- 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

SciδophenIA

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

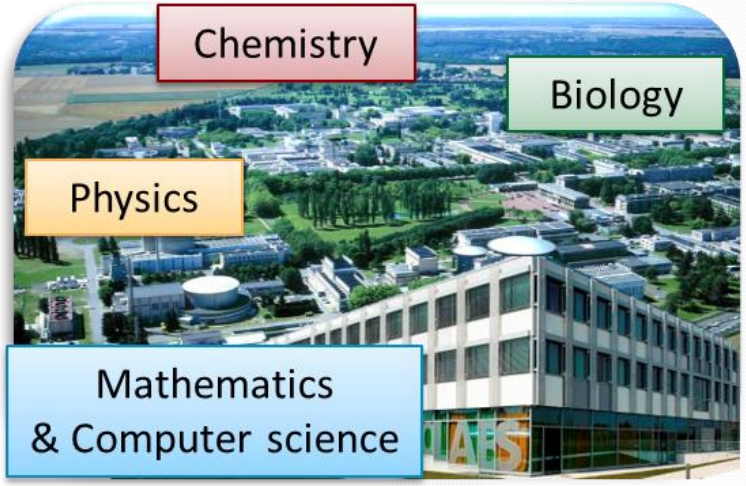
DU Bioinformatique intégrative (DUBii)

Université de Paris

ifb INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Data sciences at Paris-Saclay and CEA

- Cluster for data sciences
- Interdisciplinarity



An integrated framework in data sciences for deep phenotyping and precision medicine

biodb
Pierrick Roger

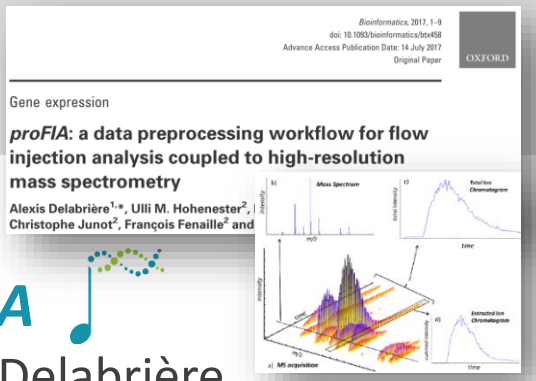
Sylvain Dechaumet

Annotation

mineMS2
Alexis Delabrière

Workflows

Signal processing



proFIA
Alexis Delabrière

ptairMS

Camille Roquencourt

Data integration

phenomis

Krystyna Biletska

ProMetIS
Alyssa Imbert

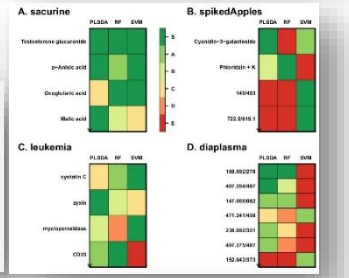
Camilo Broc

Machine learning

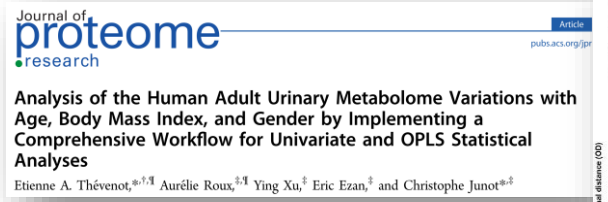
ropls
Etienne Thévenot

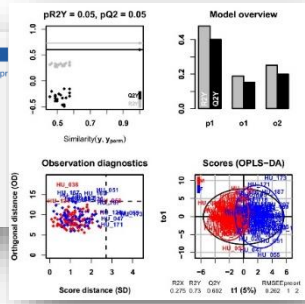
Feature selection

biosigner
Philippe Rinaudo



Eric Venot





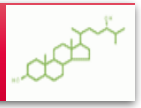
Natacha Lenuzza

Introduction

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

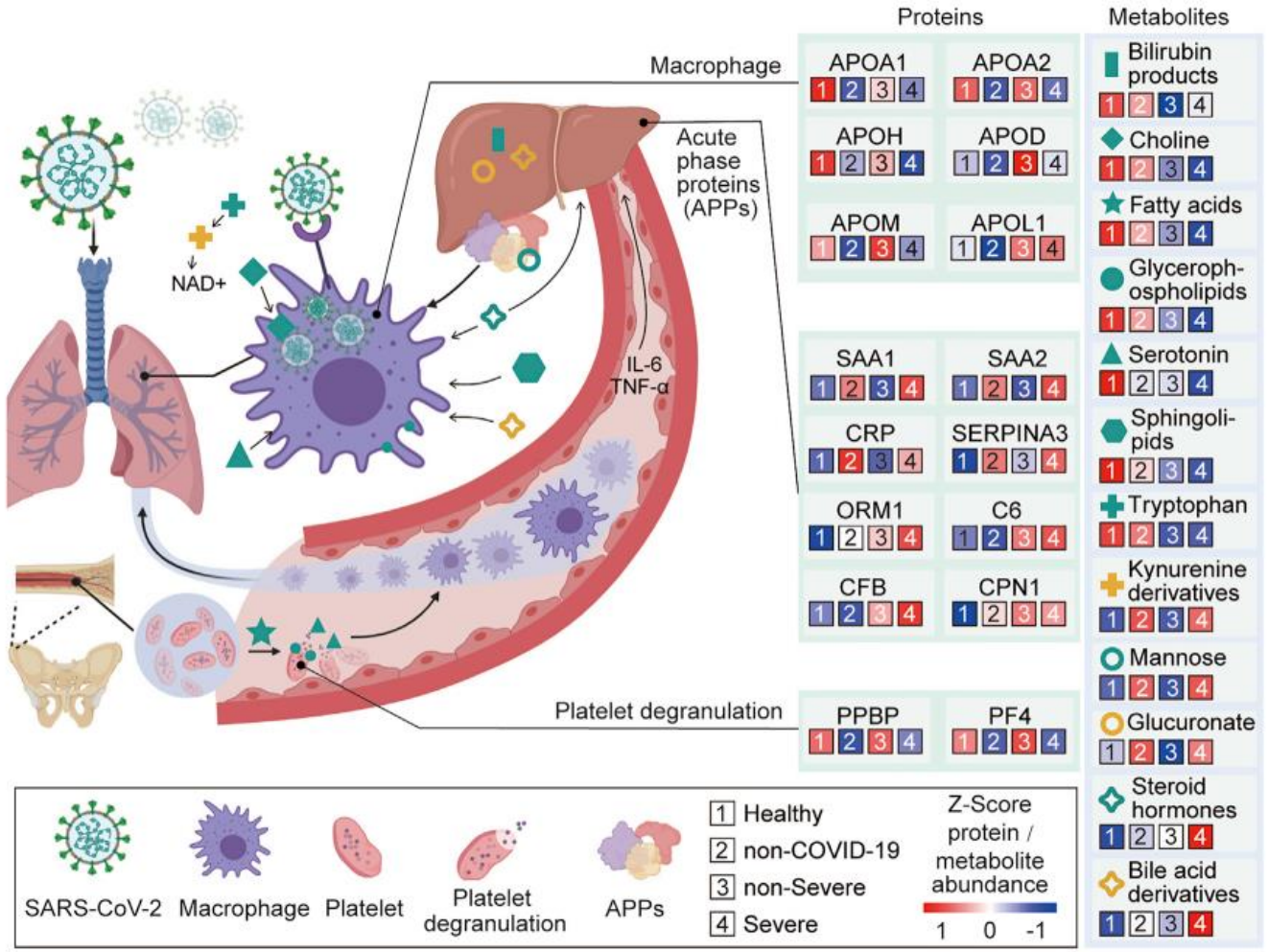
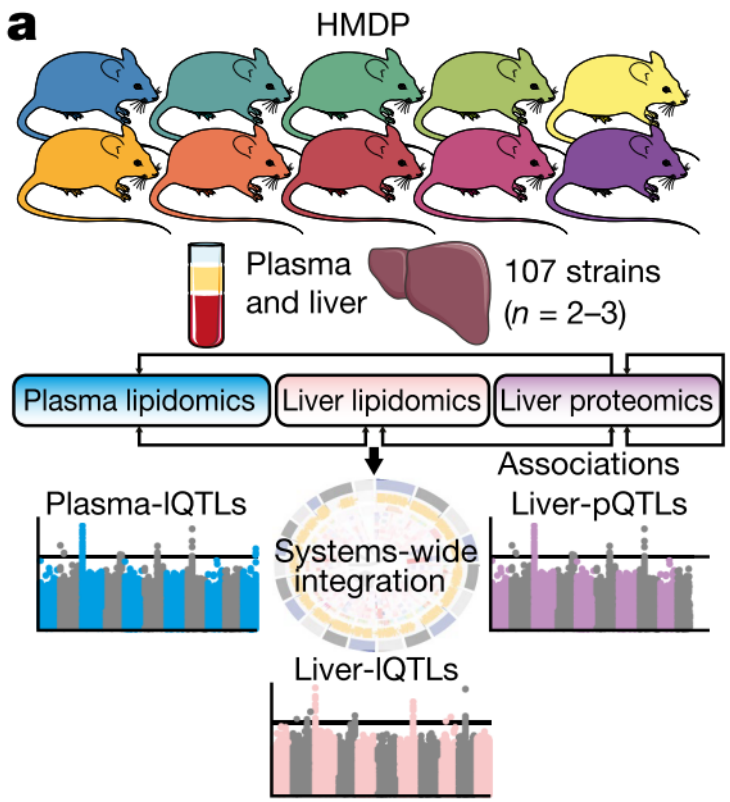


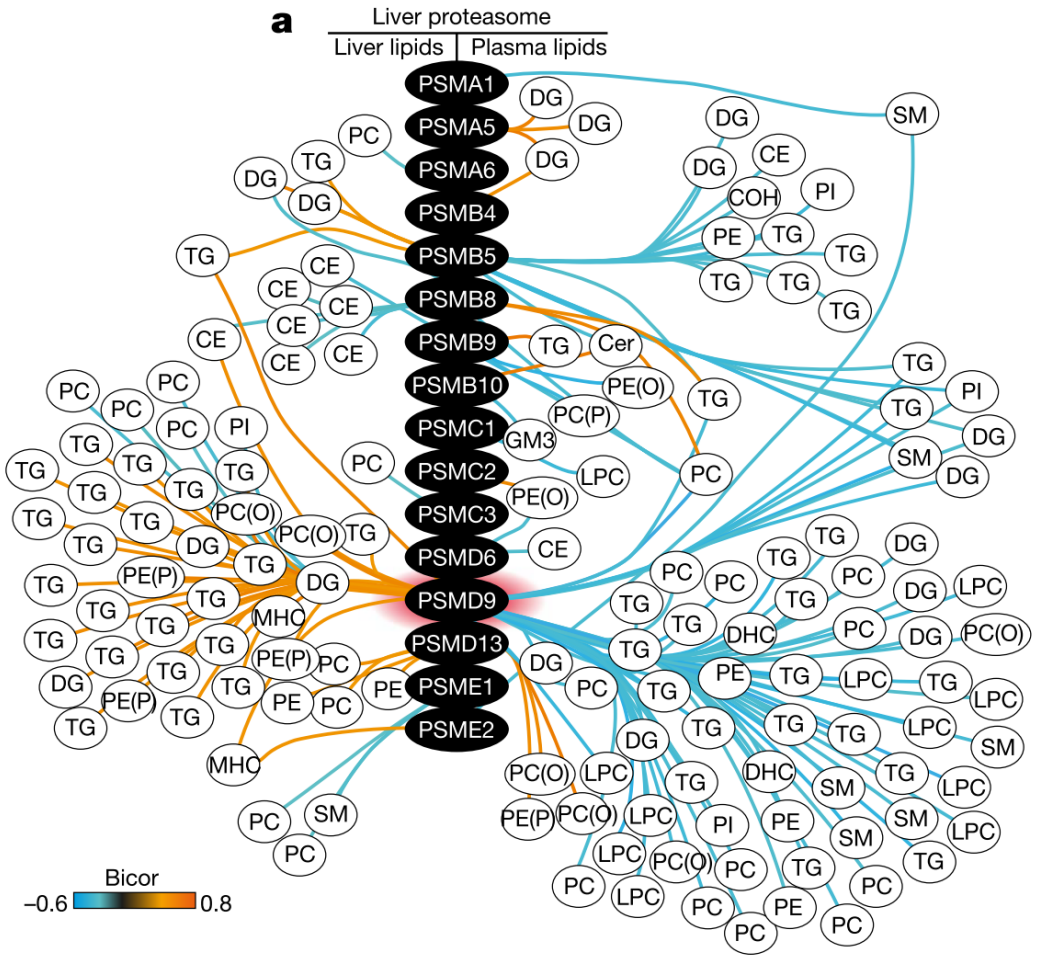
Figure 5. Key Proteins and Metabolites Characterized in Severe COVID-19 Patients in a Working Model

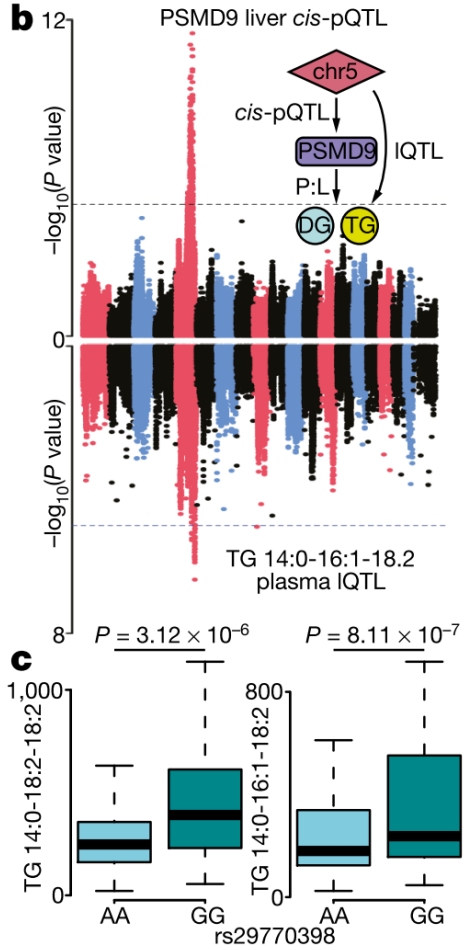
SARS-CoV-2 may target alveolar macrophages via ACE2 receptor, leading to an increase of secretion of cytokines including IL-6 and TNF-α, which subsequently induce the elevation of various APPs such as SAP, CRP, SAA1, SAA2, and C6, which are significantly upregulated in the severe group. Proteins involved 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

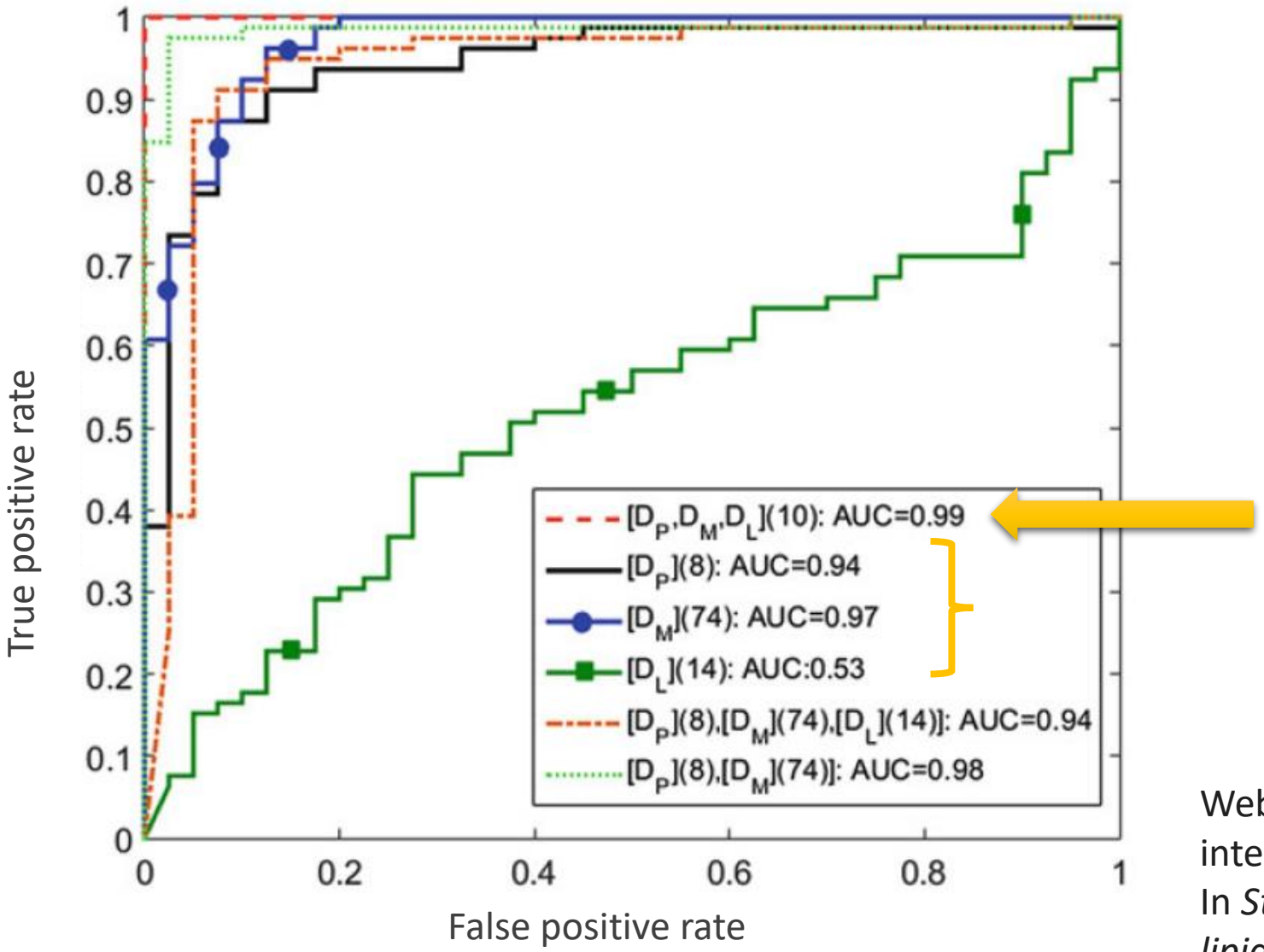






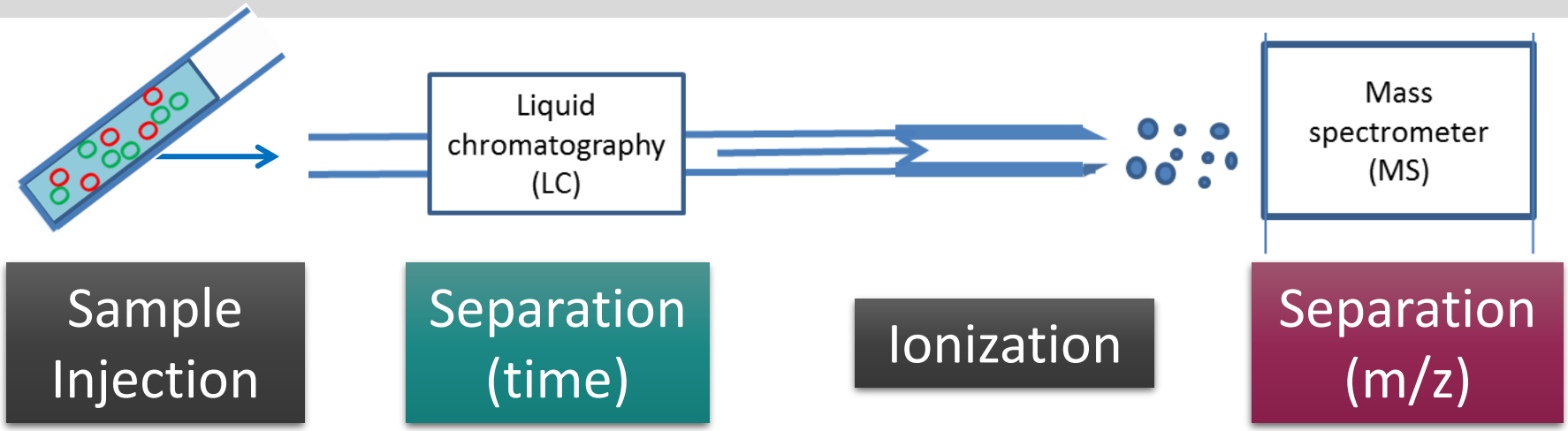
Parker *et al.* (2019). An integrative systems genetic analysis of mammalian lipid metabolism. *Nature*, **567**:187–193.

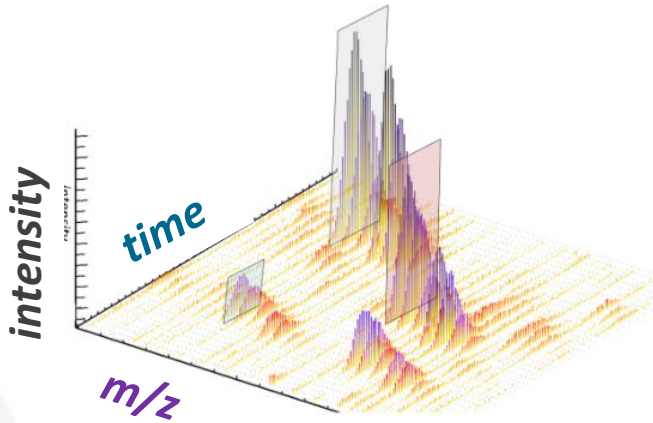
Increase the predictive performance



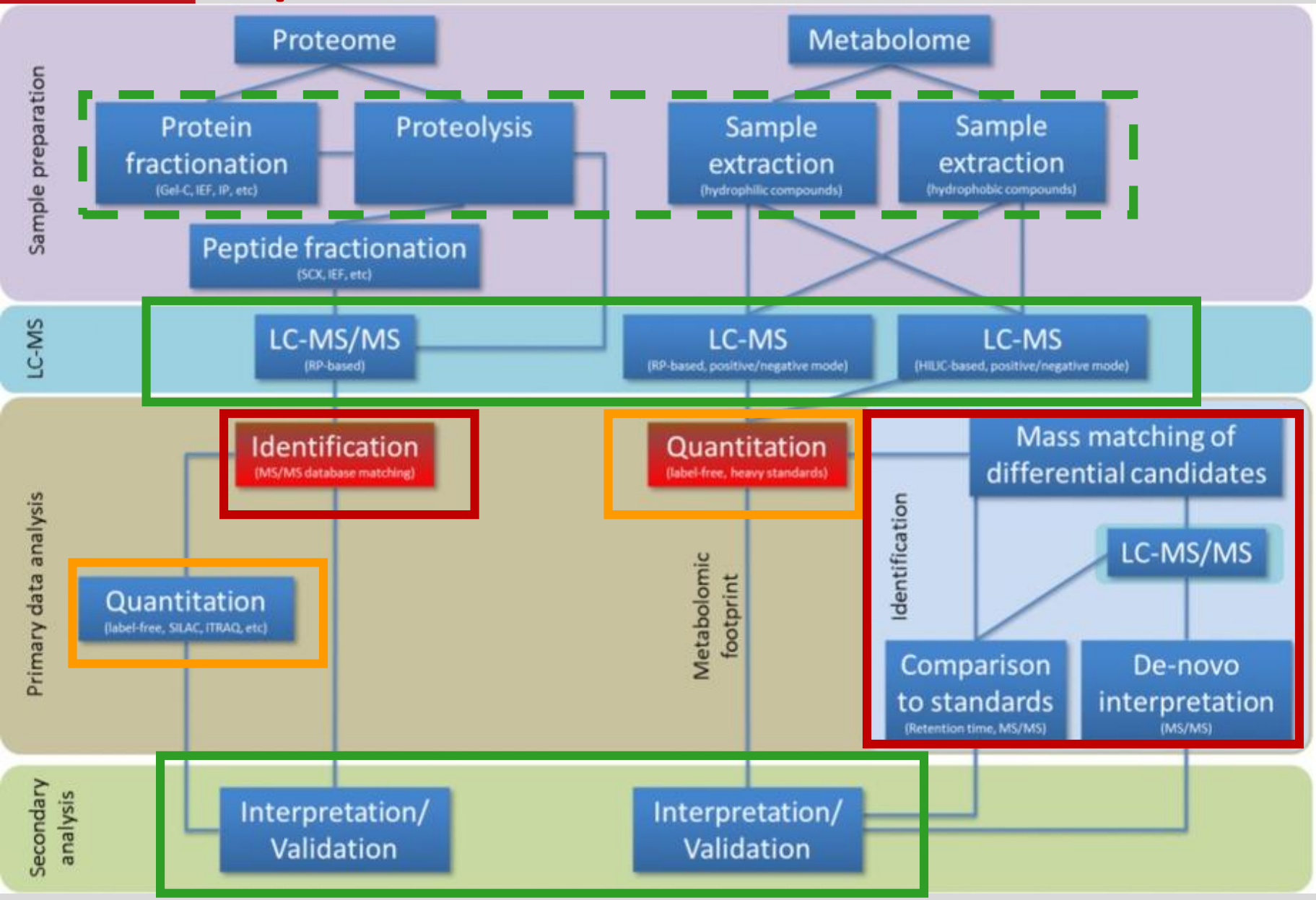
Webb-Robertson *et al.* (2016). Bayesian posterior integration for classification of mass spectrometry data. In *Statistical analysis of proteomics, metabolomics, and lipidomics data using mass spectrometry* (pp. 203–211).

Common technology: LC-HRMS





Similarities in data quantification and statistical analysis
Specificities in data annotation

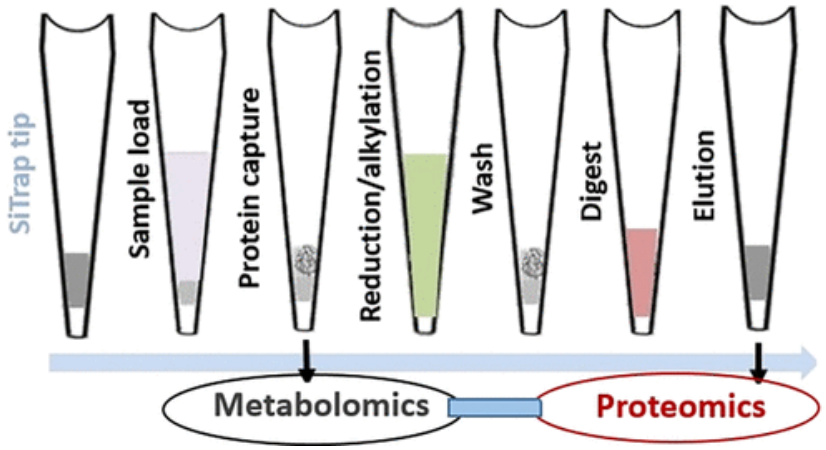


Fischer *et al.* (2013). Two birds with one stone: doing metabolomics with your proteomics kit. *Proteomics*, **13**:3371-3386.

Common sample preparation studies

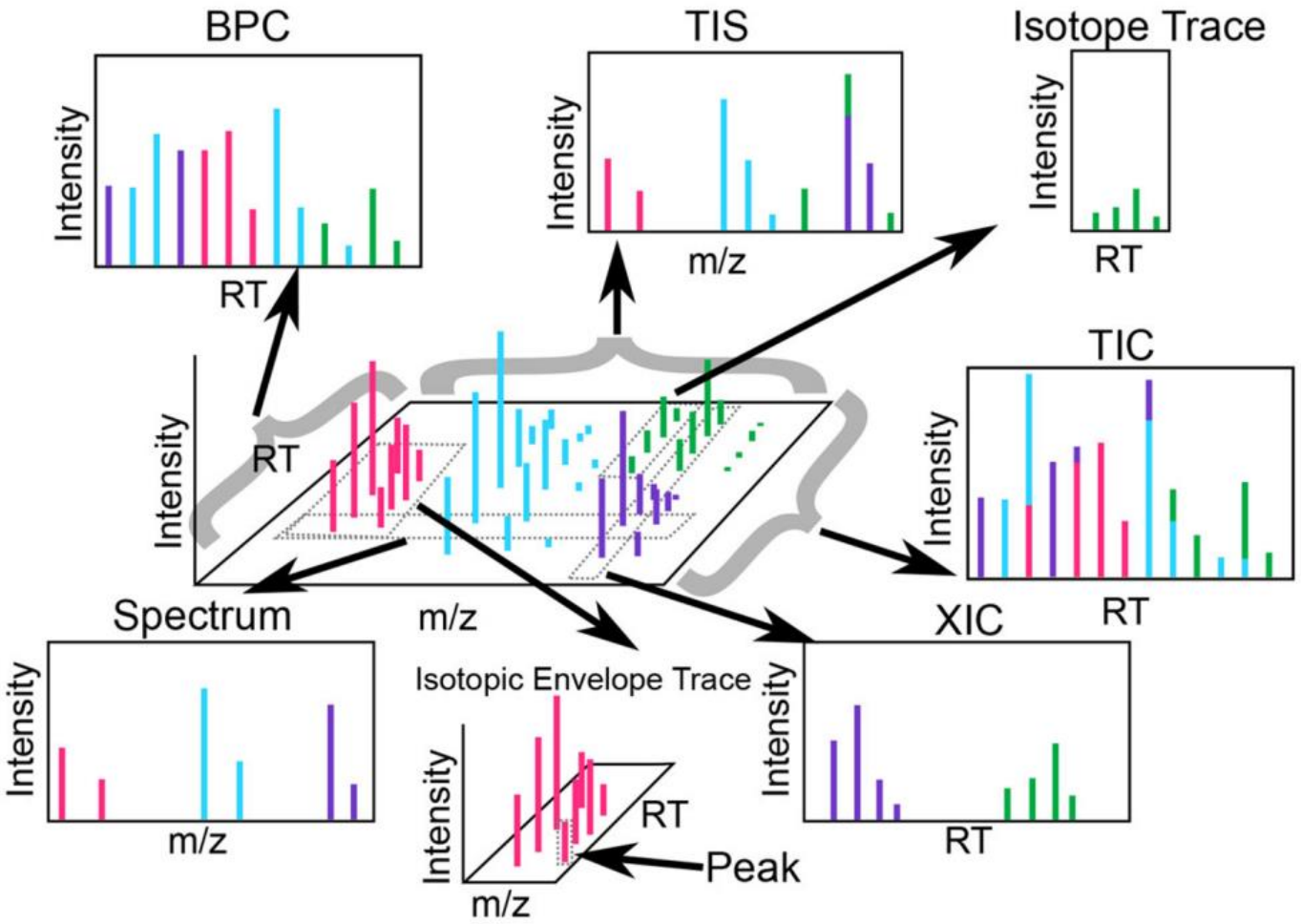
Fischer *et al.* (2013). Two birds with one stone: doing metabolomics with your proteomics kit. *PROTEOMICS*, **13**:3371-3386.

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.

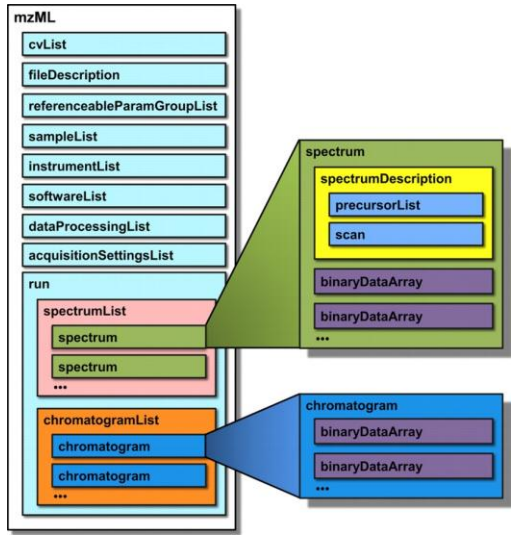
Common nomenclature



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**

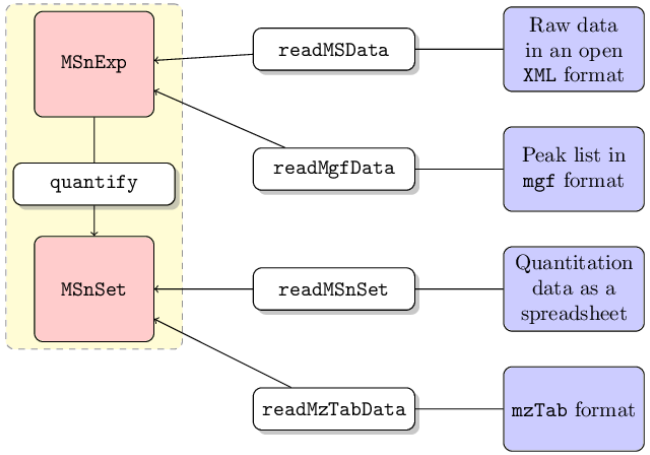


Martens *et al*. (2010). mzML - a community standard for mass spectrometry data. *Molecular & Cellular Proteomics*. **10**.

  - 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**

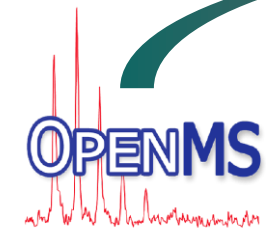


Gatto *et al*. (2020). MSnbase, efficient and elegant R-based processing and visualization of raw mass spectrometry data. *Journal of Proteome Research*.

Toward common software

Proteomics

Metabolomics



phenomis

SDA

WGCNA

ProMetIS

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.

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?

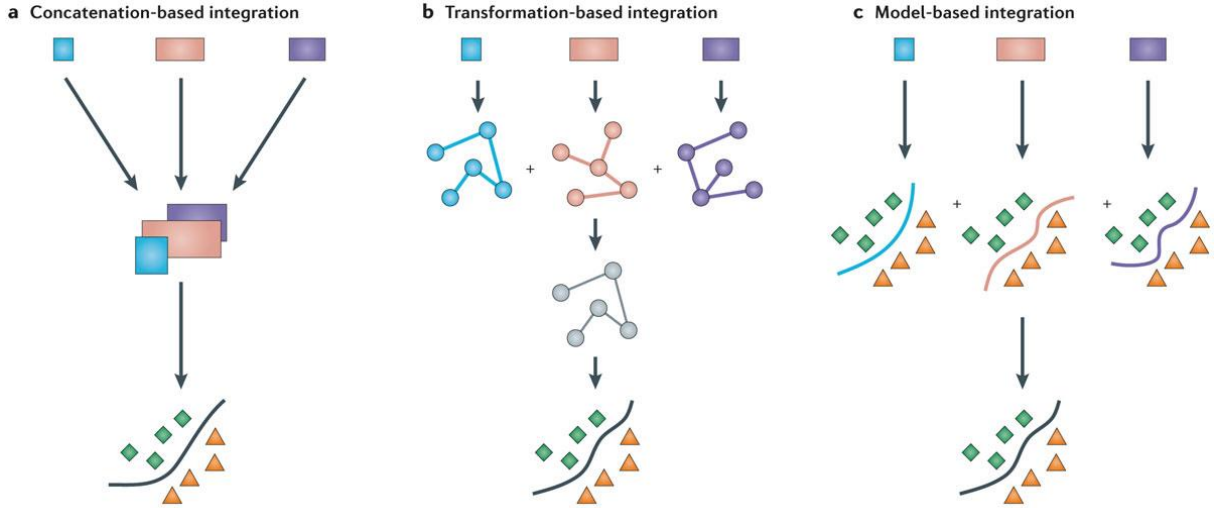
Data integration: challenges

- 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

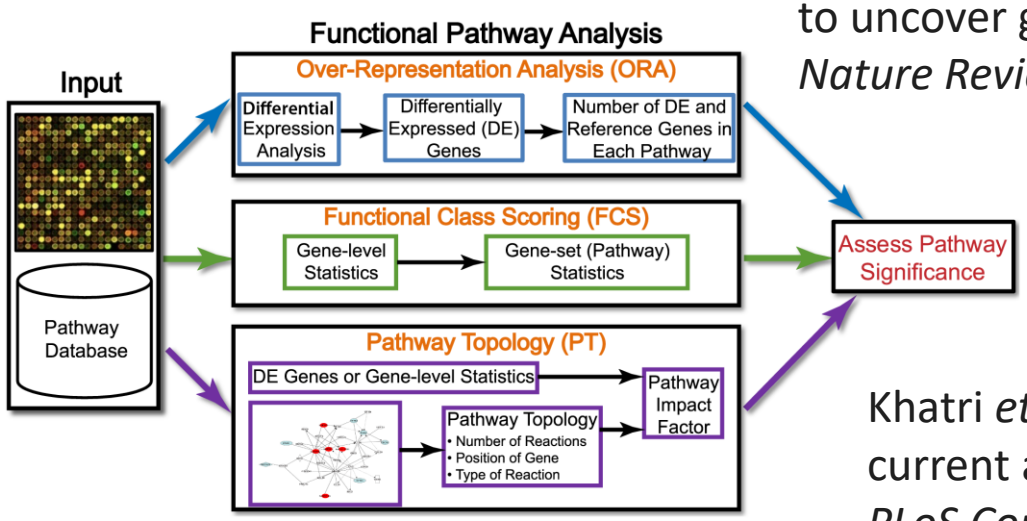
Data integration: approaches

- **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



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

- **Bioinformatics**
  - Mapping
  - Enrichment
    - Molecule set
    - Topology-based



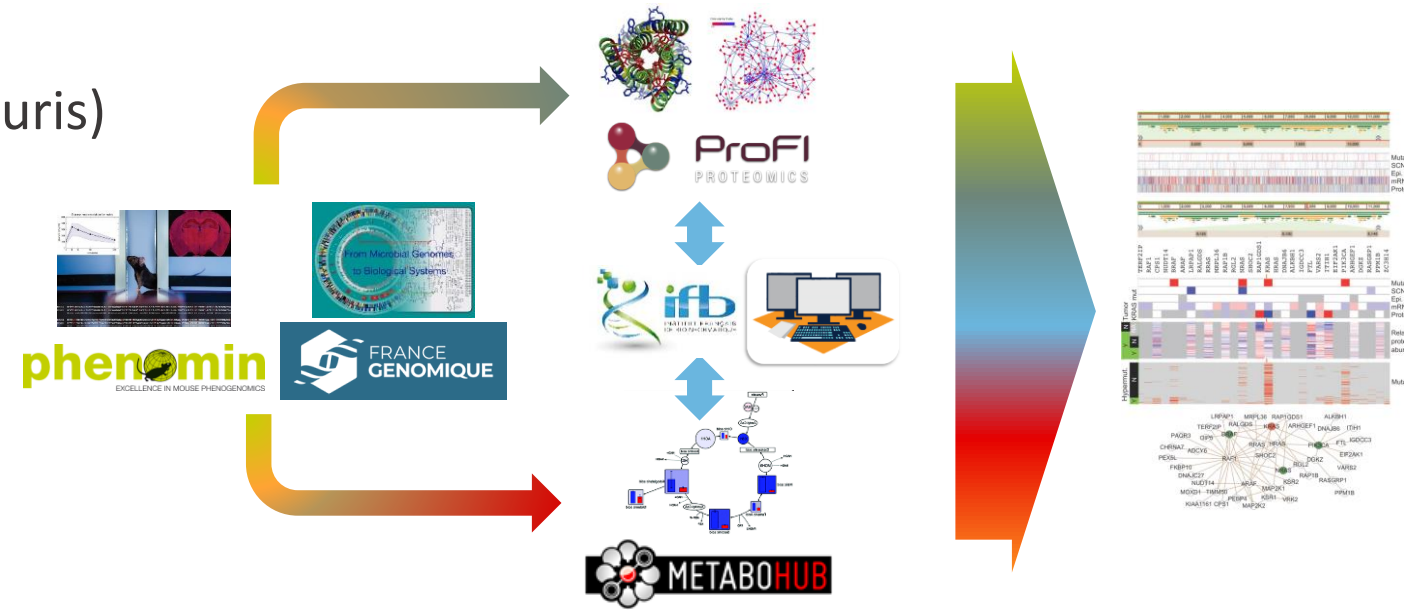
Khatri *et al.* (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology*, **8**: e1002375.

ProMetIS

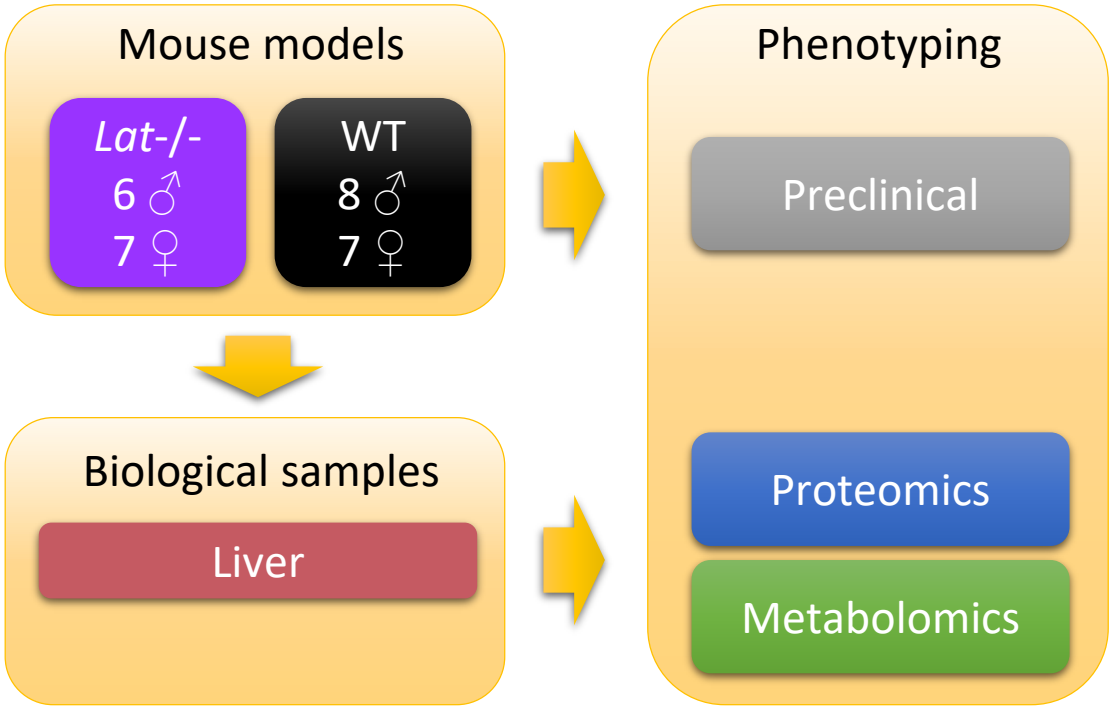
ProMetIS: deep phenotyping of mouse models by proteomics and metabolomics

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



LAT and MX2 knock-out mice



- **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.

9 preclinical, proteomic and metabolomic datasets

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

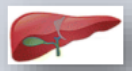
Pool CV ≤ 0.3

Pool CV / sample CV ≤ 1

Chemical redundancy

log2 transformation

preclinical 236



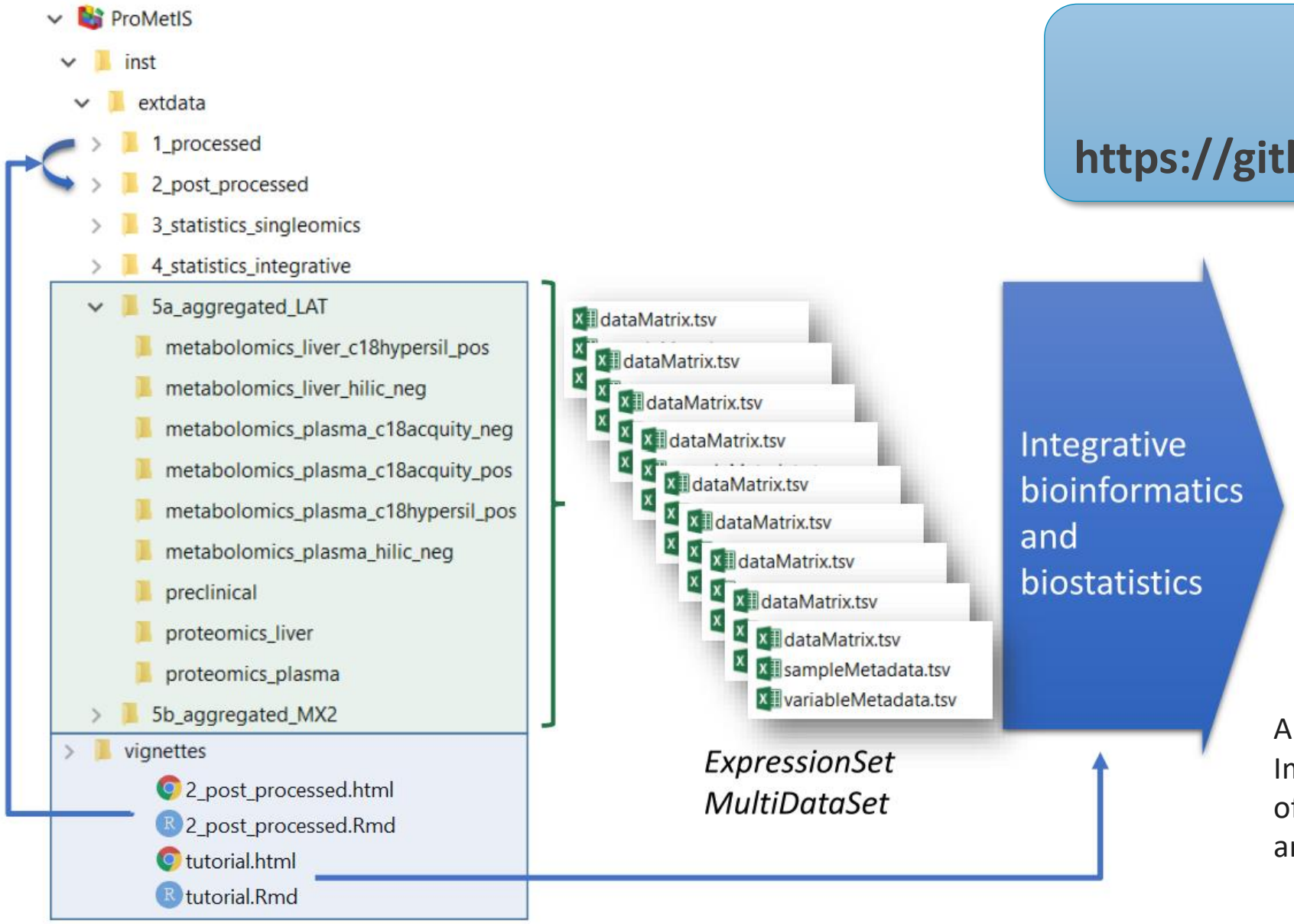
liver_proteomics 2,187

liver_metabo_c18hypersil_pos 5,665 [138]
liver_metabo_hilic_neg 2,866 [199]

plasma_proteomics 446

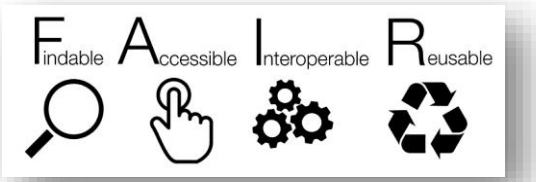
plasma_metabo_c18hypersil_pos 4,788 [113]
plasma_metabo_hilic_neg 3,131 [191]
plasma_metabo_c18acquity_pos 6,104 [78]
plasma_metabo_c18acquity_neg 1,584 [49]

Datasets & code availability: the *ProMetIS* package



ProMetIS

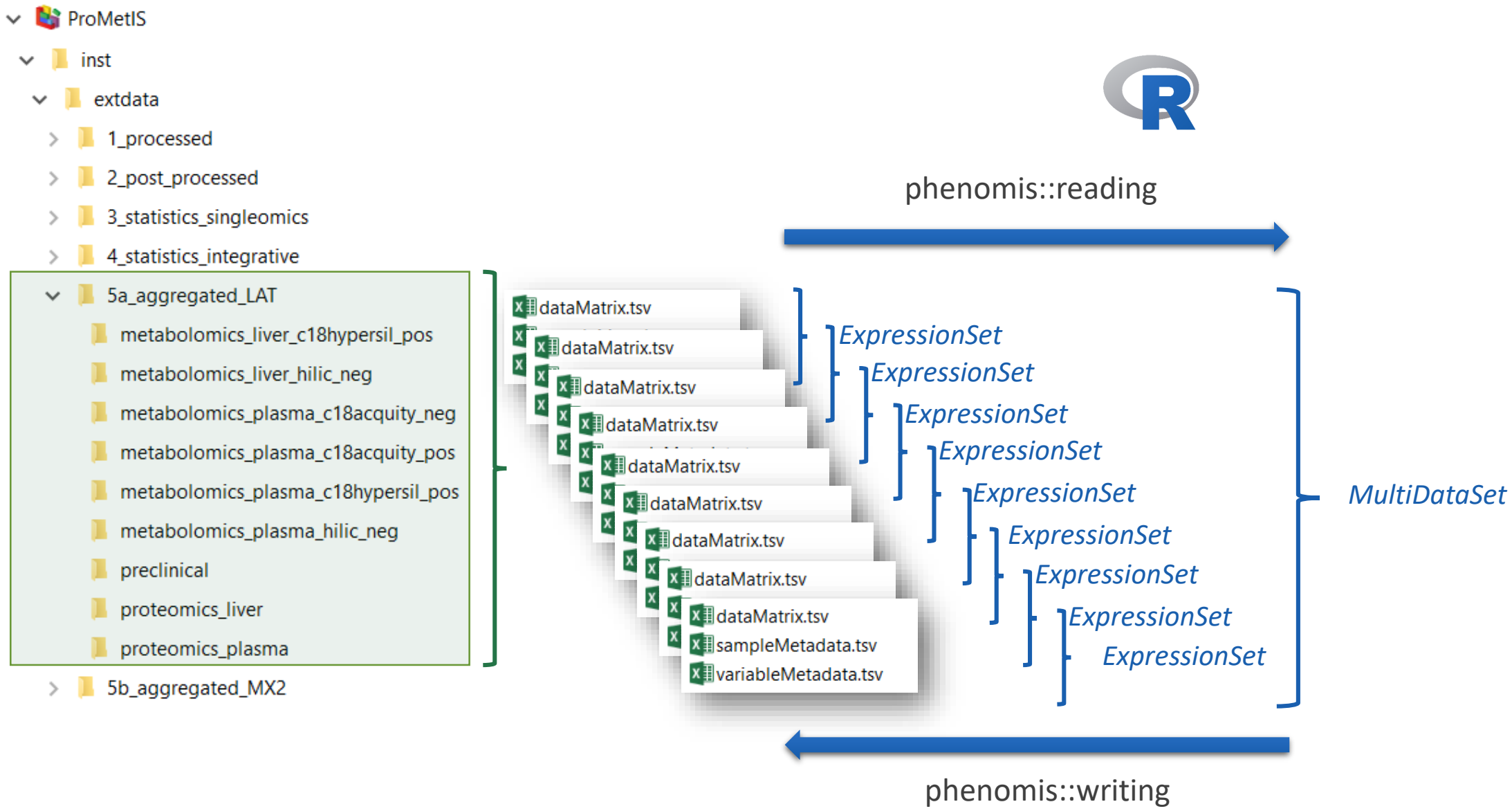
https://github.com/IFB-ElixirFr/ProMetIS



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

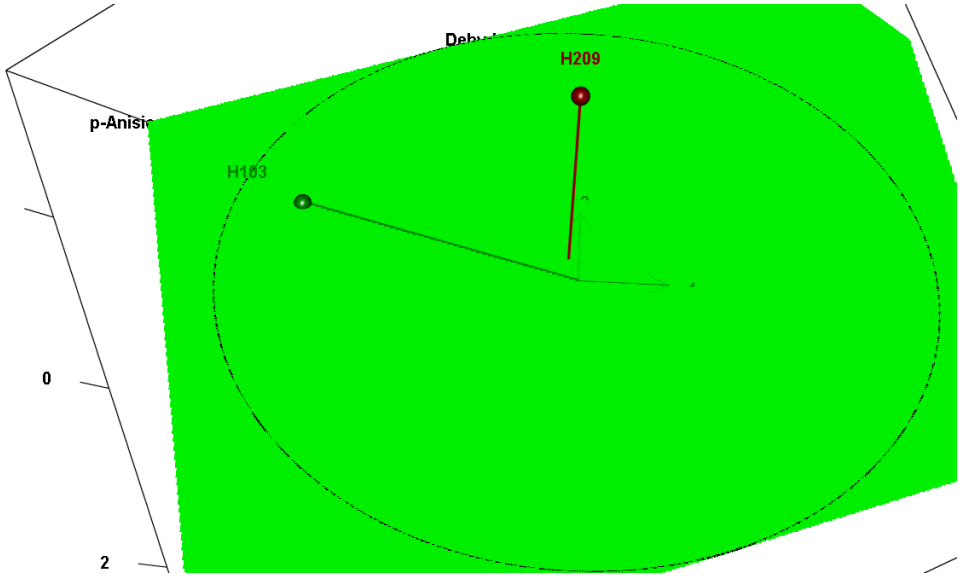
Hands-on

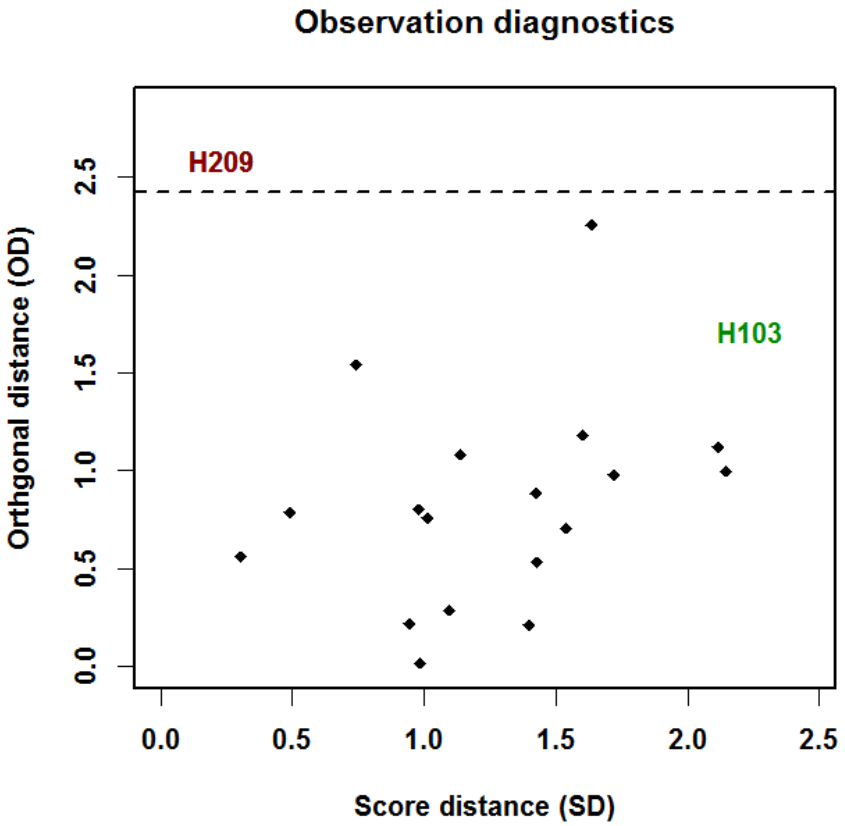
Reading the data



PCA: Observation diagnostics

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

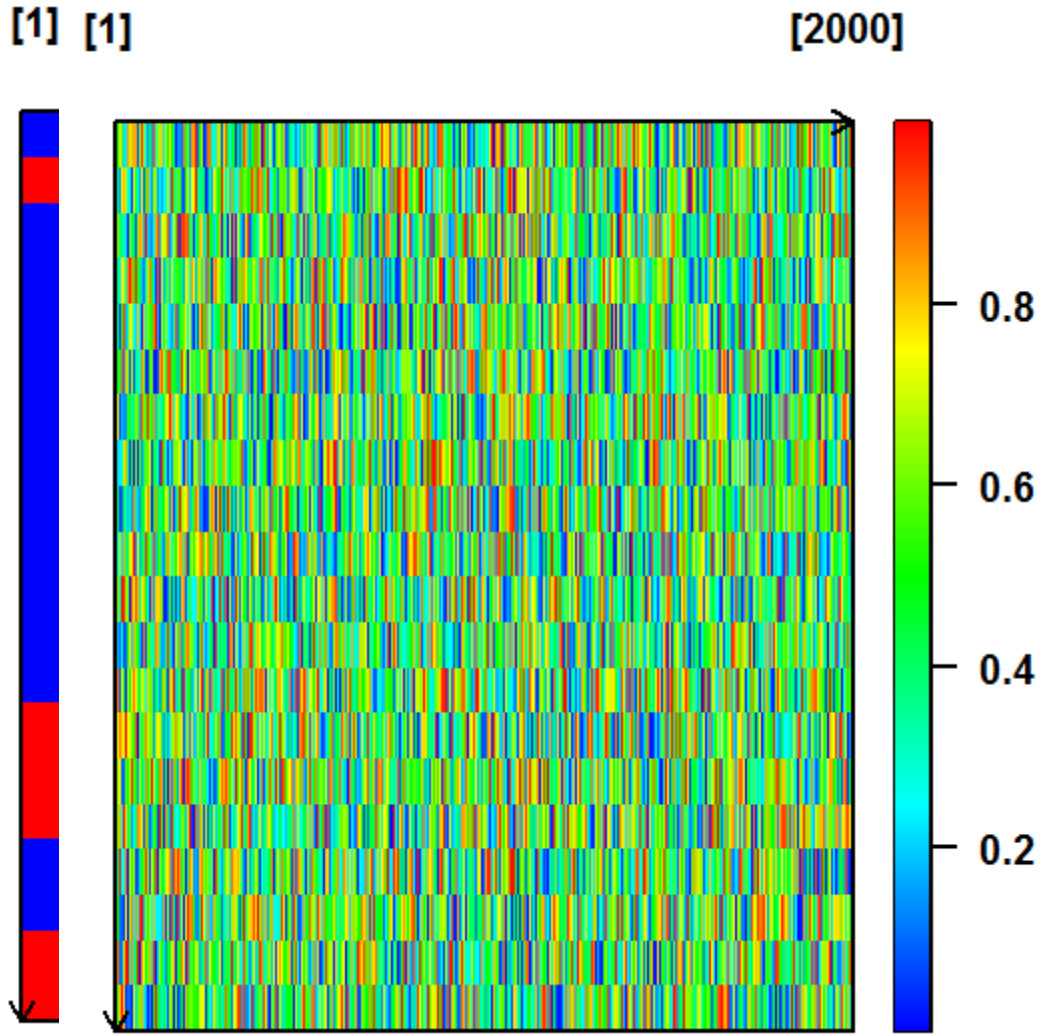




Hubert M., Rousseeuw P. and Vanden Branden K. (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics,* **47:**64-79. DOI:10.1198/0040170040000000563

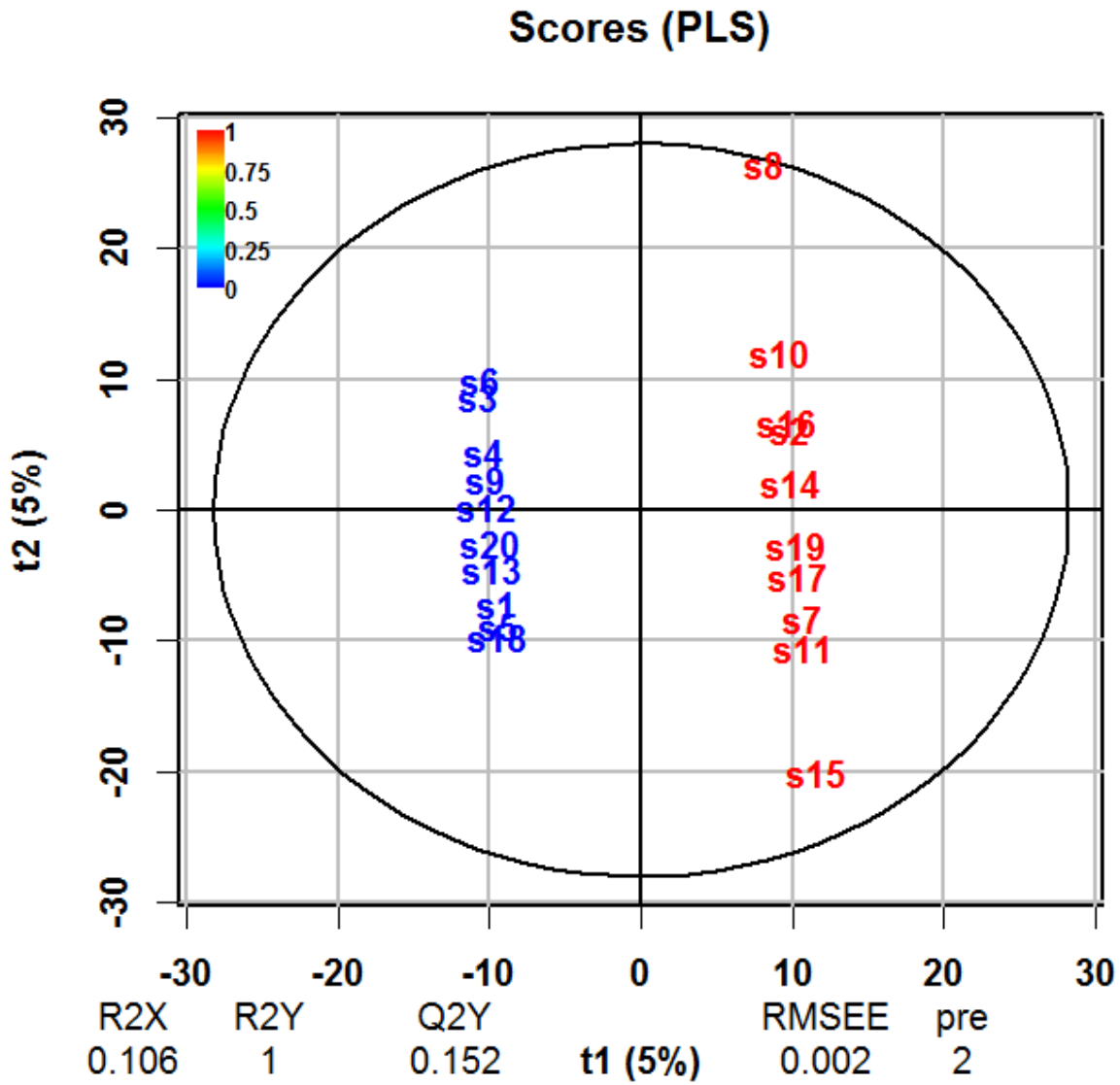
PLS: overfitting

- **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



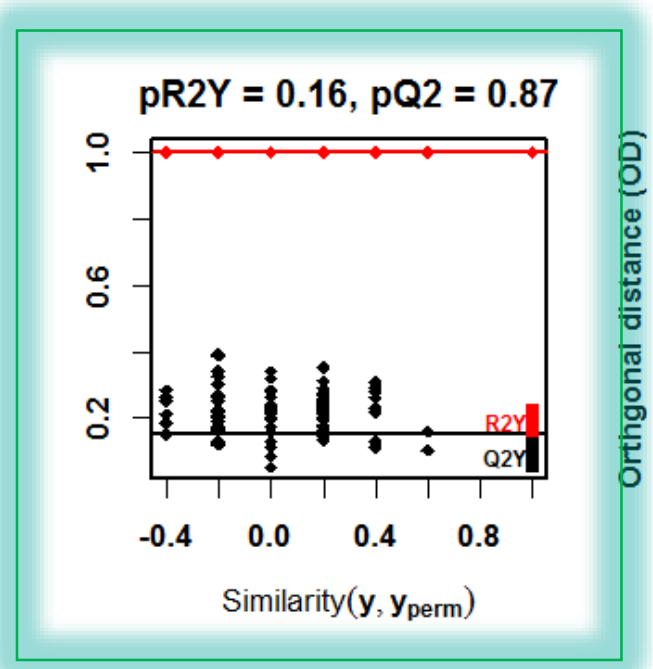
adpated from Wehrens (2011). Chemometrics with R. Springer.

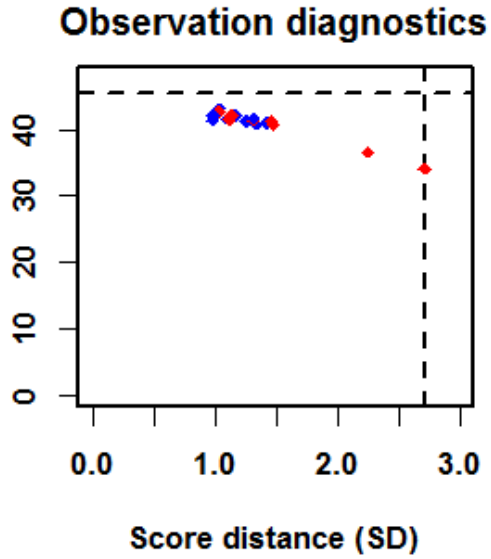
Score plot!

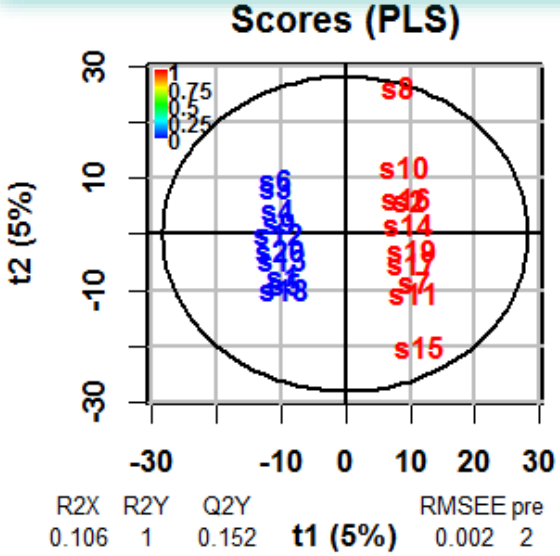

Scores (PLS)
1
0.75
0.5
0.25
0
s8
s10
s6
s3
s16
s2
s4
s9
s14
s12
s20
s19
s13
s17
s1
s7
s18
s11
s15
t2 (5%)
30
20
10
0
-10
-20
-30
-30
-20
-10
0
10
20
30
R2X
R2Y
Q2Y
RMSEE
pre
0.106
1
0.152
t1 (5%)
0.002
2

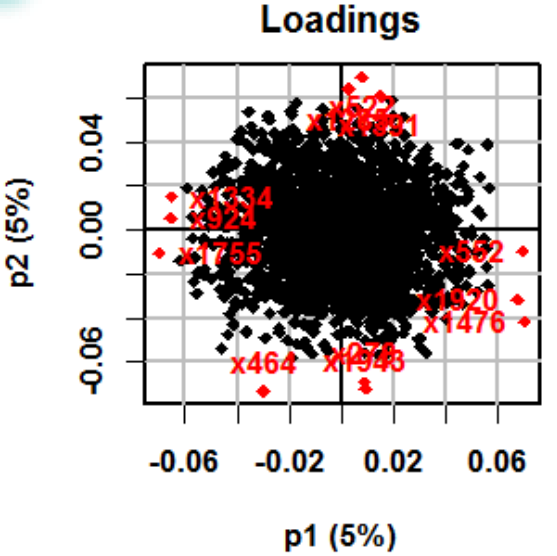
Importance of diagnostics

- **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**





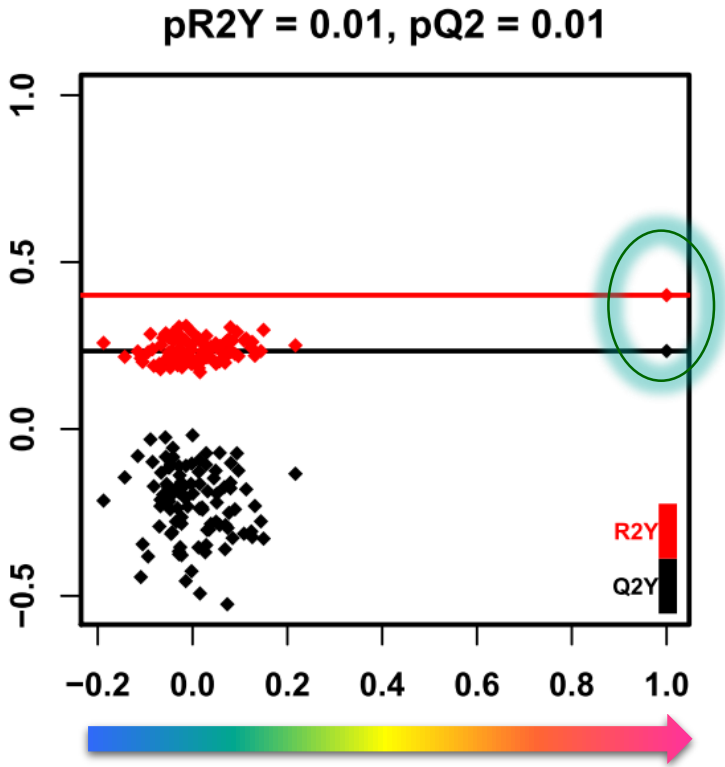




Szymanska E., Saccenti E., Smilde A. and Westerhuis J. (2012). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics,* **8**:3-16. DOI:

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**



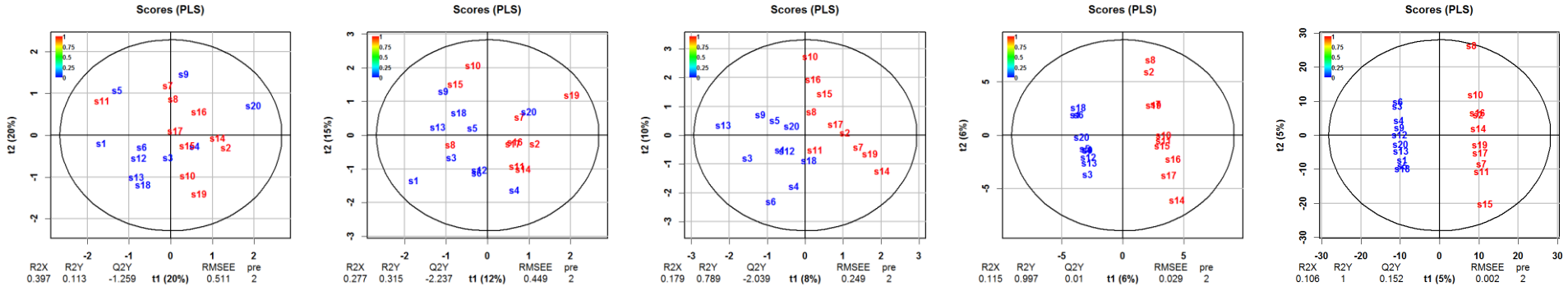
$R2Y$ and $Q2Y$ of the model with the true $\mathbf{y}$ values

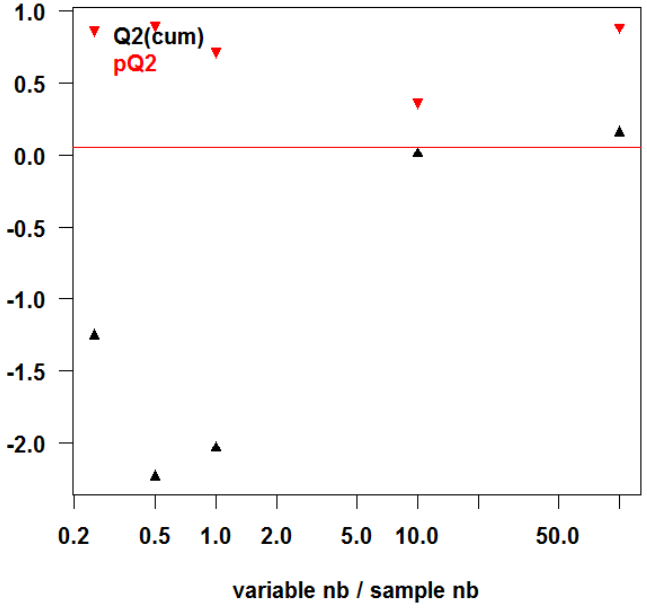
'permutation'

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

Risk of overfitting when *n* < *p*

$$\frac{\textbf{variables}}{\textbf{samples}} = \quad 0.2 \qquad 0.5 \qquad 1 \qquad 10 \qquad 100$$





Single omics data analysis workflow

SciδophenIA

phenomis
Thévenot *et al.*

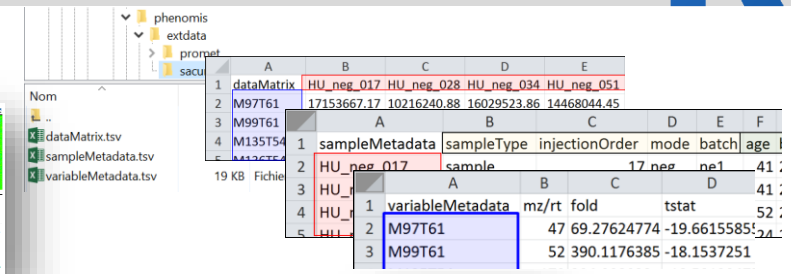
ropls
Thévenot *et al.*, 2015

biosigner
Rinaudo *et al.*, 2016

biodb
Roger *et al.*

Importing

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



Post-processing

Quality control

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

Signal drift and batch effect correction

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

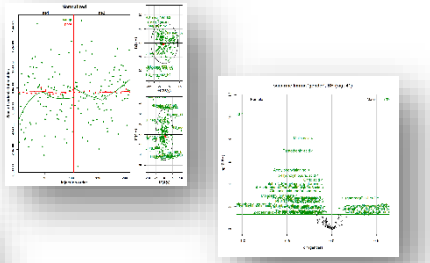
Transformation

eSet <- **phenomis::transforming**(eSet, methodC = 'log2')

Statistics

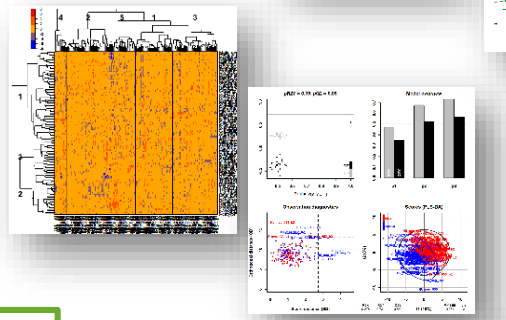
Univariate hypothesis testing

eSet <- **phenomis::hypotesting**(eSet, testC = 'limma',
factorNamesVc = 'gender', adjustC = 'BH')



PCA

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



Heatmap

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

(O)PLS(-DA)

setPlsda <- **ropls::opls**(eSet, 'gender')
eSet <- **ropls::getEset**(setPlsda)

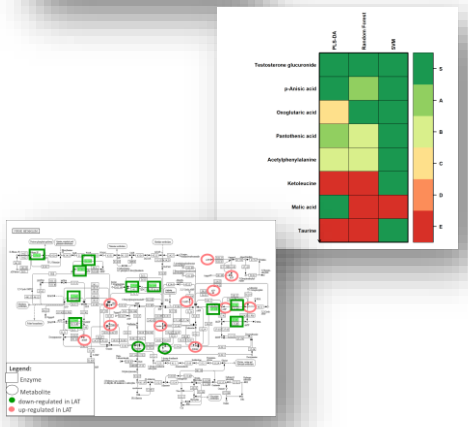
Feature selection

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())

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

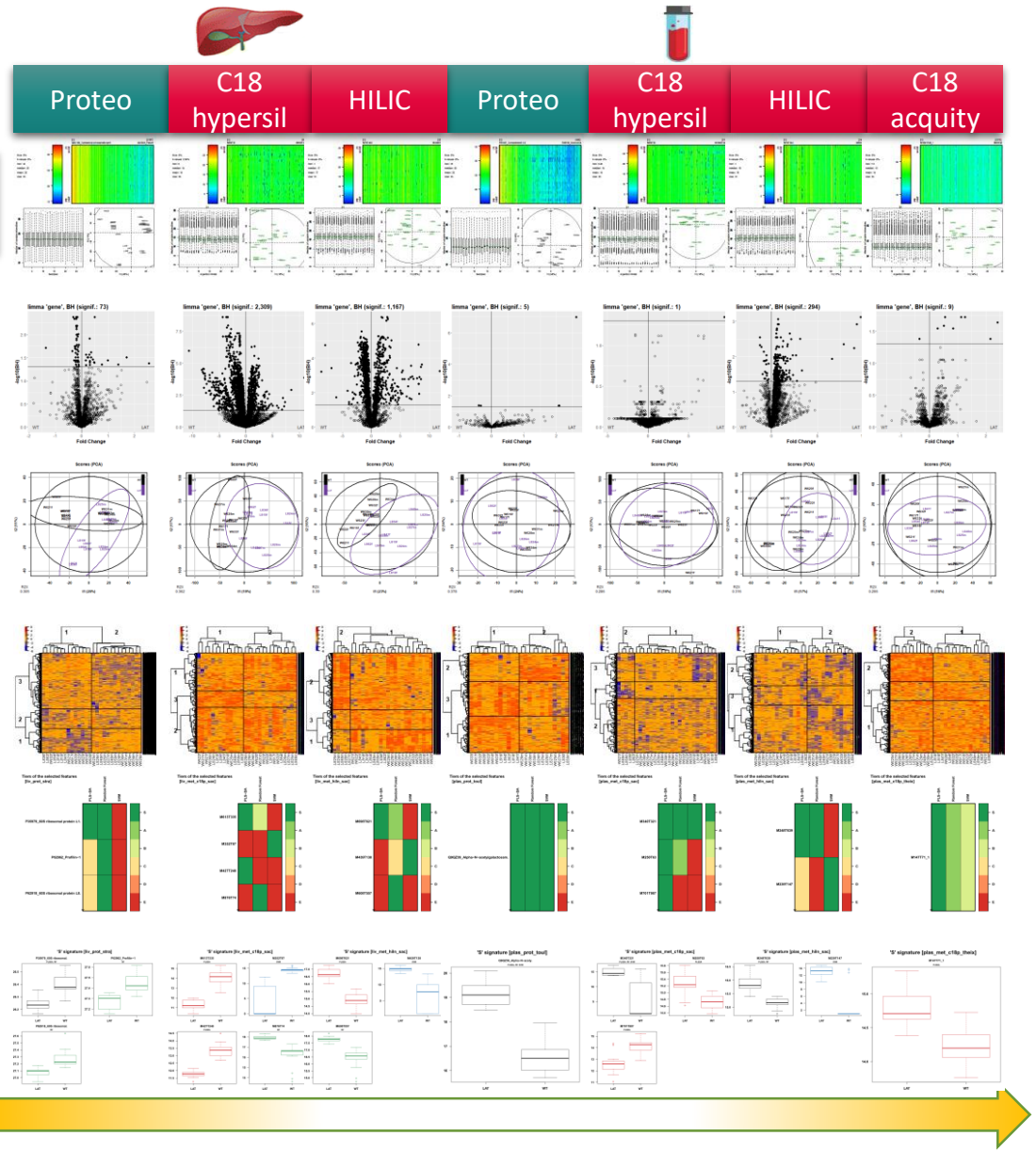
Importing

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

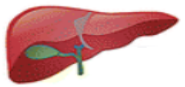
Statistics
- Univariate hypothesis testing
- PCA
- Heatmap
- (O)PLS(-DA)
- Feature selection

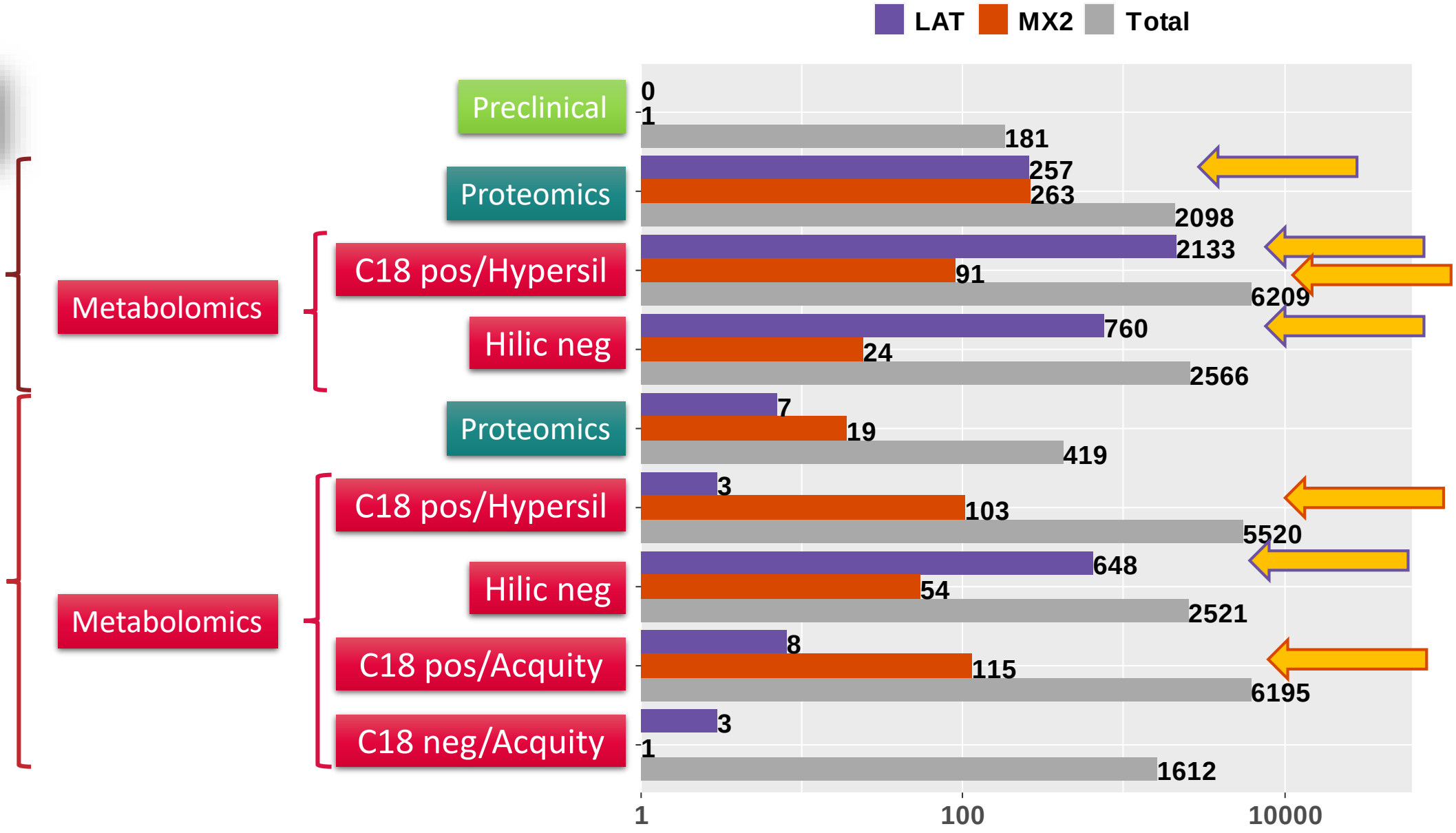
Annotation
- Chemical annotation

Exporting

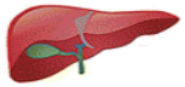


Significant features KO vs WT (limma)



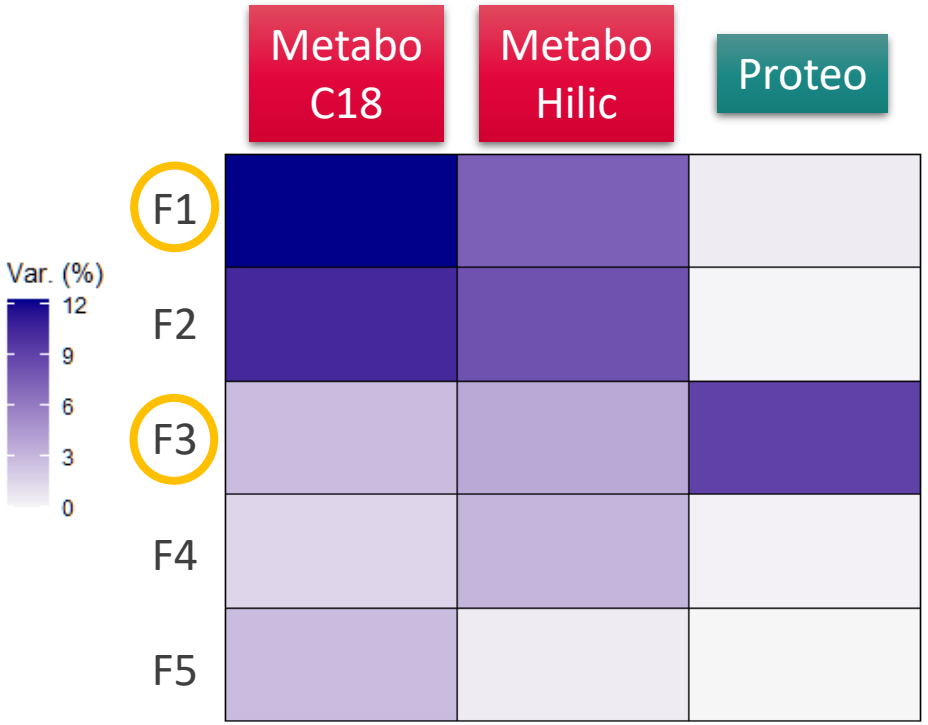


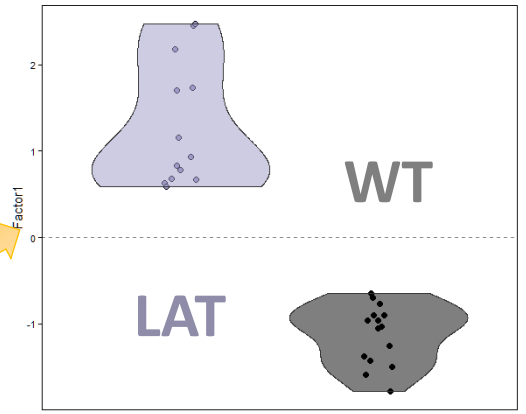
Multi-Omics Factor Analysis

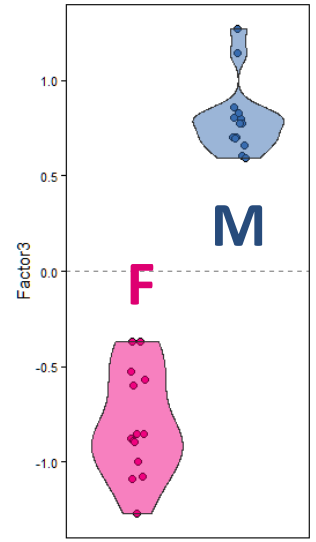


LAT vs WT

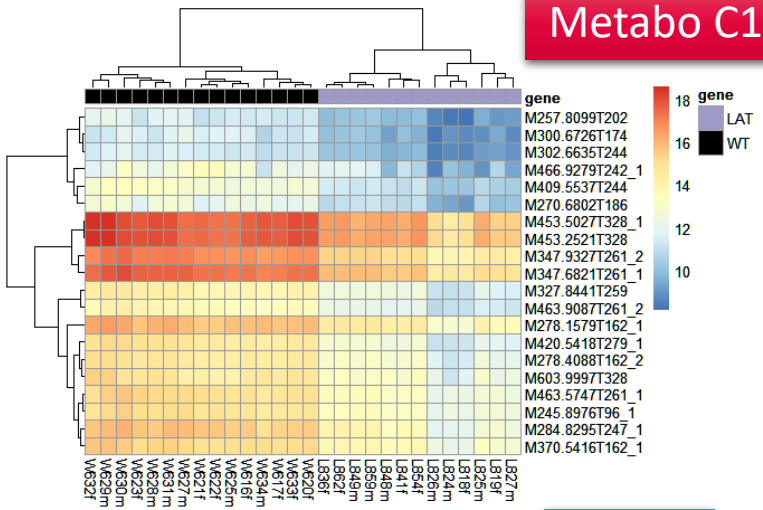
Argelaguet *et al.* (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, **14**.



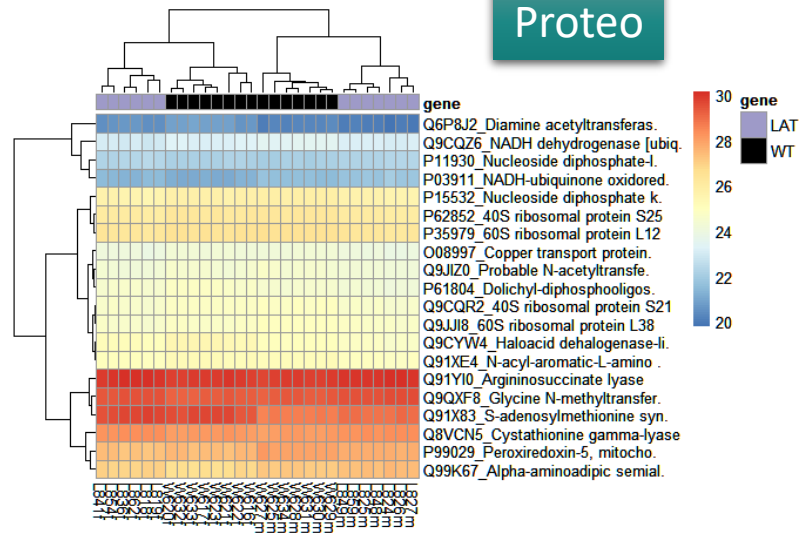




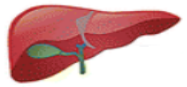
Metabo C18



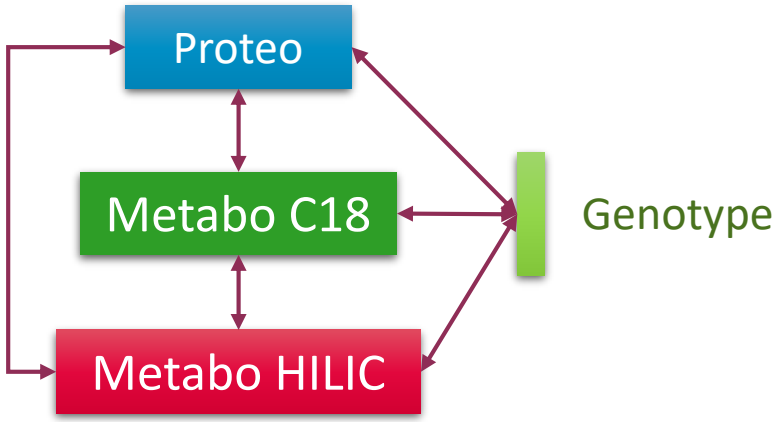
Proteo

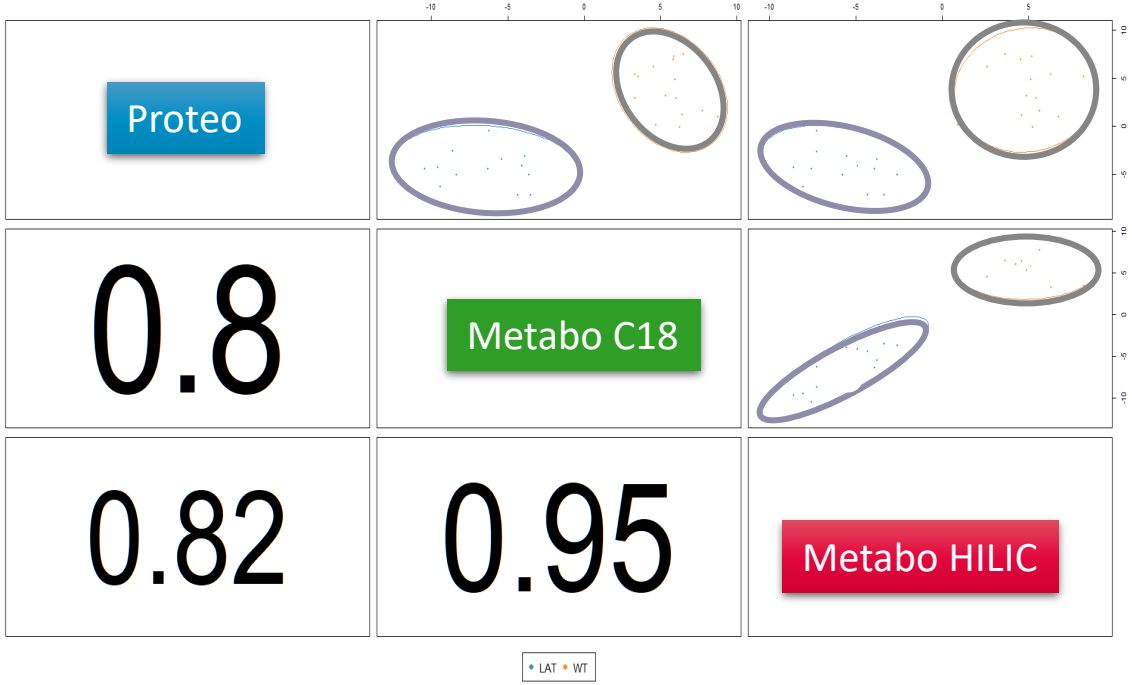


Sparse Generalized Canonical Correlation Analysis - Discriminant Analysis



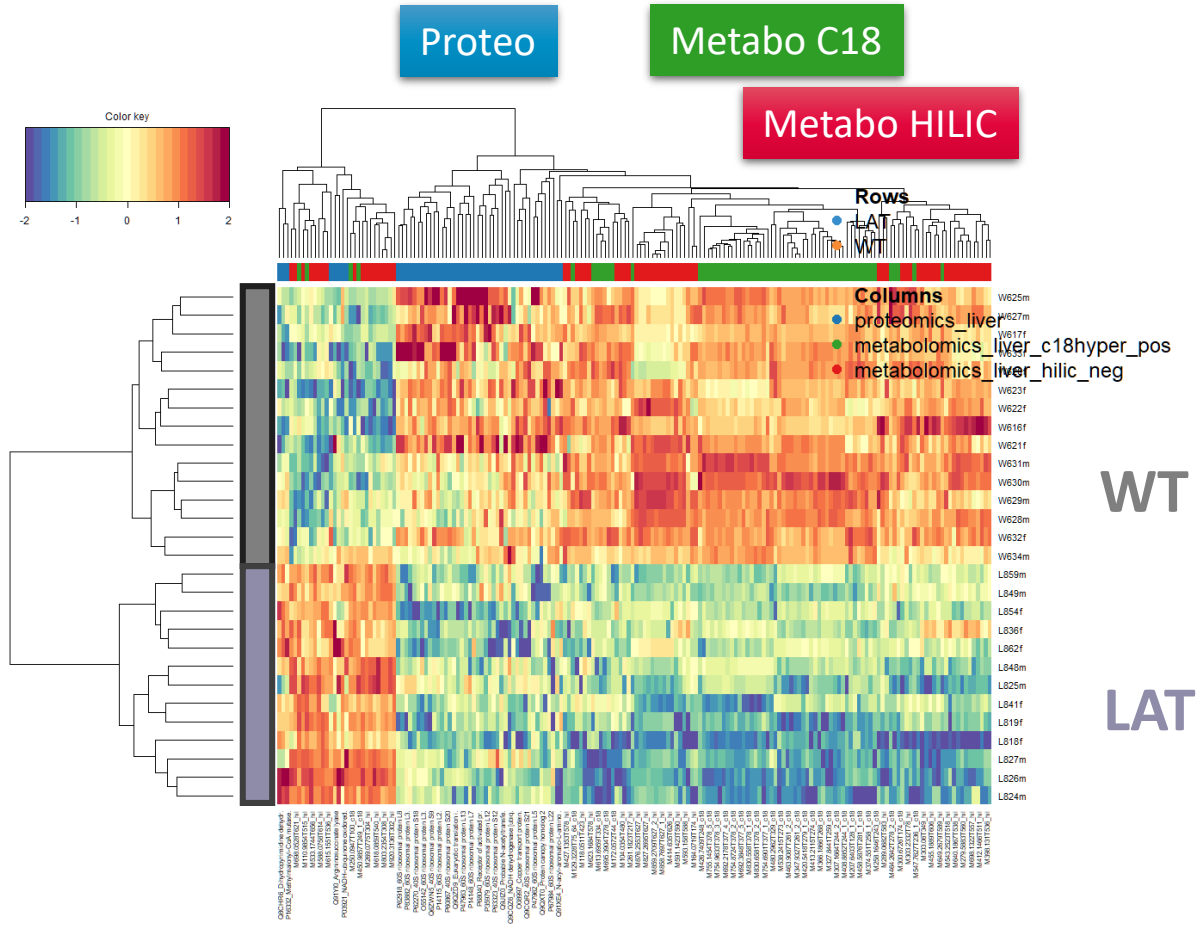
LAT vs WT





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.



Conclusions

- **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

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

Marion Brandolini
Charlotte Joly
Estelle Pujos-Guillot

Magali Rompais
Christine Carapito

Emmanuelle Mouton
Anne Gonzalez de Peredo

Thomas Burger
Yves Vandenbrouck
Myriam Ferro

https://scidophenia.github.io/

SciδophenIA

Mohammed Selloum
Tania Sorg
Sophie Leblanc
Yann Herault

Olivier Sand
Jacques van Helden
Claudine Médigue