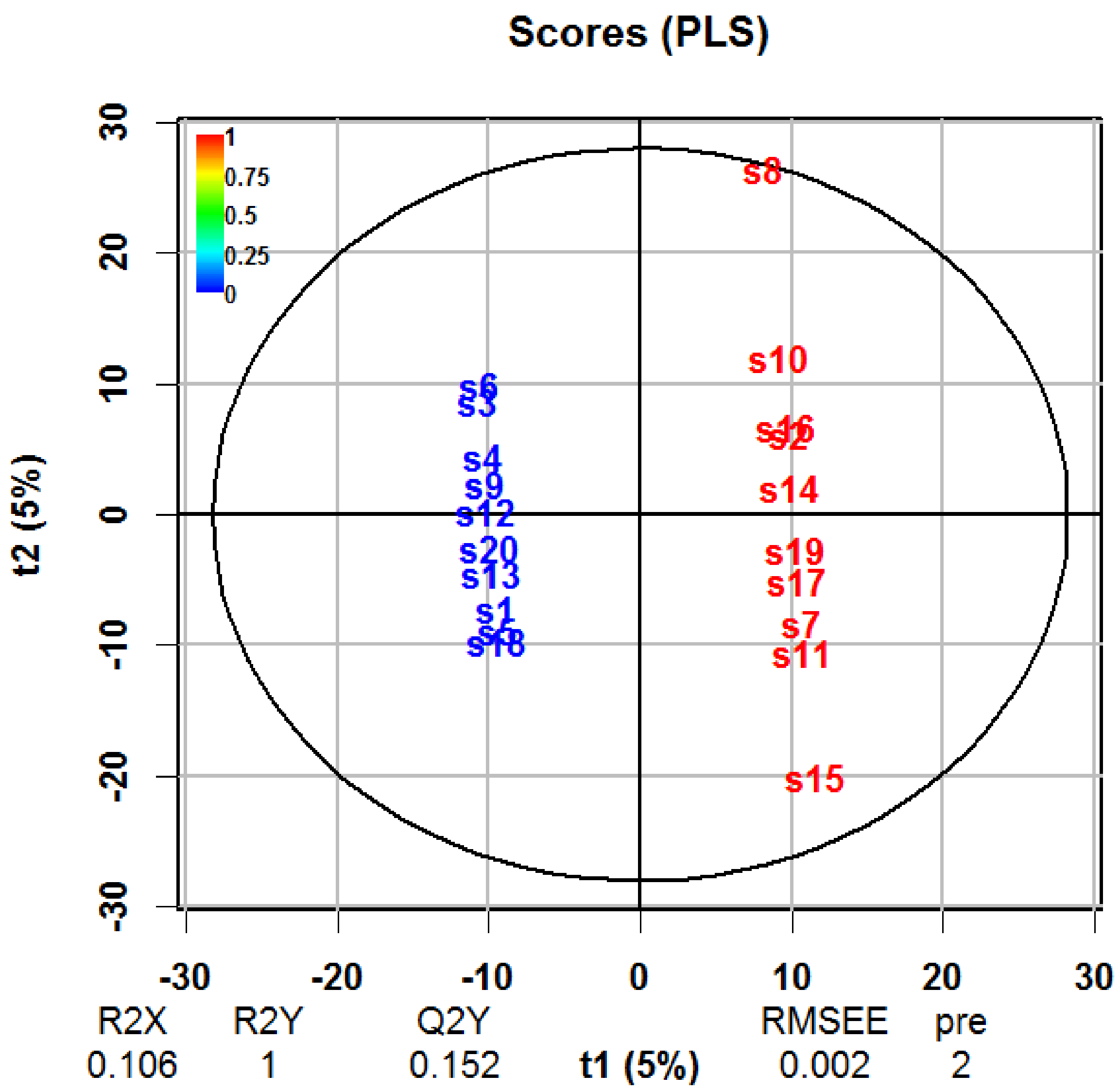
Overfitting

- **X: 20 × 2,000 matrix of random numbers**
  - Uniform distribution between 0 and 1
- **Y: 20 x 1 matrix of random labels**
  - 0 or 1 values



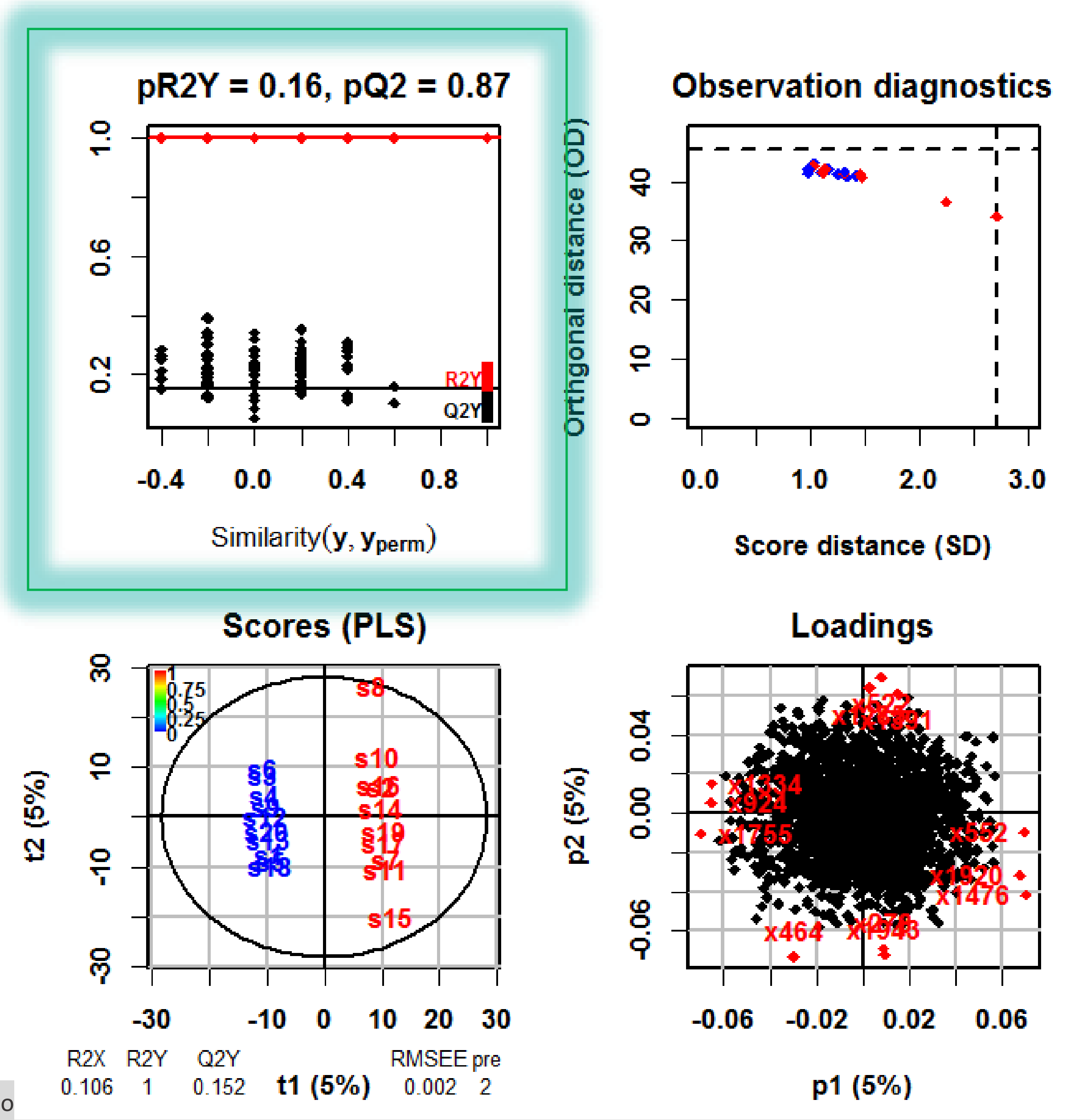
adpated from Wehrens (2011). Chemometrics with R. Springer.

Score plot!


Scores (PLS)
1
0.75
0.5
0.25
0
30
20
10
0
-10
-20
-30
t2 (5%)
s8
s10
s16
s2
s14
s19
s17
s7
s11
s15
s6
s3
s4
s9
s12
s20
s13
s1
s5
s18
-30
-20
-10
0
10
20
30
t1 (5%)
R2X
0.106
R2Y
1
Q2Y
0.152
RMSEE
0.002
pre
2

Importance of diagnostics

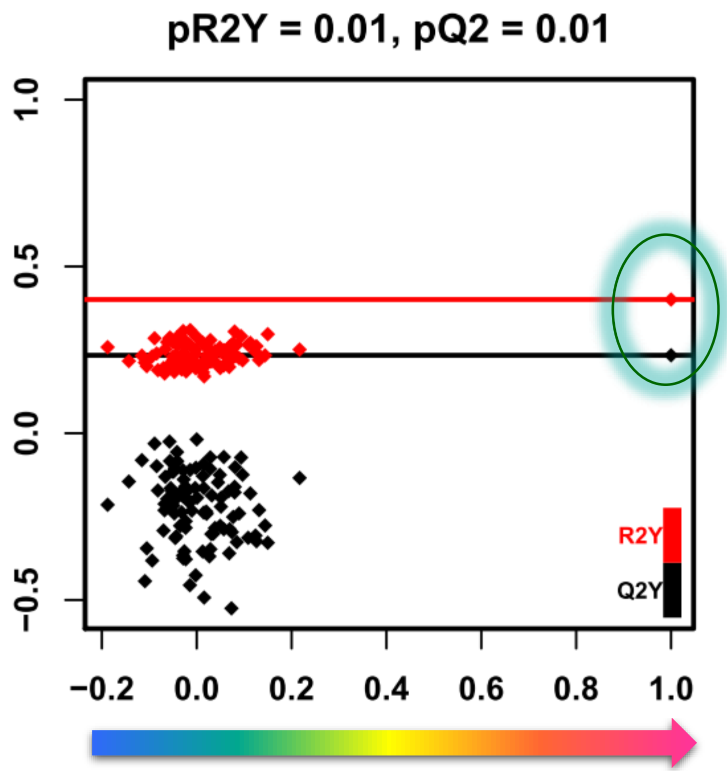
- **Permutation testing: comparing the R2Y and Q2Y values of the model built with the true Y labels with n_{perm} models built with random permutation of Y labels**



Szymanska E., Saccenti E., Smilde A. and Westerhuis J. (2012). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics,* **8:**3-16. DOI:

Significance of the model

- **Counting the number of $R2Y$ (and $Q2Y$) metrics from random models which are superior to the values of the true model gives an indication of the significance of the PLS modelling**



$R2Y$ and $Q2Y$ of the model with the true $\mathbf{y}$ values

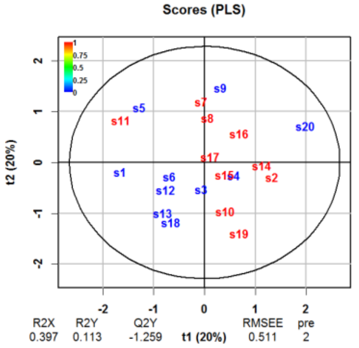
'permutation'

Similarity between $\mathbf{y}_{true}$ and $\mathbf{y}_{random}$

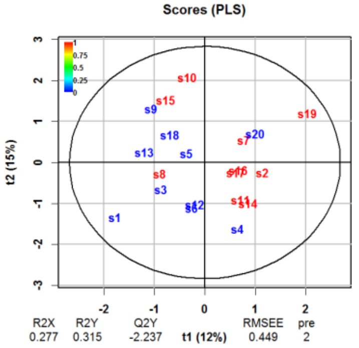
Risk of overfitting when *n* < *p*

$$\frac{\textbf{variables}}{\textbf{samples}} =$$

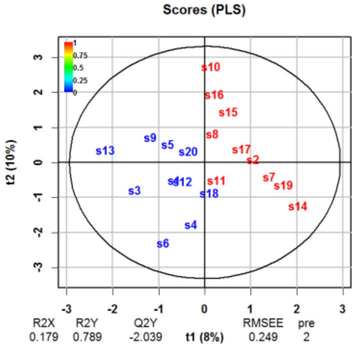
0.2



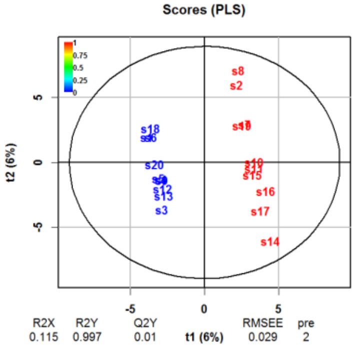
0.5



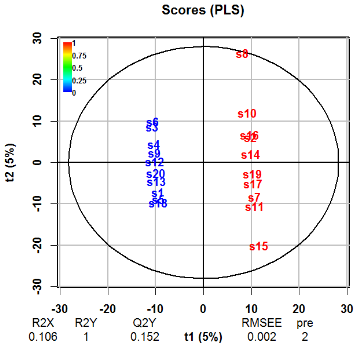
1

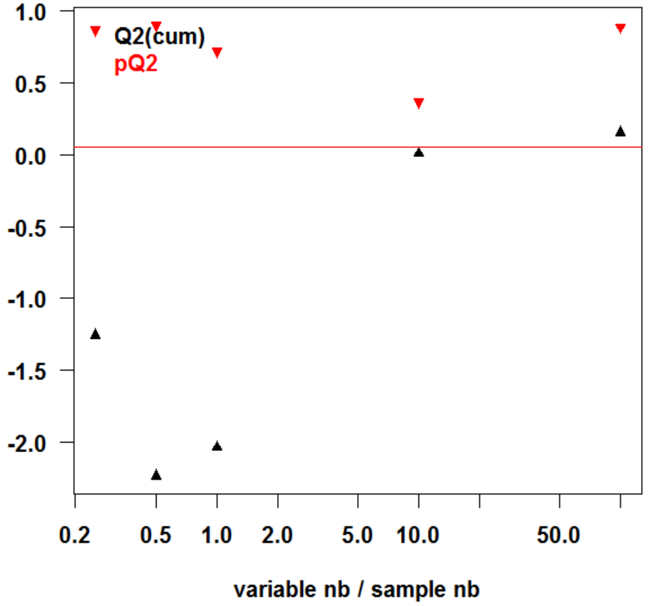


10



100





References

- Wold S., Sjöström M. and Eriksson L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems,* **58:**109-130.

- Trygg J., Holmes E. and Lundstedt T. (2007). Chemometrics in Metabonomics. *Journal of Proteome Research*, **6**:469-479.

- Brereton R.G. and Lloyd G.R. (2014). Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics,* **28:**213-225.

Significance of the model

	(2-methoxyethoxy)propanoic acid isomer	(gamma)Glu-Leu/Ile	1-Methyluric acid	1-Methylxanthine	1,3-Dimethyluric acid	…	Threonic acid/Erythronic acid	Tryptophan	Valerylglycine isomer 1	Valerylglycine isomer 2	Xanthosine
HU_011	3.02	3.89	3.87	3.72	3.54	…	4.31	4.01	4.02	3.89	4.08
HU_014	3.81	4.28	3.84	3.78	3.93	…	4.47	4.42	3.88	4.18	4.20
HU_015	3.52	4.20	4.10	4.29	3.96	…	4.12	4.44	4.19	4.25	4.12
HU_017	2.56	4.32	4.54	4.43	4.23	…	4.56	4.54	4.15	4.29	4.25
HU_018	3.78	4.63	4.18	4.12	4.01	…	4.45	4.22	4.10	4.14	4.36
…	…	…	…	…	…	…	…	…	…	…	…
HU_205	3.86	4.54	4.24	4.19	4.38	…	4.36	4.12	4.16	4.22	4.41
HU_206	1.32	4.34	4.62	4.61	4.82	…	4.27	4.04	3.93	4.28	4.39
HU_207	4.19	4.28	4.48	4.46	4.45	…	4.64	4.00	4.74	4.65	4.26
HU_208	3.75	4.52	4.36	4.36	4.23	…	4.70	4.69	4.44	4.63	4.49
HU_209	4.21	4.68	4.19	4.21	4.15	…	4.52	4.50	4.47	4.47	4.22

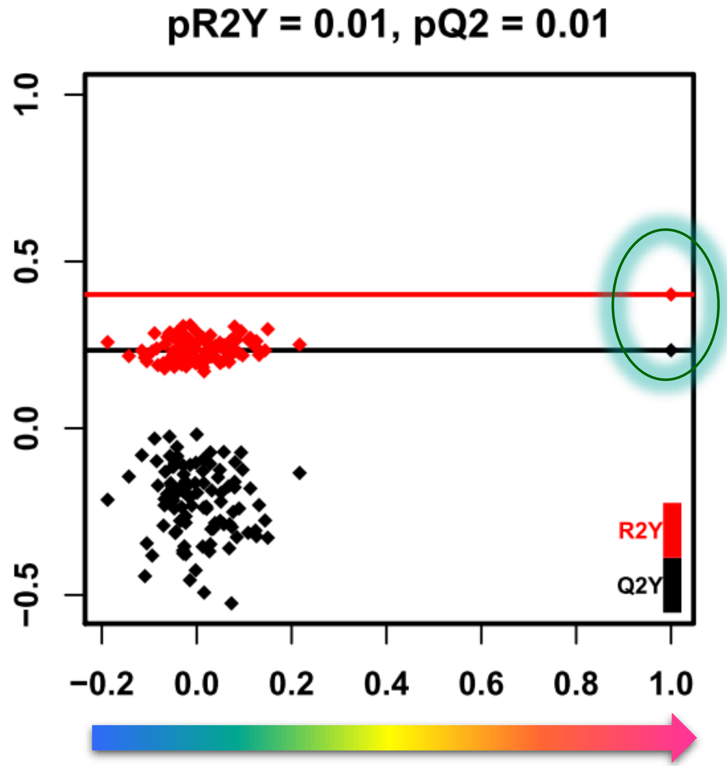
$\mathbf{X}$

	bmi
HU_067	19.31
HU_168	20.76
HU_090	19.79
HU_185	21.09
HU_195	18.42
…	…
HU_127	25.62
HU_105	22.96
HU_126	24.7
HU_081	17.84
HU_014	22.64

y_{random}

Significance of the model

- **Counting the number of $R2Y$ (and $Q2Y$) metrics from random models which are superior to the values of the true model gives an indication of the significance of the PLS modelling**



$R2Y$ and $Q2Y$ of the model with the true $\mathbf{y}$ values

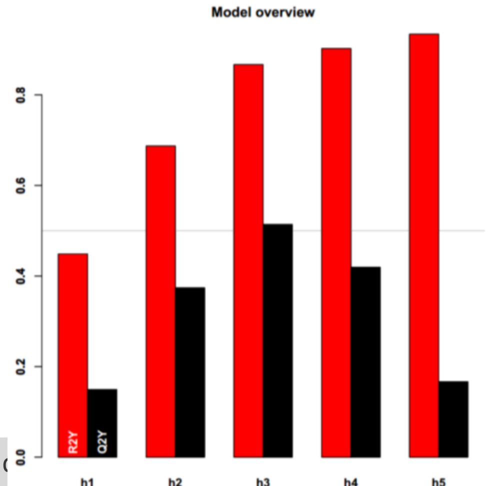
'permutation'

Similarity between $\mathbf{y}_{true}$ and $\mathbf{y}_{random}$

Diagnostic metrics

- $0 \leq R2X \leq 1$: **percentage of X inertia explained by the model**
- $0 \leq R2Y \leq 1$: **percentage of Y inertia explained by the model**
- $0 \leq Q2Y \leq 1$: **estimation of the predictive performance of the model by cross-validation**
- $R2X$ **and** $R2Y$ **increase with the number of components while** $Q2Y$ **reaches a maximum (due to overfitting):**

'overview'



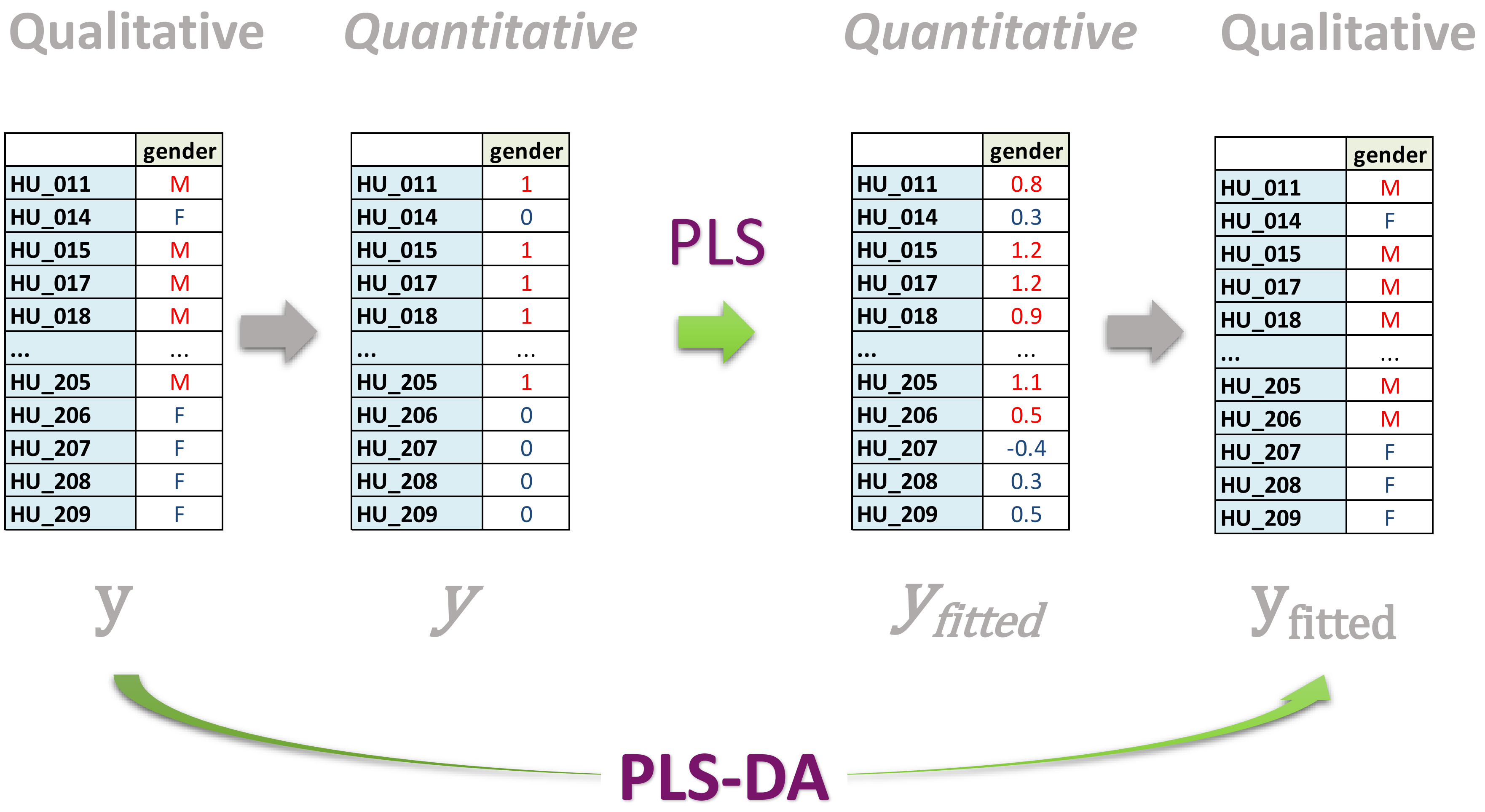
Partial Least Squares – Discriminant Analysis (PLS-DA)

Regression and classification

Response y	Example	Approach	PLS method
Quantitative	BMI	regression	PLS
Qualitative	gender	classification	PLS-DA

PLS-DA

▶ The two response levels are encoded as numbers



Qualitative

	gender
HU_011	M
HU_014	F
HU_015	M
HU_017	M
HU_018	M
...	...
HU_205	M
HU_206	F
HU_207	F
HU_208	F
HU_209	F

y

Quantitative

	gender
HU_011	1
HU_014	0
HU_015	1
HU_017	1
HU_018	1
...	...
HU_205	1
HU_206	0
HU_207	0
HU_208	0
HU_209	0

y

PLS

Quantitative

	gender
HU_011	0.8
HU_014	0.3
HU_015	1.2
HU_017	1.2
HU_018	0.9
...	...
HU_205	1.1
HU_206	0.5
HU_207	-0.4
HU_208	0.3
HU_209	0.5

y_{fitted}

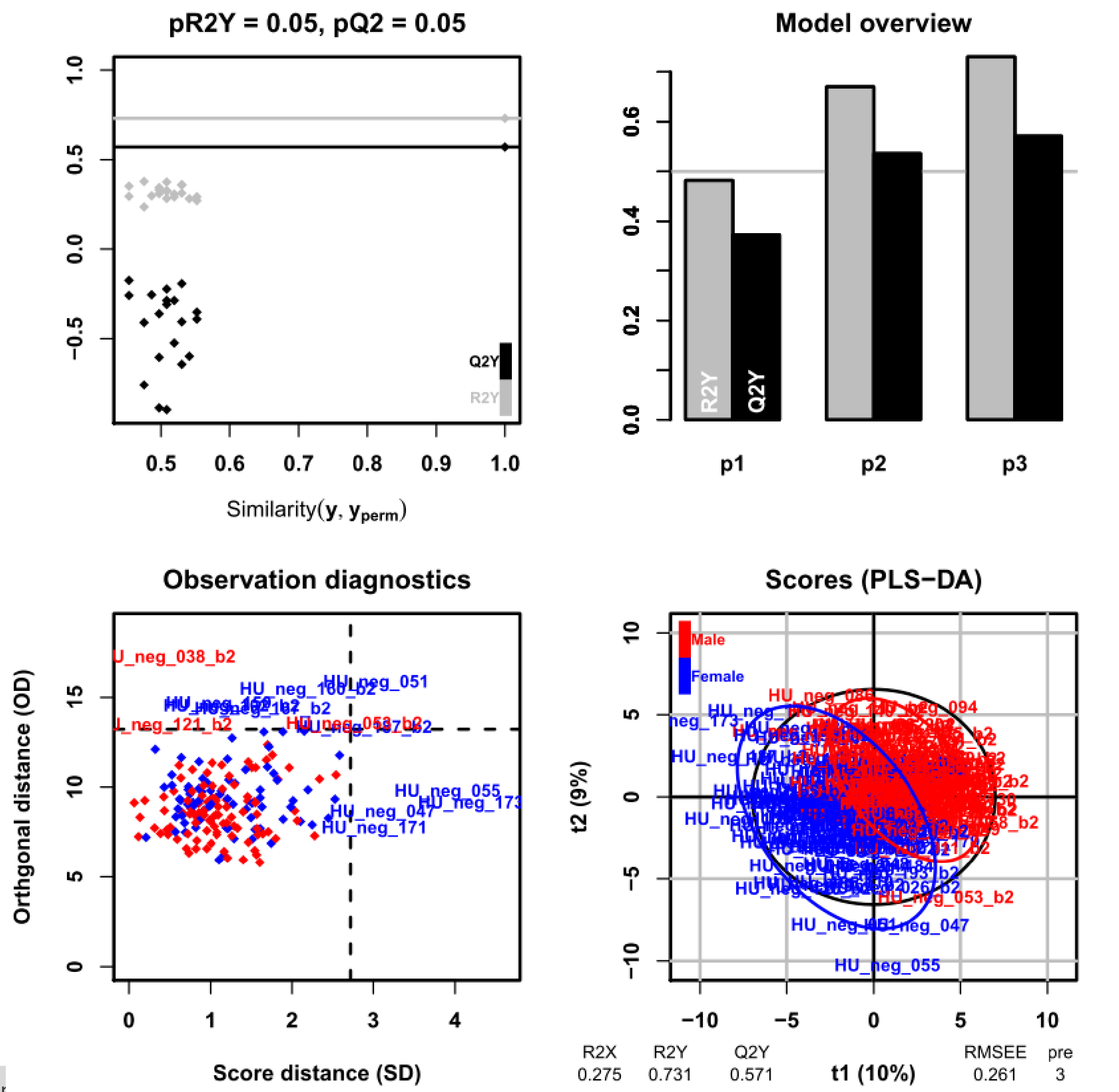
Qualitative

	gender
HU_011	M
HU_014	F
HU_015	M
HU_017	M
HU_018	M
...	...
HU_205	M
HU_206	M
HU_207	F
HU_208	F
HU_209	F

y_{fitted}

PLS-DA

PLS-DA


pR2Y = 0.05, pQ2 = 0.05
1.0
0.5
0.0
−0.5
Q2Y
R2Y
0.5 0.6 0.7 0.8 0.9 1.0
Similarity(y, y_perm)
Model overview
0.0
0.2
0.4
0.6
R2Y
Q2Y
p1
p2
p3
Observation diagnostics
Orthgonal distance (OD)
Score distance (SD)
0 1 2 3 4
0 5 10 15
U_neg_038_b2
L_neg_121_b2
HU_neg_051
HU_neg_055
HU_neg_047
HU_neg_171
HU_neg_173
Scores (PLS−DA)
Male
Female
t2 (9%)
t1 (10%)
−10 −5 0 5 10
HU_neg_094
HU_neg_053_b2
HU_neg_047
HU_neg_055
R2X
0.275
R2Y
0.731
Q2Y
0.571
RMSEE
0.261
pre
3

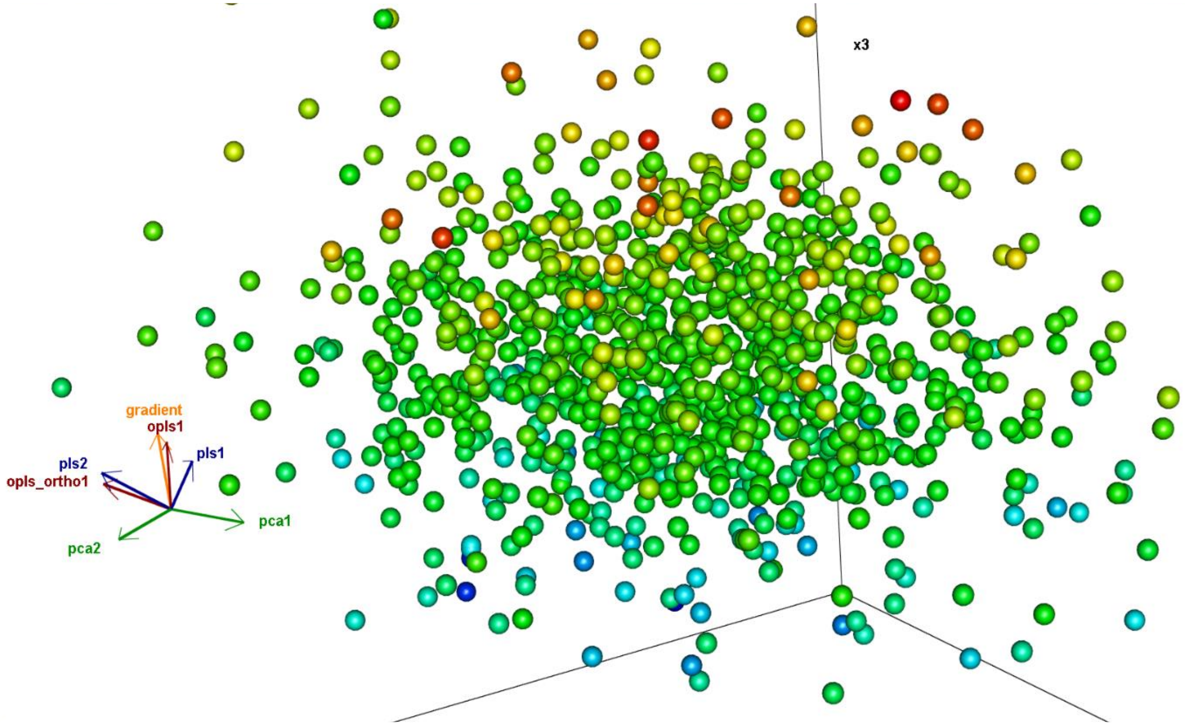
Orthogonal Partial Least Squares regression (OPLS) and Discriminant Analysis (OPLS-DA)

- Separately models the variations of the predictors correlated and orthogonal to the response
- Improves the interpretation of the components but not the overall predictive performance of the model
- Only one predictive component required for single response models
- Note: As with PLS, care should be taken to avoid too many (orthogonal) components (which would result in overfitting)

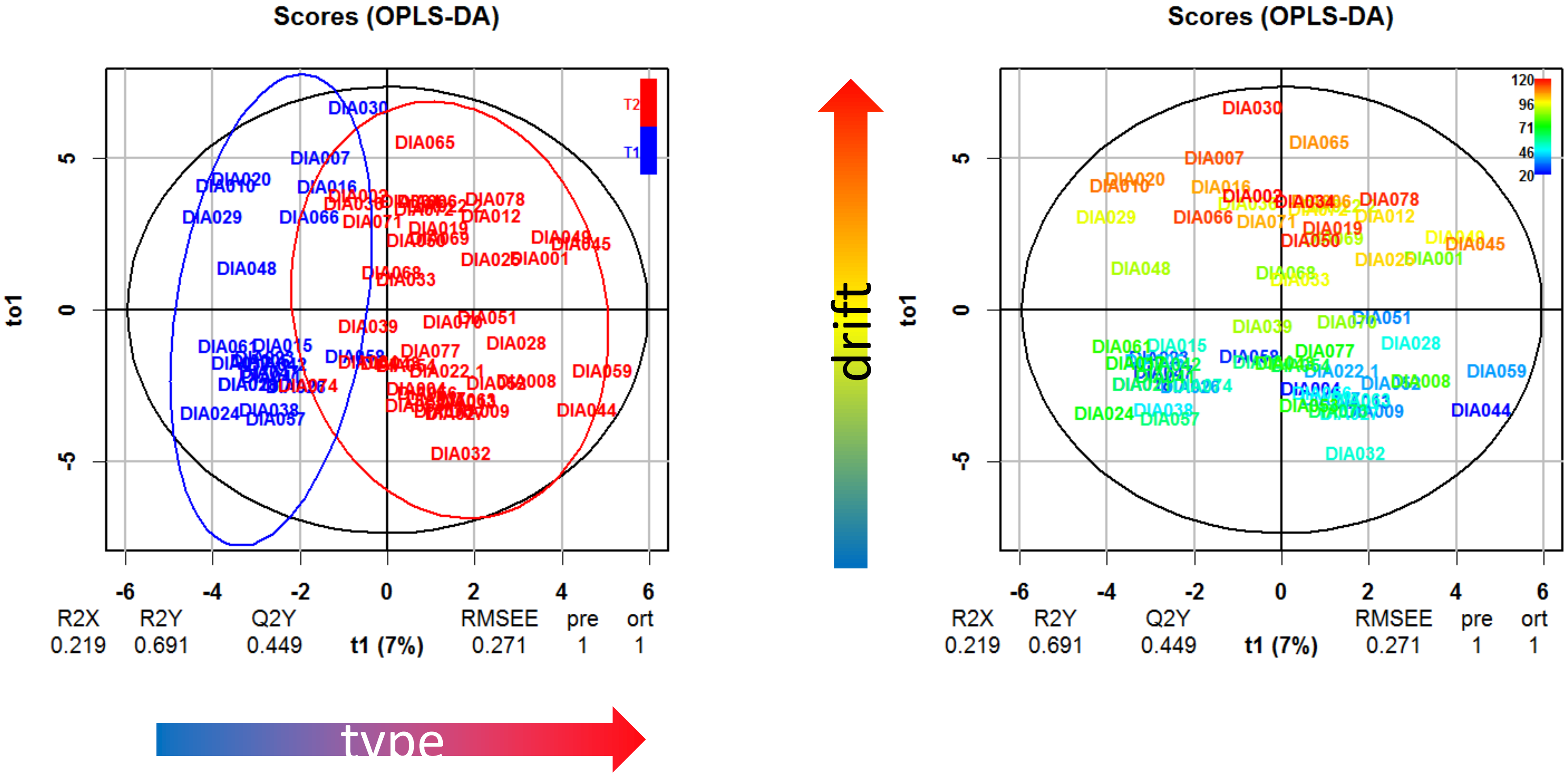
OPLS vs PLS

- Variation not correlated to the response (e.g., technical bias) is modelled separately by the orthogonal component(s)

=> The first predictive component is strongly correlated to the response



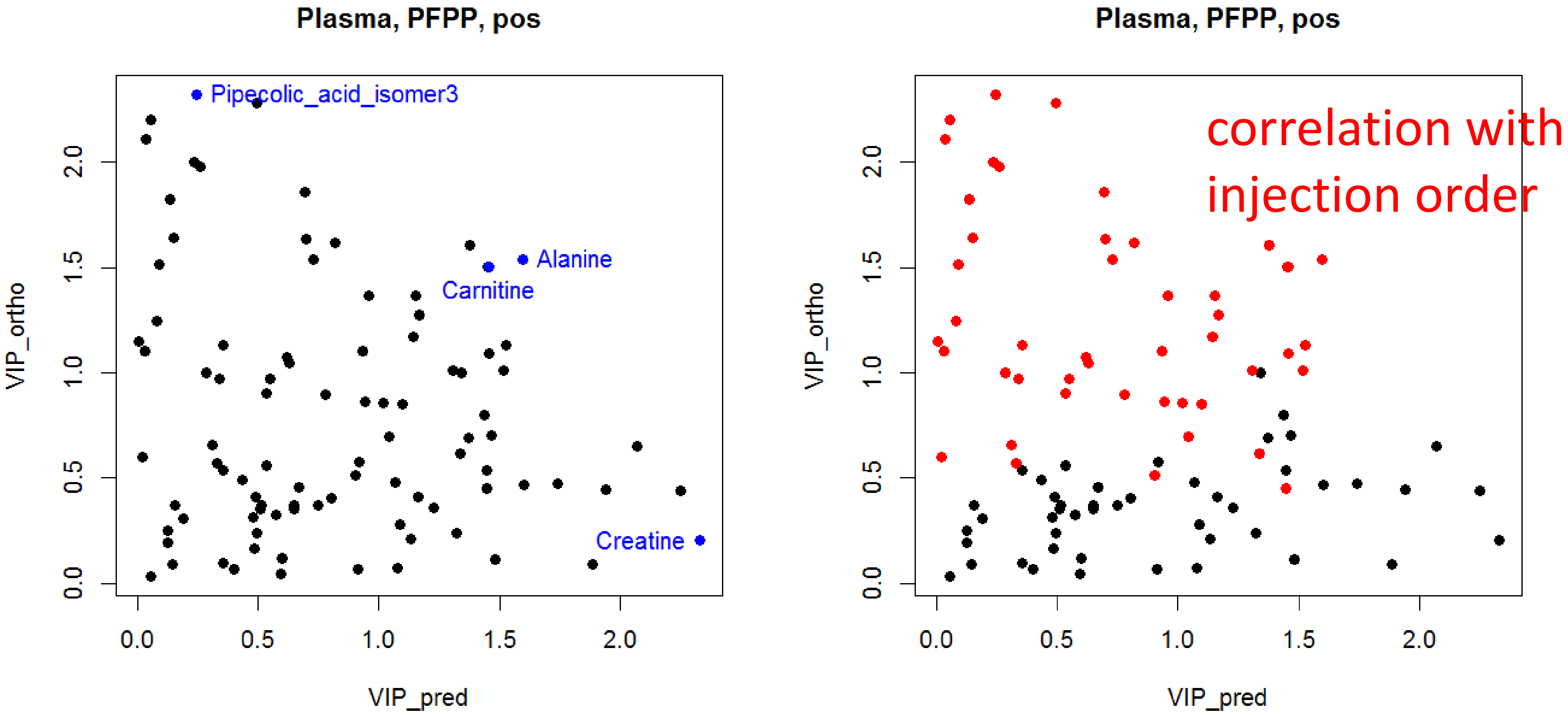
Predictive and orthogonal variations



Scores (OPLS-DA)

Scores (OPLS-DA)

Predictive and orthogonal VIP

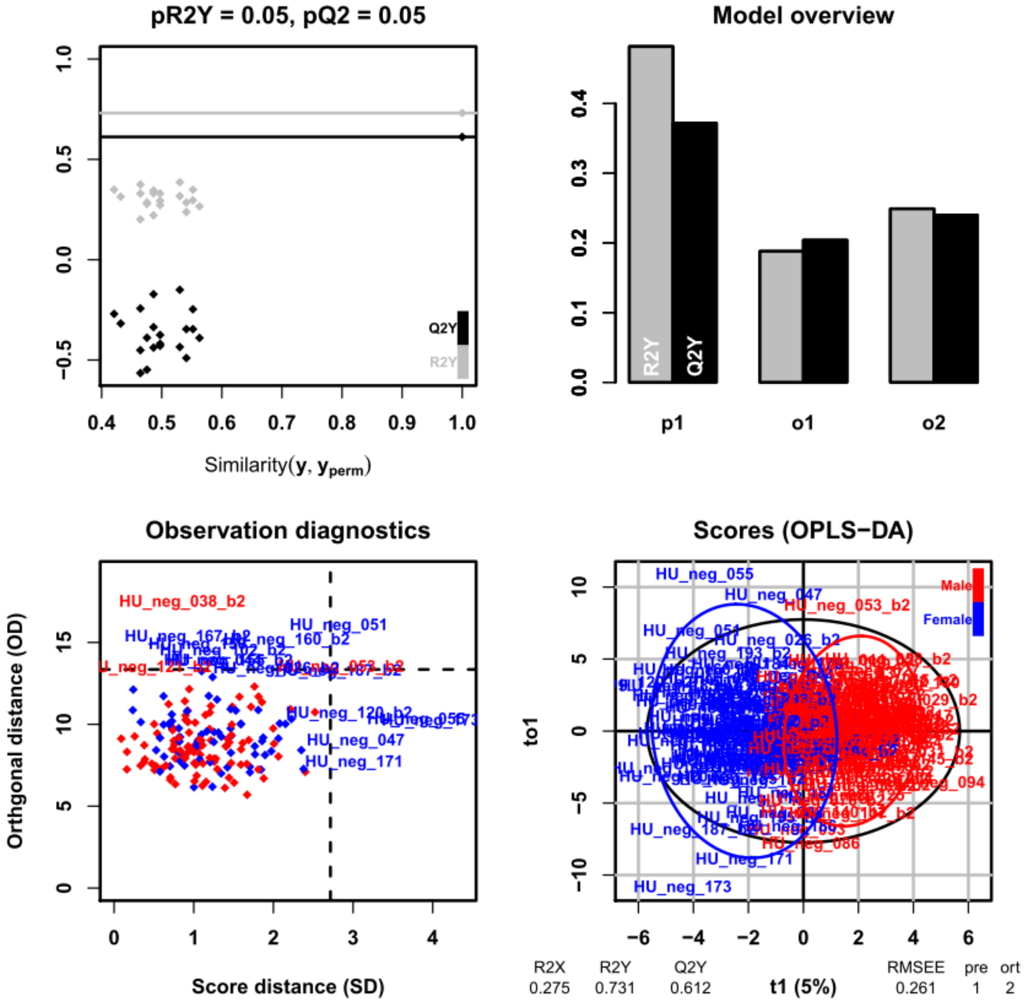


correlation with injection order

Galindo-Prieto et al (2014). *Journal of Chemometrics*, 28, 623-632.

Graphical results

- permutation, overview, outlier, and score plots displayed as the default ('summary')



pR2Y = 0.05, pQ2 = 0.05

Model overview

Observation diagnostics

Scores (OPLS−DA)

R code (*ropls* package)

- **Loading**

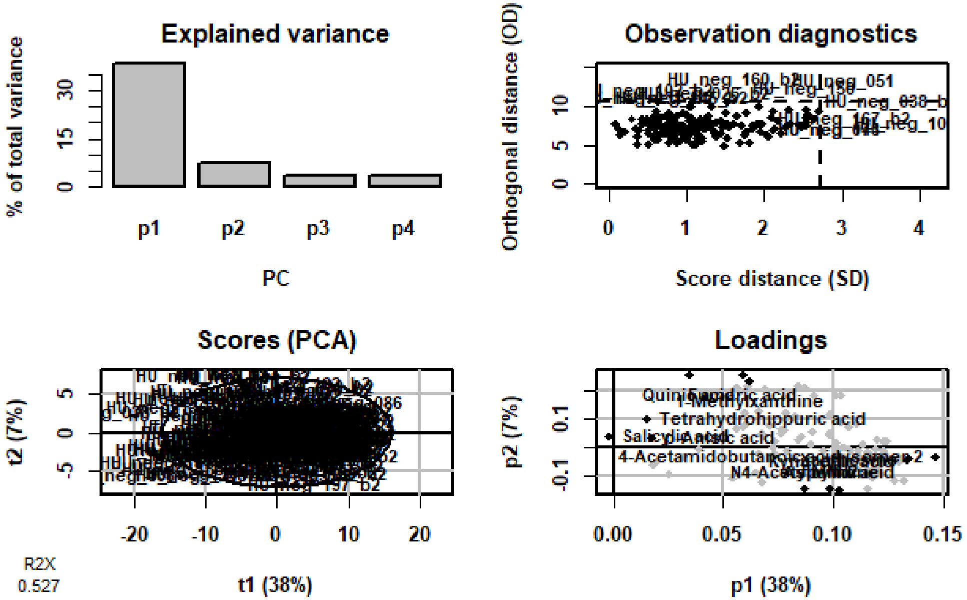
```
sacurine_dir.c <-
"C:/Users/et207099/Documents/sources/training/inst/extdata/sacurine_annotated_postprocessed"
sacurine.eset <- phenomis::reading(sacurine_dir.c)
```

- **Inspecting**

```
sacurine.eset <- phenomis::inspecting(sacurine.eset)
```

- **Computing the PCA**

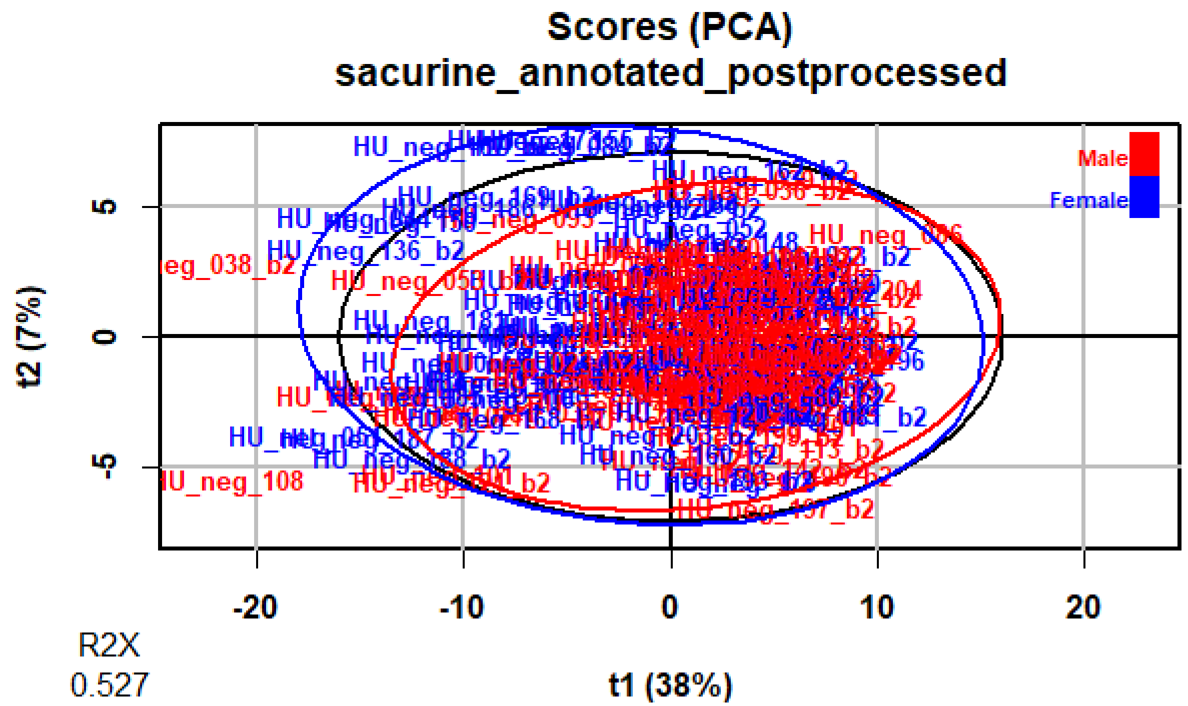
```
sacurine.pca <- ropls::opls(sacurine.eset)
```



Coloring the score plot according to the metadata

- **Coloring the score plot according to ‘gender’ (column of the sampleMetadata)**

```
ropls::plot(sacurine.pca,
            typeVc = "x-score",
            parAsColFcVn = Biobase::pData(sacurine.eset)[, "gender"]))
```



Getting back the ExpressionSet object

- **Getting back the ExpressionSet object**

```
sacurine.eset <- ropls::getEset(sacurine.pca)
```

- **The scores and loadings values have been added to the sampleMetadata and variableMetadata:**

```
head(Biobase::pData(sacurine.eset)[, c("PCA_xscor-p1", "PCA_xscor-p2")])
head(Biobase::fData(sacurine.eset)[, c("PCA_xload-p1", "PCA_xload-p2")])
```

References

- **Wold S., Sjöström M. and Eriksson L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58:109-130. http://dx.doi.org/10.1016/S0169-7439(01)00155-1**

- **Trygg J., Holmes E. and Lundstedt T. (2007). Chemometrics in Metabonomics. *Journal of Proteome Research*, 6:469-479. http://dx.doi.org/10.1021/pr060594q**

- **Brereton R.G. and Lloyd G.R. (2014). Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics*, 28:213-225.**

Multi-omics analysis and integration

Multi-steps and Multi-datasets (platforms, tissues, omics)

Importing

Post-processing

- Quality control
- Signal drift and batch effect correction
- Transformation

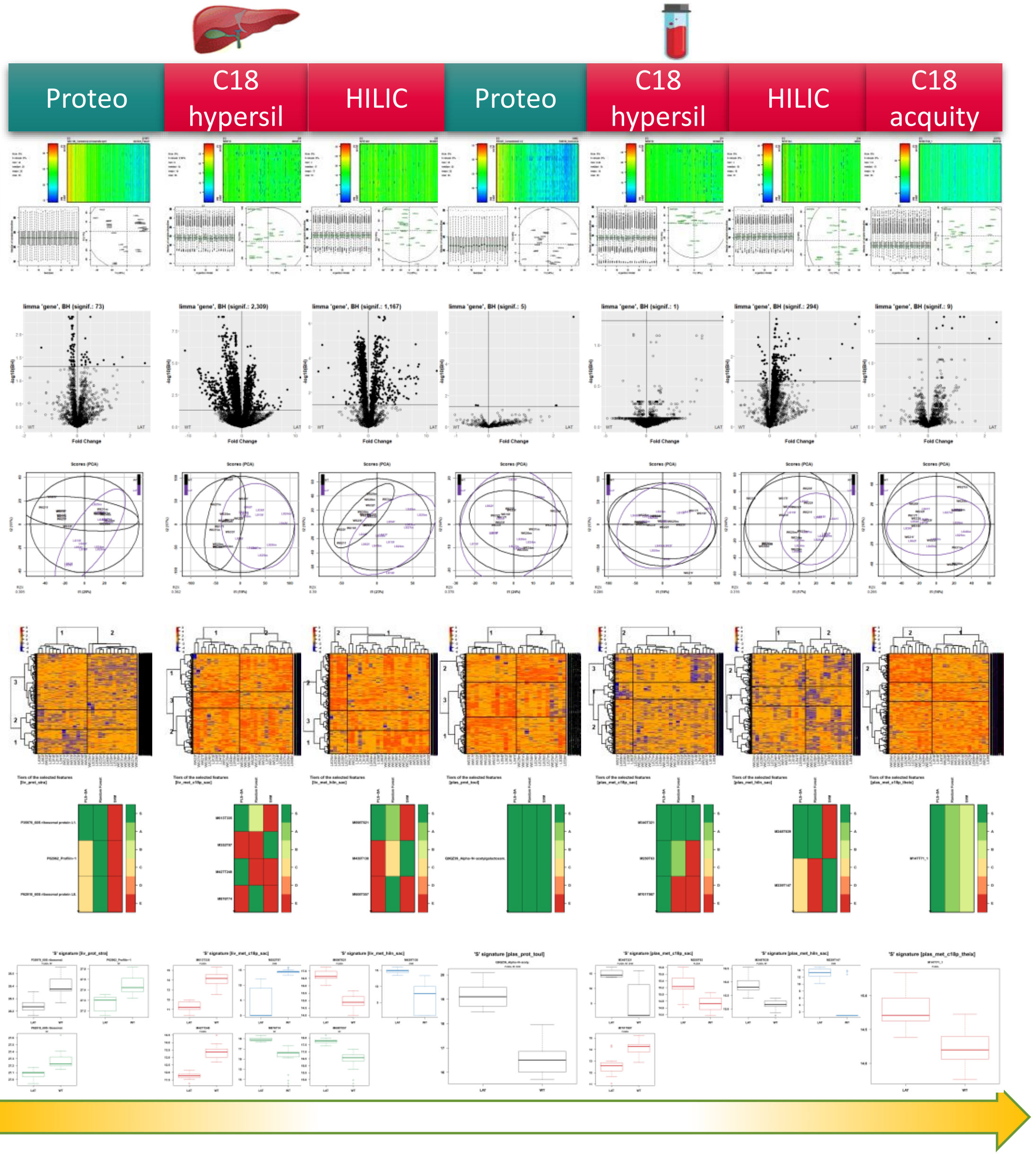
Statistics

- Univariate hypothesis testing
- PCA
- Heatmap
- (O)PLS(-DA)
- Feature selection

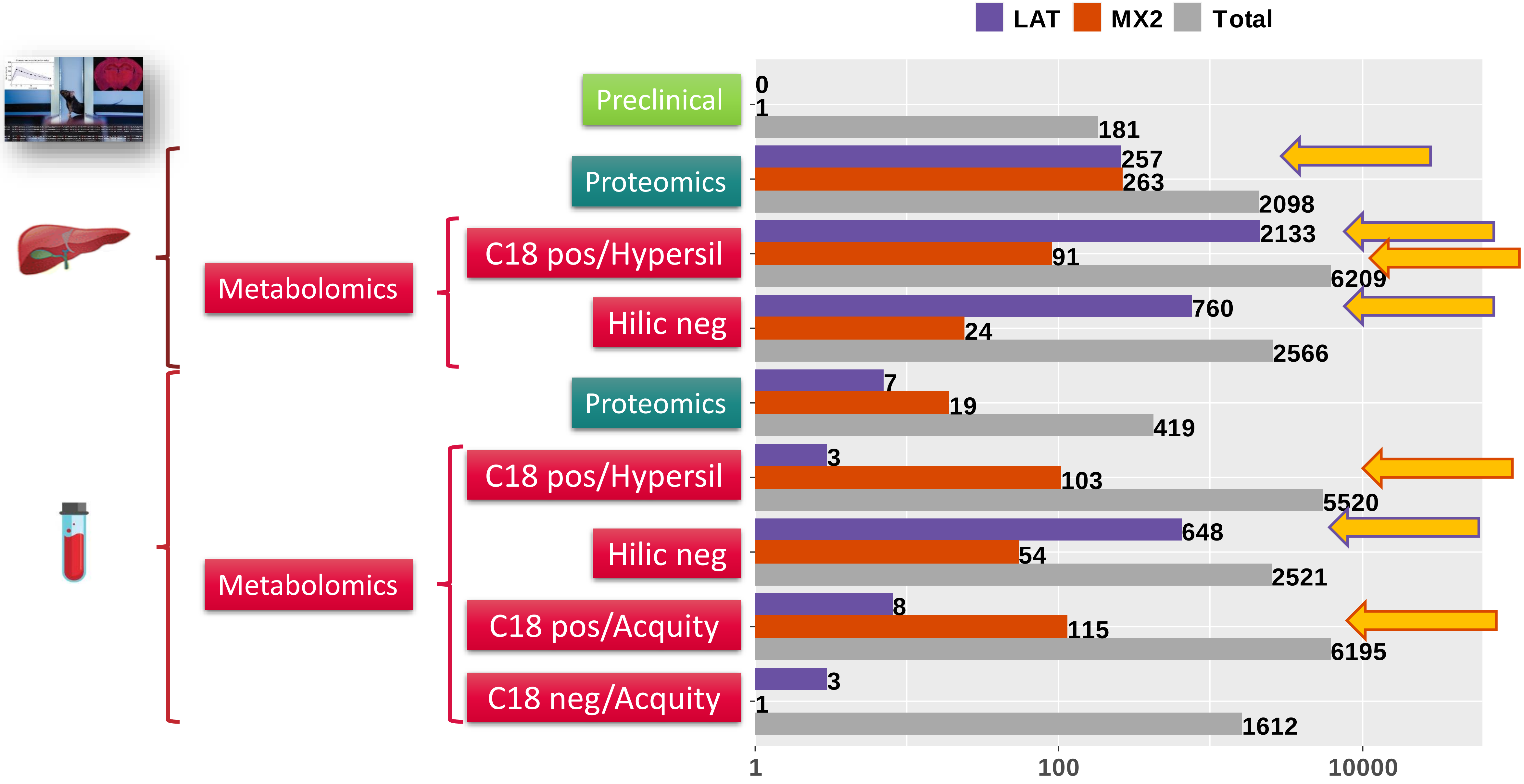
Annotation

- Chemical annotation

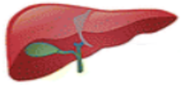
Exporting



Significant features KO vs WT (limma)

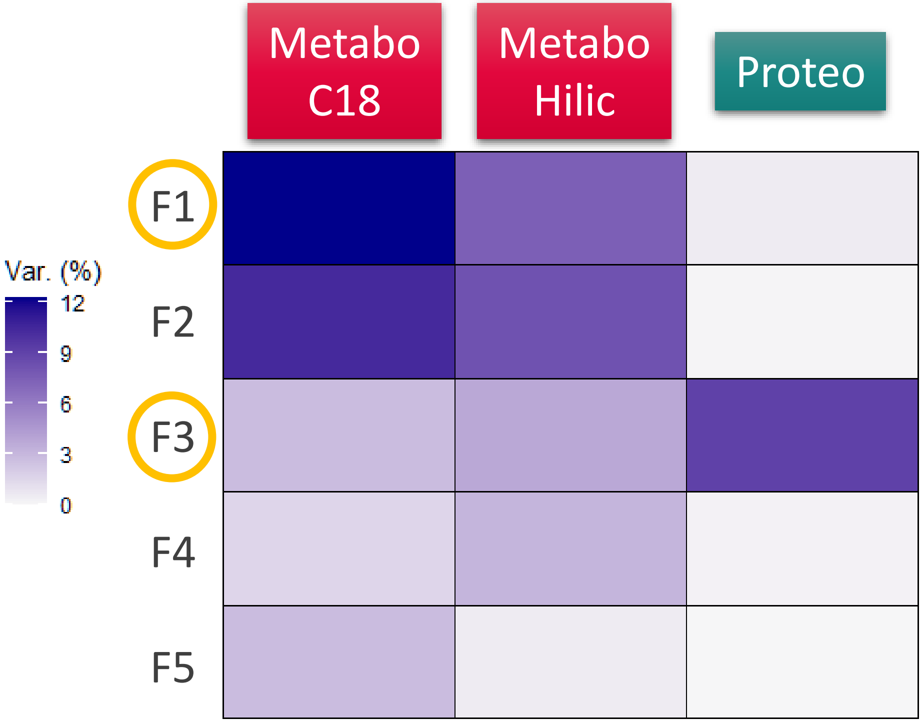


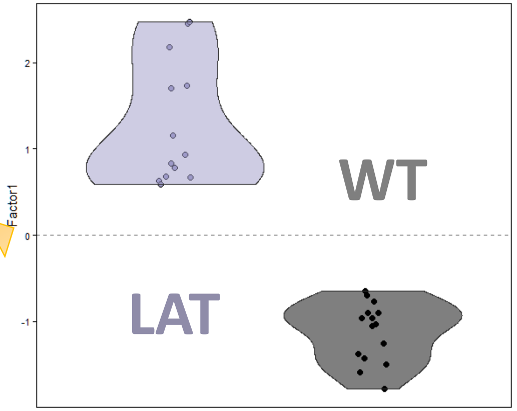
Multi-Omics Factor Analysis

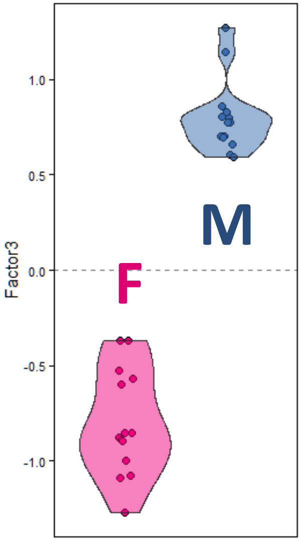


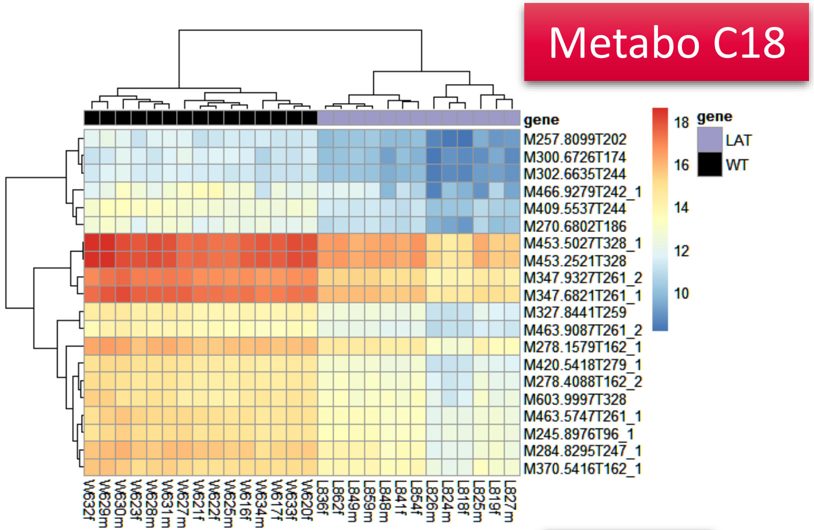
LAT vs WT

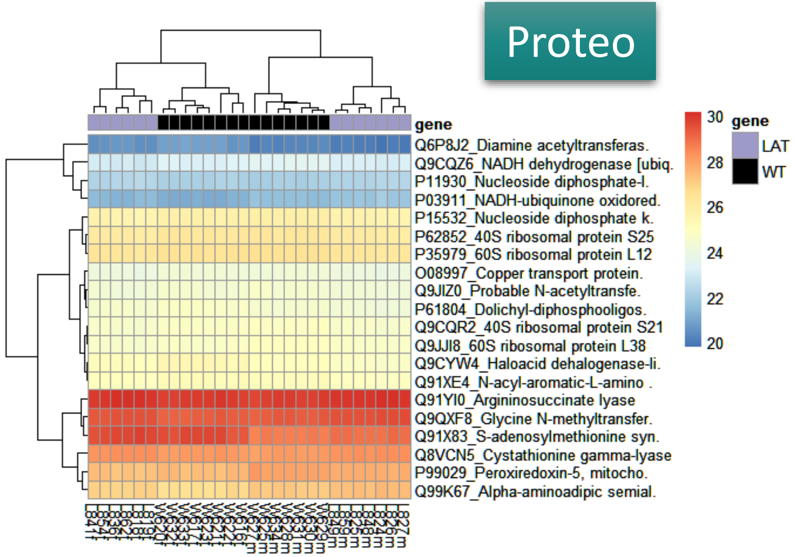
Argelaguet *et al.* (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, **14**.



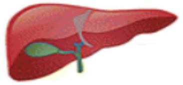




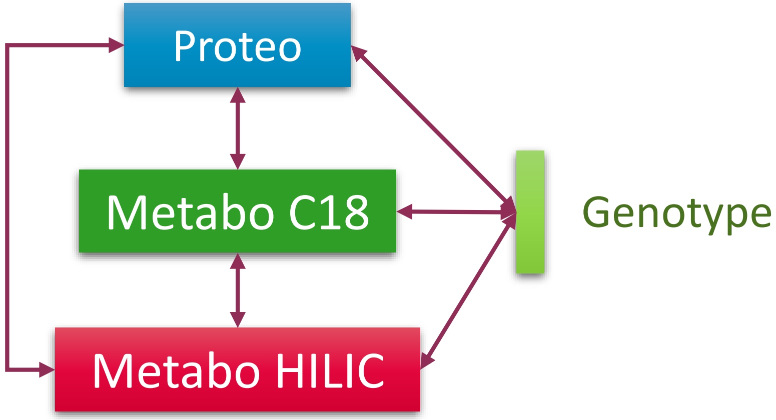




Sparse Generalized Canonical Correlation Analysis - *ProMetIS* Discriminant Analysis

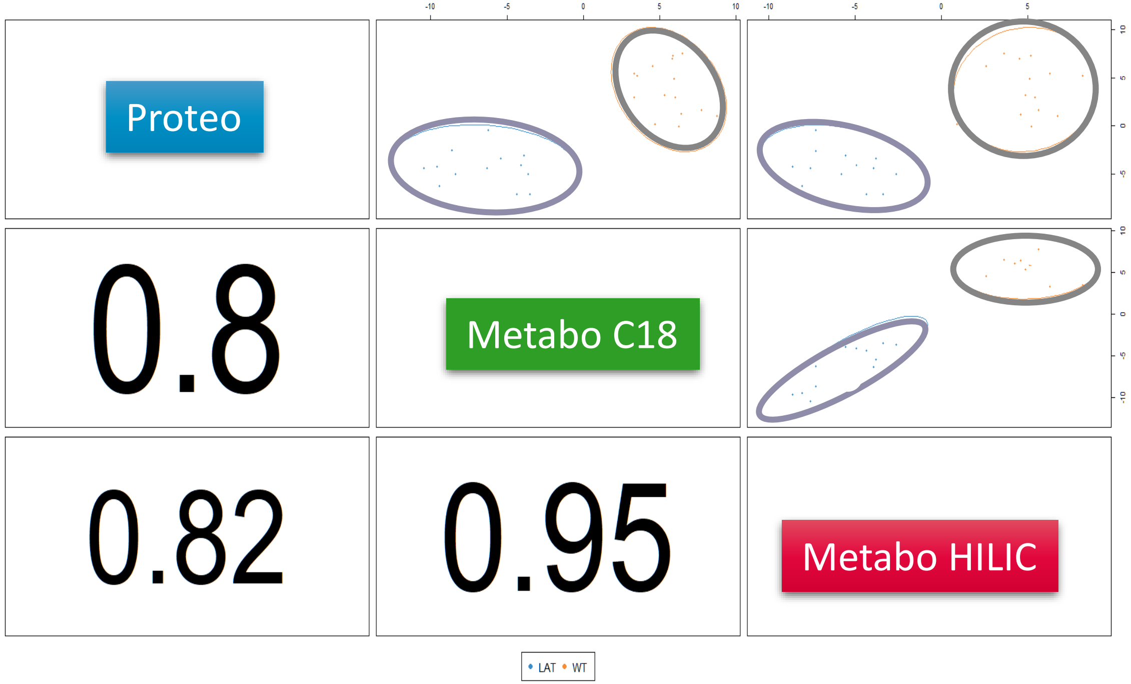


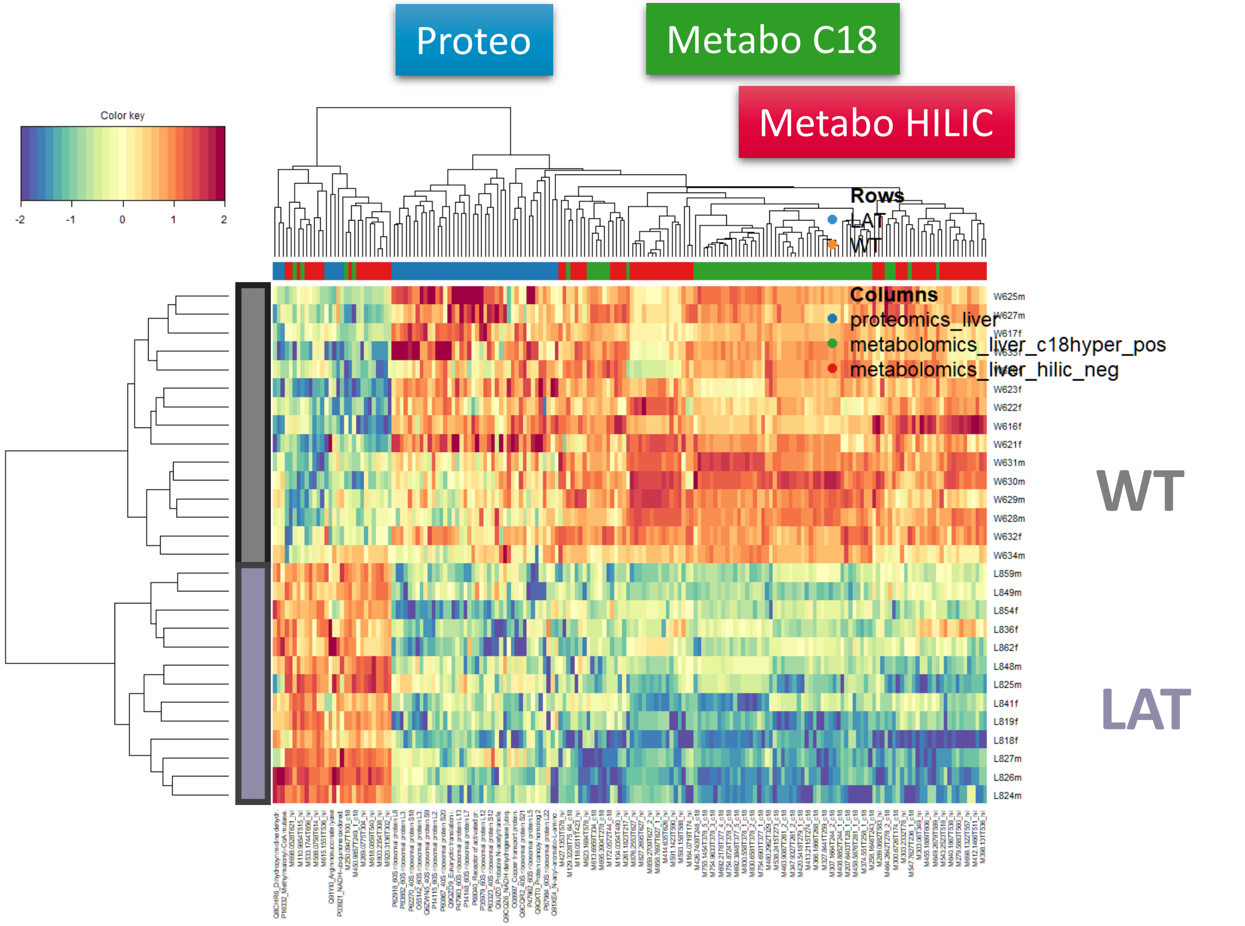
LAT vs WT



Tenenhaus *et al.* (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics*, **15**:569–583.

Singh *et al.* (2019). DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, **35**:3055–3062.





Take home message

- Value of combining proteomics and metabolomics for fundamental and applied research
- Proteomics and metabolomics data analysis is mature enough to build common pipelines
- Major challenges remain
  - Limited number of public datasets
  - Limited metabolite annotation
  - Multidisciplinarity

Alyssa Imbert
Camille Roquencourt
Camilo Broc
Krystyna Biletska
Pierrick Roger
Eric Venot
Etienne Thévenot

https://scidophenia.github.io/

SciδophenIA

METABOHUB

Florence Castelli
Emeline Chu-Van
Sadia Ouzia
Christophe Junot
François Fenaille

Marion Brandolini
Charlotte Joly
Estelle Pujos-Guillot

phenomin EXCELLENCE IN MOUSE PHENOGENOMICS

Mohammed Selloum
Tania Sorg
Sophie Leblanc
Yann Herault

ProFI PROTEOMICS

Magali Rompais
Christine Carapito

Emmanuelle Mouton
Anne Gonzalez de Peredo

Thomas Burger
Yves Vandenbrouck
Myriam Ferro

ifb

Olivier Sand
Jacques van Helden
Claudine Médigue

BIOMARGIN PhenoMeNal chopin METABOHUB ifb ProMetIS SoftwAiR

SEVENTH FRAMEWORK PROGRAMME
European Commission HORIZON 2020
INVESTISSEMENTS D'AVENIR
AGENCE NATIONALE DE LA RECHERCHE ANR