https://jipromet-2020.sciencesconf.org

3e Journée Intégrative de Protéomique et Métabolomique
(8 octobre 2020 en visioconférence)

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

etienne.thevenot@cea.fr

DE LA RECHERCHE À L'INDUSTRIE

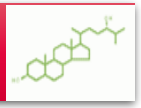
Proteomics and metabolomics data integration: Where do we stand?

What?

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”

Why?

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

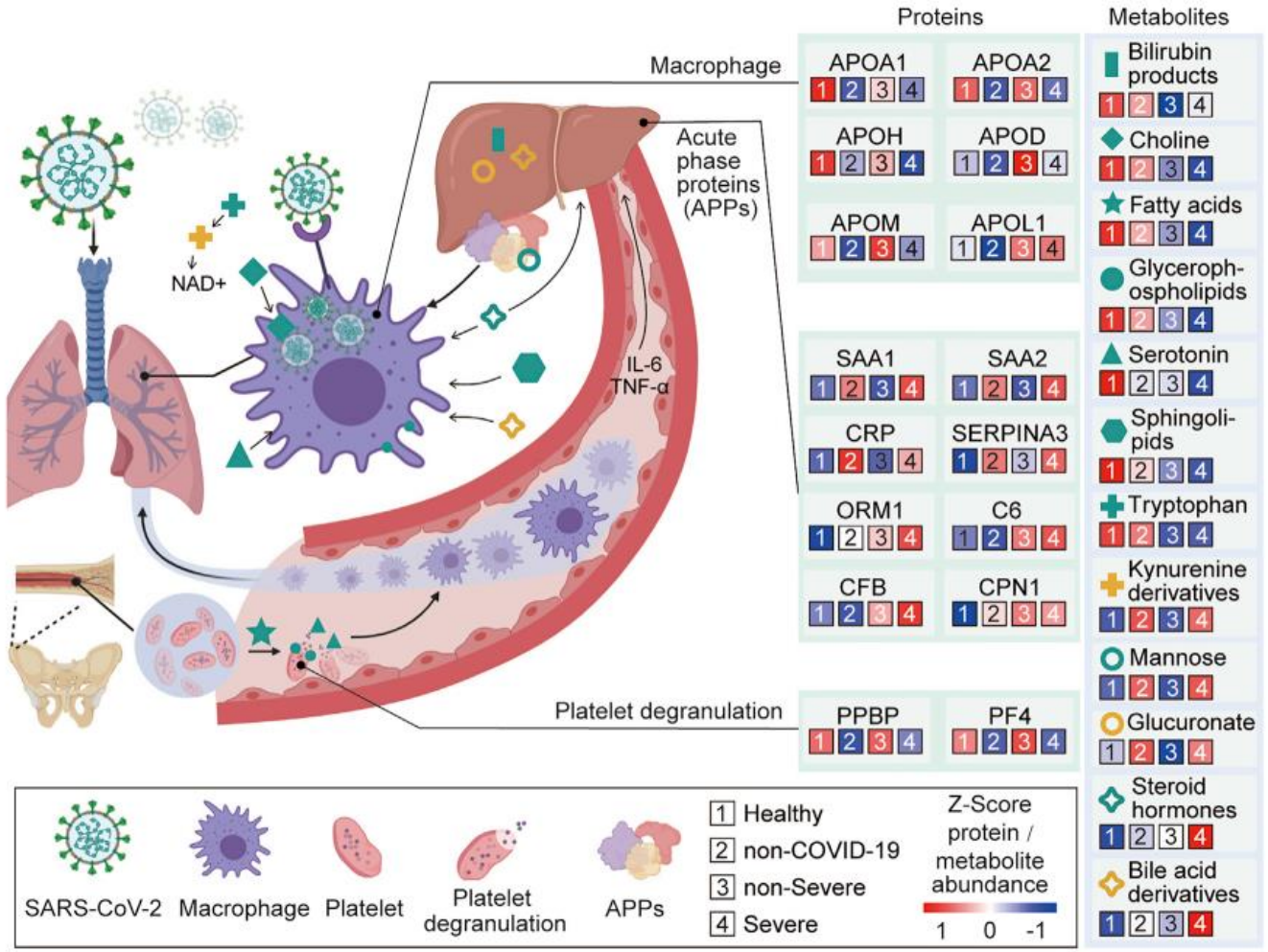
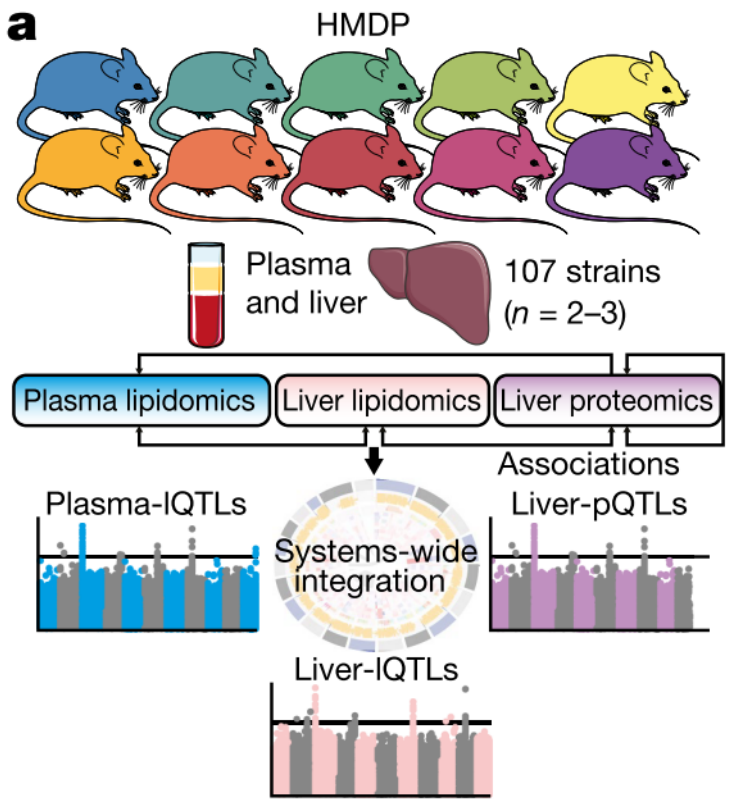


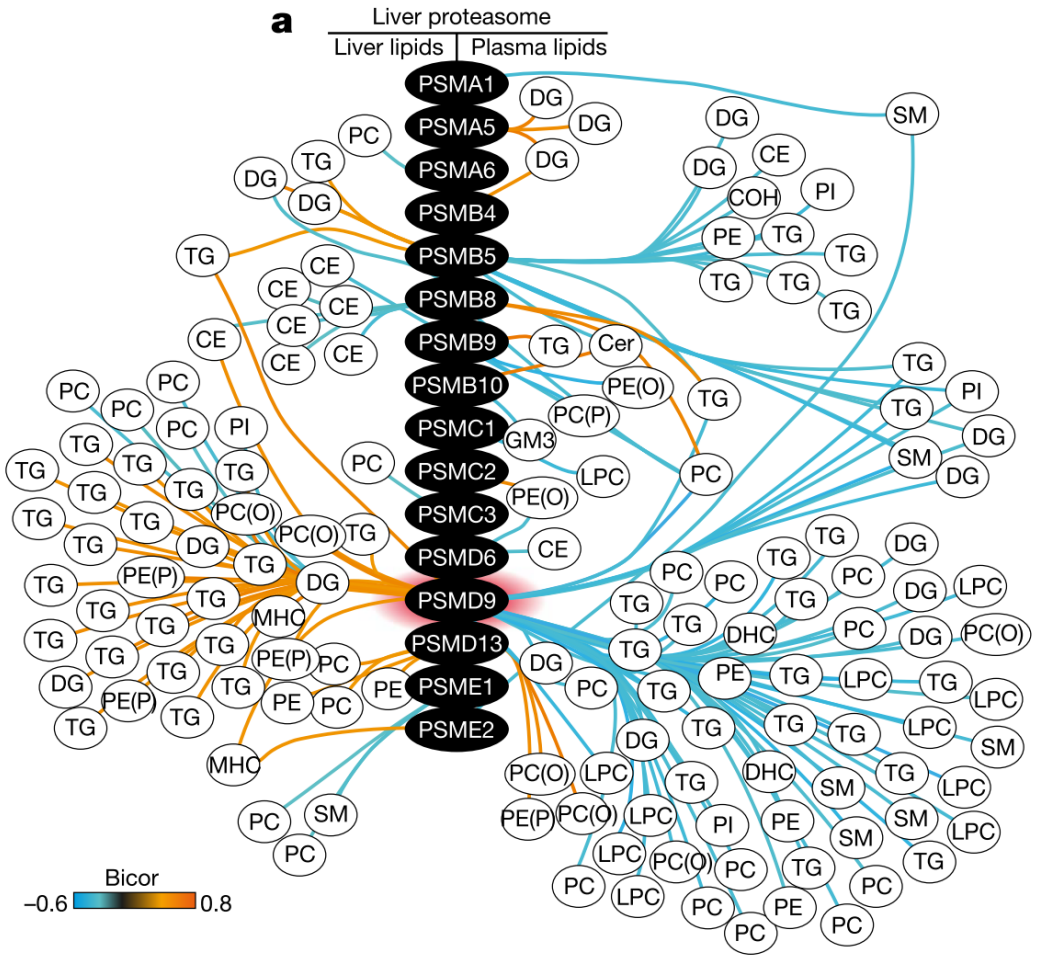
Figure 5. Key Proteins and Metabolites Characterized in Severe COVID-19 Patients in a Working Model

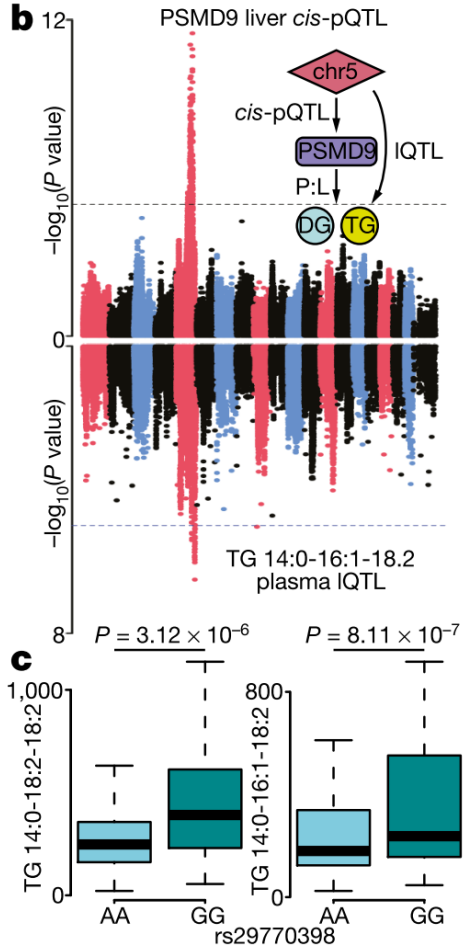
SARS-CoV-2 may target alveolar macrophages via ACE2 receptor, leading to an increase of secretion of cytokines including IL-6 and TNF-α, which subsequently induce the elevation of various APPs such as SAP, CRP, SAA1, SAA2, and C6, which are significantly upregulated in the severe group. Proteins involved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

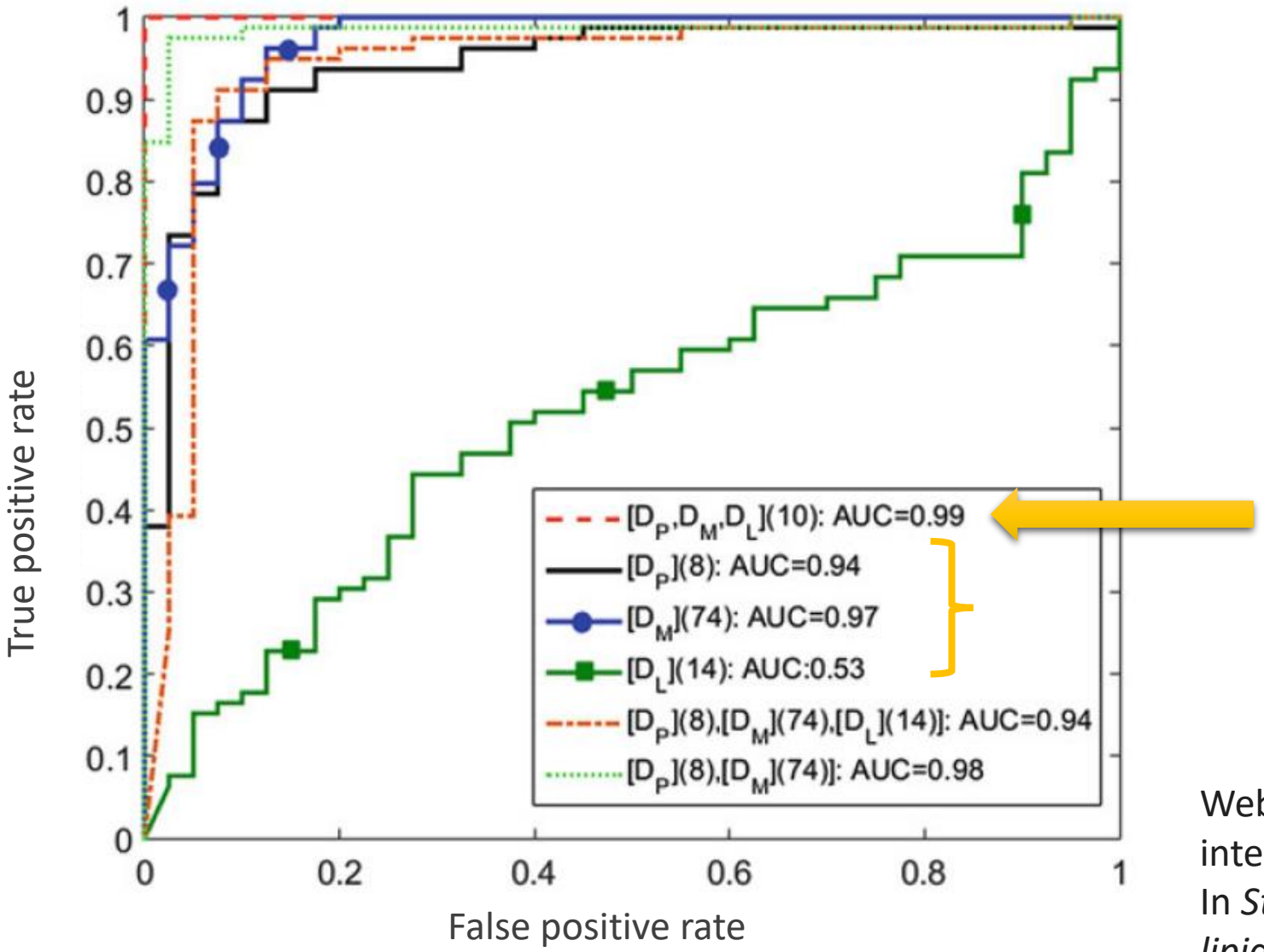






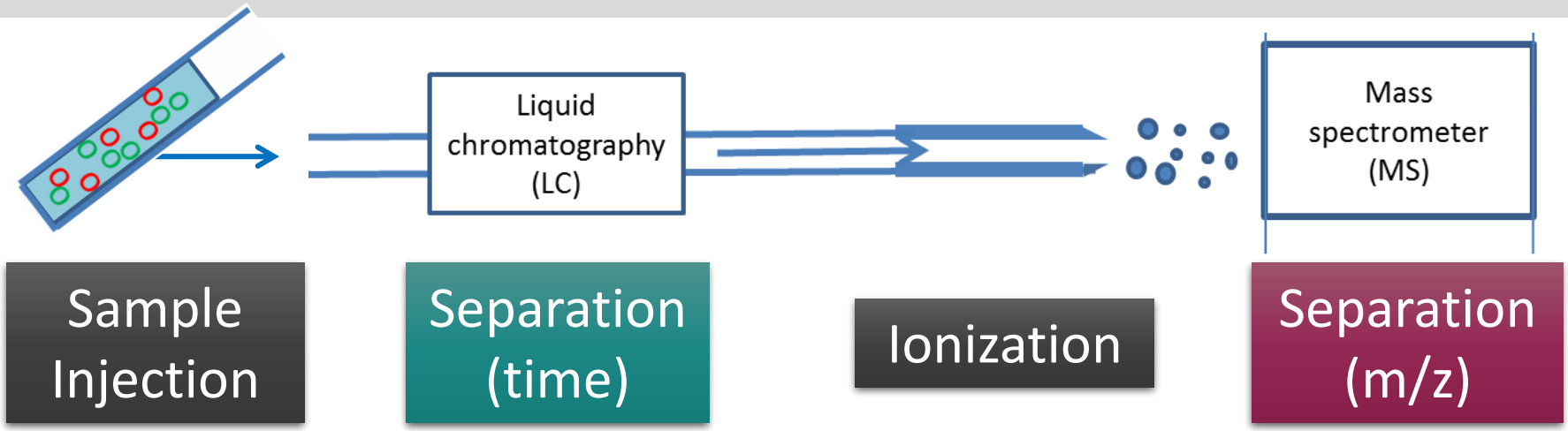
Parker *et al.* (2019). An integrative systems genetic analysis of mammalian lipid metabolism. *Nature*, **567**:187–193.

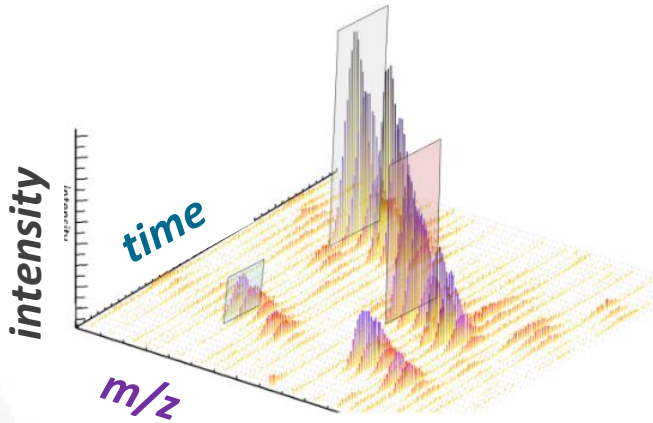
Increase the predictive performance



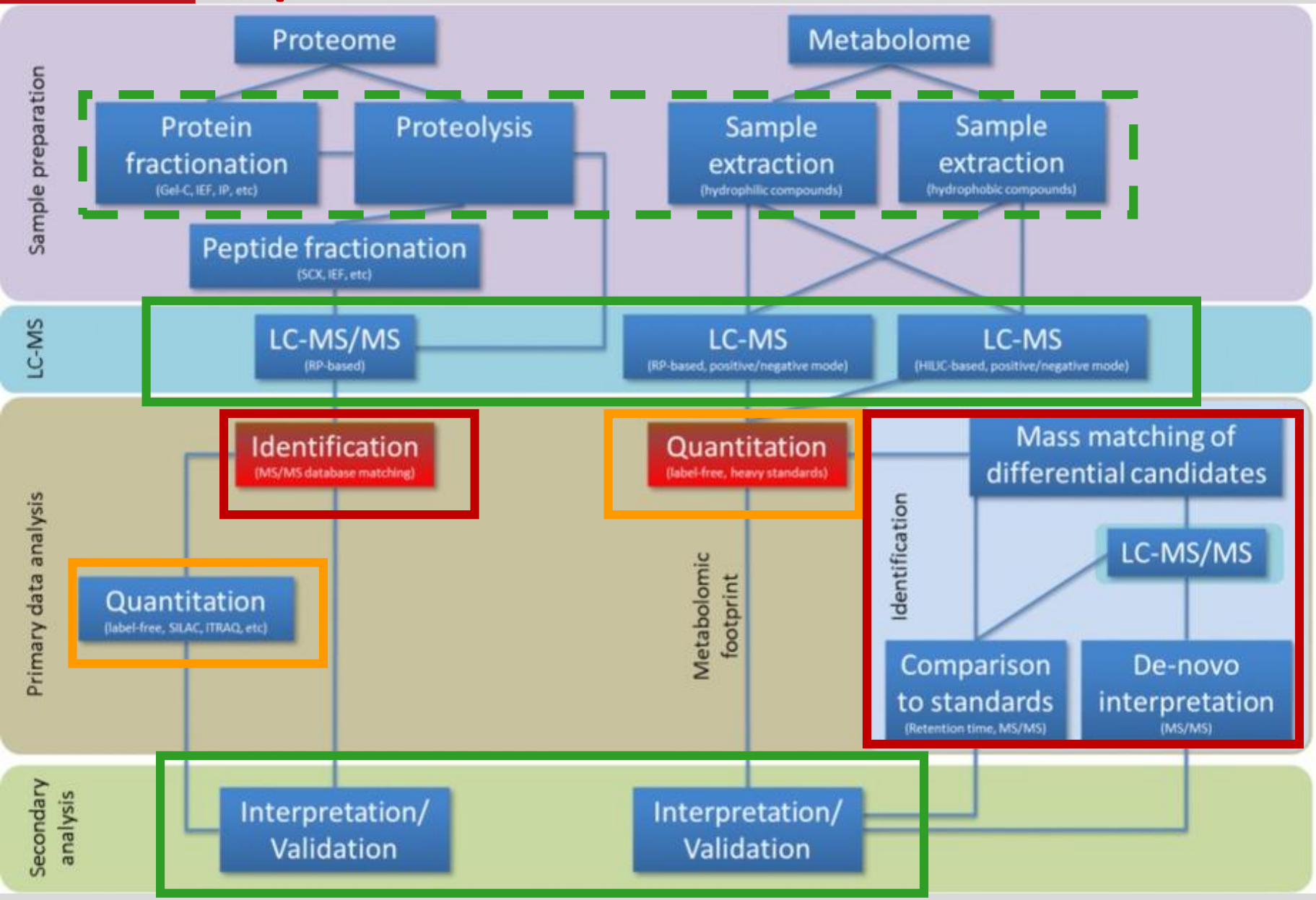
Webb-Robertson *et al.* (2016). Bayesian posterior integration for classification of mass spectrometry data. In *Statistical analysis of proteomics, metabolomics, and lipidomics data using mass spectrometry* (pp. 203–211).

Common technology: LC-HRMS





Similarities in data quantification and statistical analysis
Specificities in data annotation

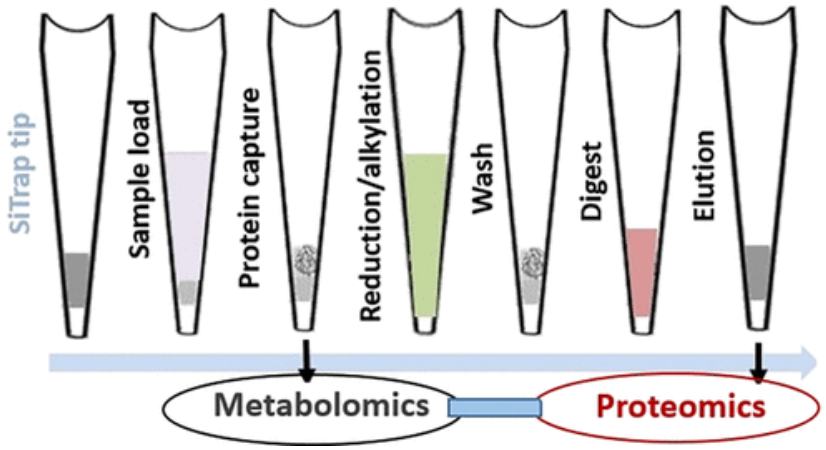


Fischer *et al.* (2013). Two birds with one stone: doing metabolomics with your proteomics kit. *Proteomics*, **13**:3371-3386.

Common sample preparation studies

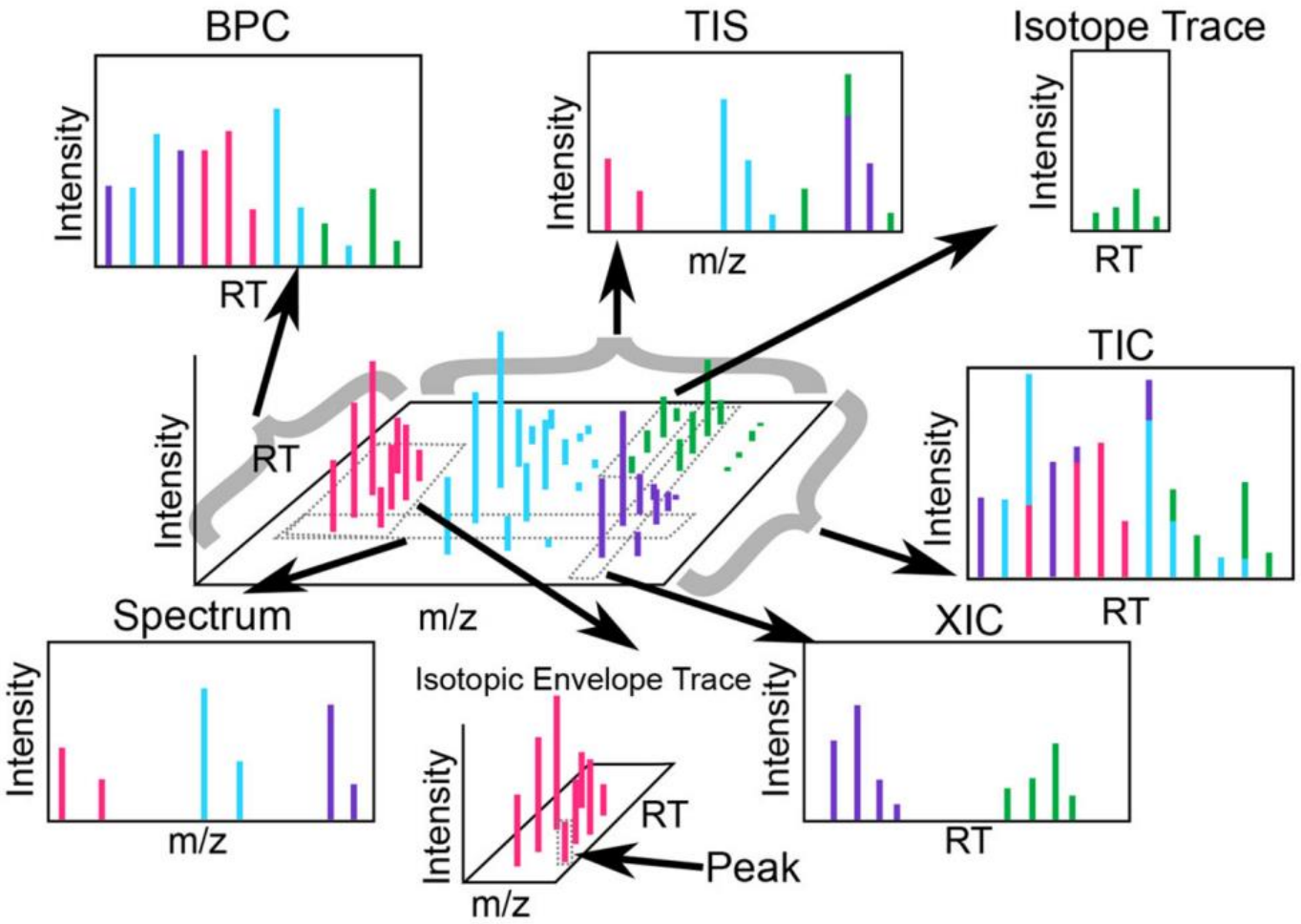
Fischer *et al.* (2013). Two birds with one stone: doing metabolomics with your proteomics kit. *PROTEOMICS*, **13**:3371-3386.

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.

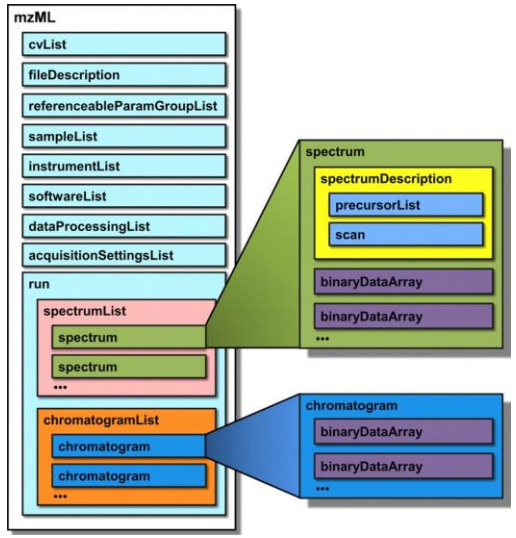
Common nomenclature



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**

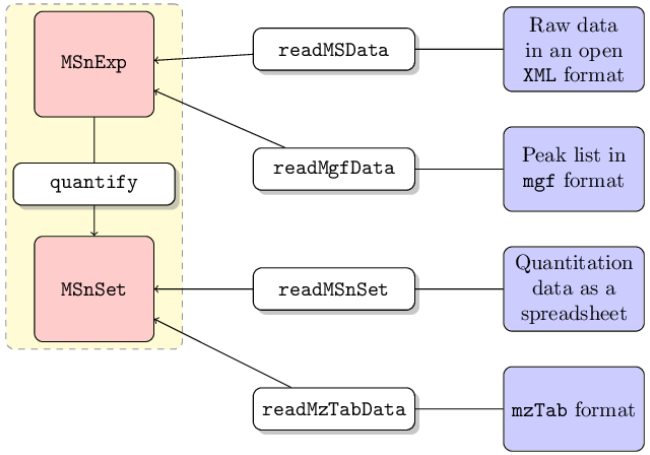


Martens *et al*. (2010). mzML - a community standard for mass spectrometry data. *Molecular & Cellular Proteomics*. **10**.

  - 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**

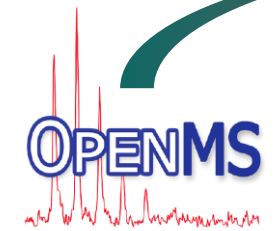


Gatto *et al*. (2020). MSnbase, efficient and elegant R-based processing and visualization of raw mass spectrometry data. *Journal of Proteome Research*.

Toward common software

Proteomics

Metabolomics



phenomis

SDA

WGCNA

ProMetIS

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

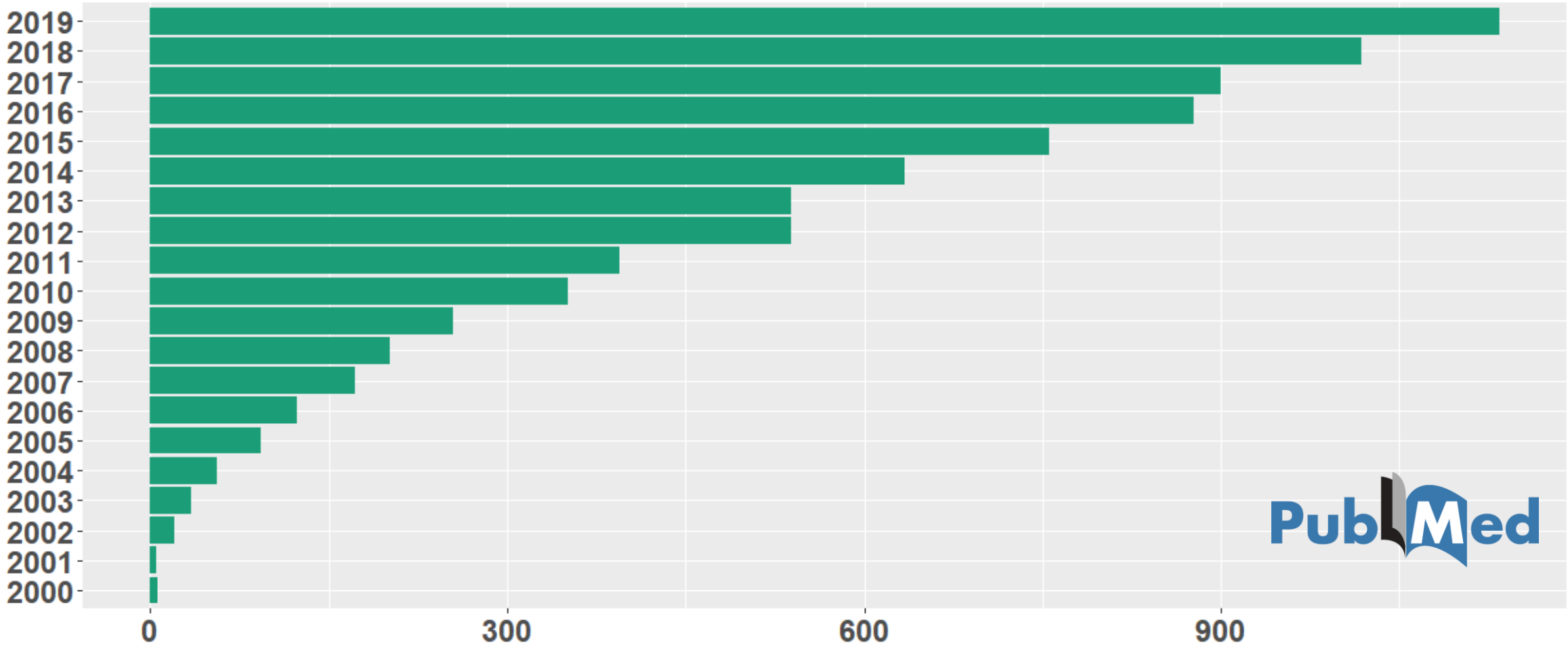
Adams *et al.* (2020). Skyline for small molecules: A unifying software package for quantitative metabolomics. *Journal of Proteome Research*, **19**:1447-1458.

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

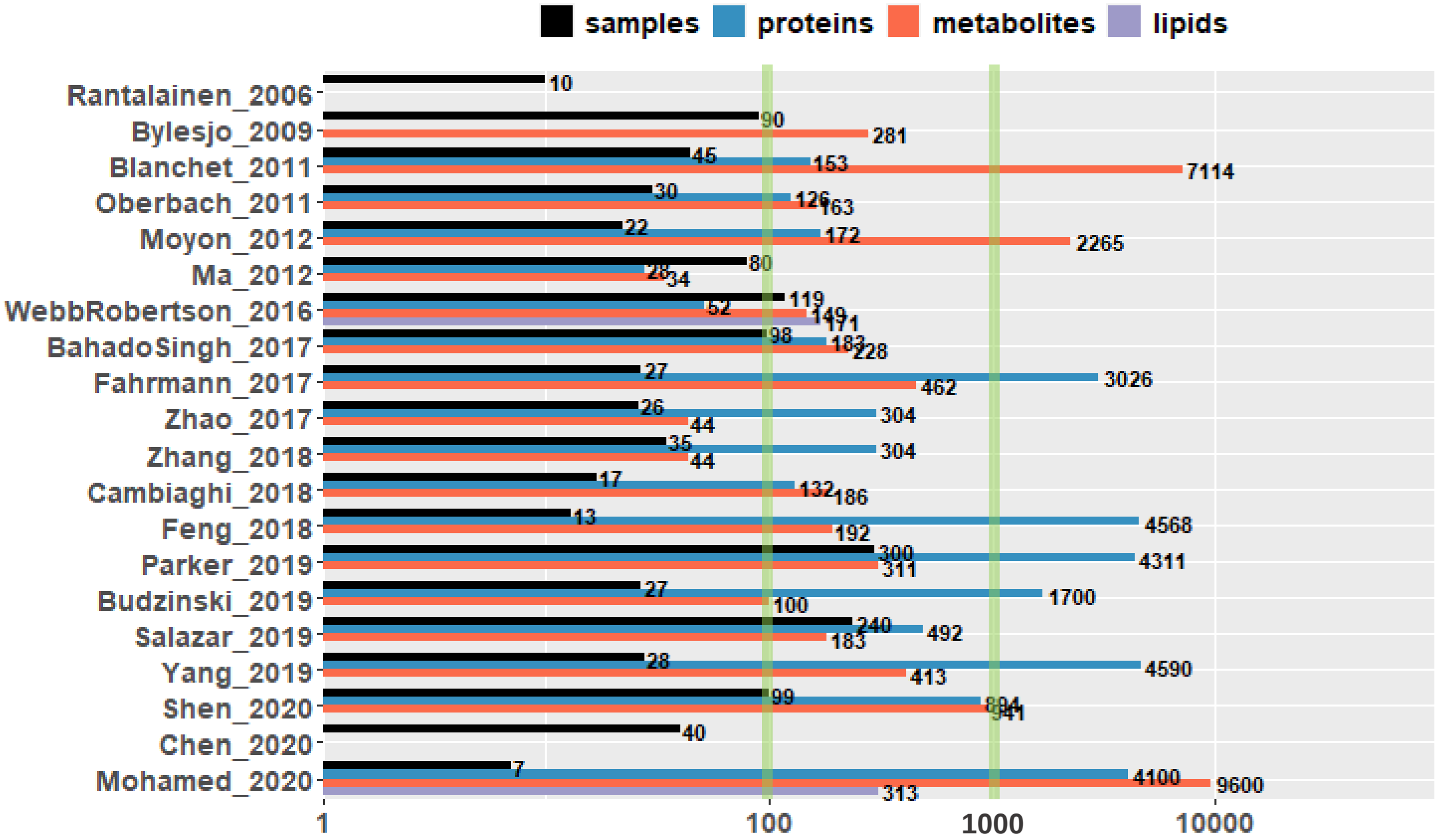
Pei *et al.* (2017). WGCNA application to proteomic and metabolomic data analysis. *Methods in Enzymology*. 135-158.

Who?

Joint proteomics & metabolomics publications



Bibliography



Bibliographic focus

- **criteria:**
  - untargeted proteomics and metabolomics
  - include a statistical data integration methodology
  - open-source implementation

- **selection:**

- Moyon *et al.* (2012). Statistical strategies for relating metabolomics and proteomics data: a real case study in nutrition research area. *Metabolomics*, **8**, 1090–1101.
- 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).
- Cambiaghi *et al.* (2018). An innovative approach for the integration of proteomics and metabolomics data in severe septic shock patients stratified for mortality. *Scientific Reports*, **8**.
- Parker *et al.* (2019). An integrative systems genetic analysis of mammalian lipid metabolism. *Nature*, **567**, 187–193.
- Shen *et al.* (2020). Proteomic and metabolomic characterization of COVID-19 patient sera. *Cell*, **9**, 59–72.

Bibliographic focus: overview

Reference	Disease	Species	Sample	Samples	Clinics	Proteins	Metabolites	Lipids	Data integration
Moyon et al., 2012	maternal diet impact on offspring	rat	plasma, hypothalamus	22		172	2265		PCA, RCCA, MB-PLS
WebbRobertson et al., 2016	diabetes	human	serum	119		52	149	171	concatenation + (RFE +) NB; LDA and RF fusion
Cambiaghi et al., 2018	severe septic shock	human	plasma	17	17	132	186		feature selection (mRMR) + concatenation + classification (elastic net, LDA, PLS-DA)
Parker et al., 2019	hepatosteatosis, metabolic syndrome	mouse	liver, plasma	300		4311	311		clustering based representative feature selection, lasso
Shen et al., 2020	COVID-19	human	serum	99	12	894	941		random forest, mFuzz, IPA

How?

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?

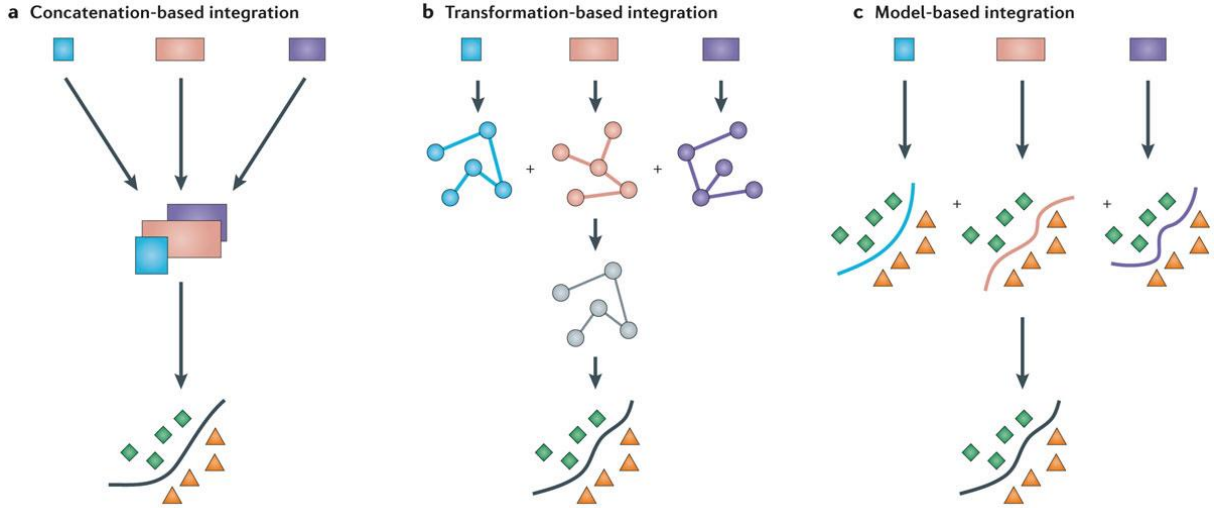
Data integration: challenges

- 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

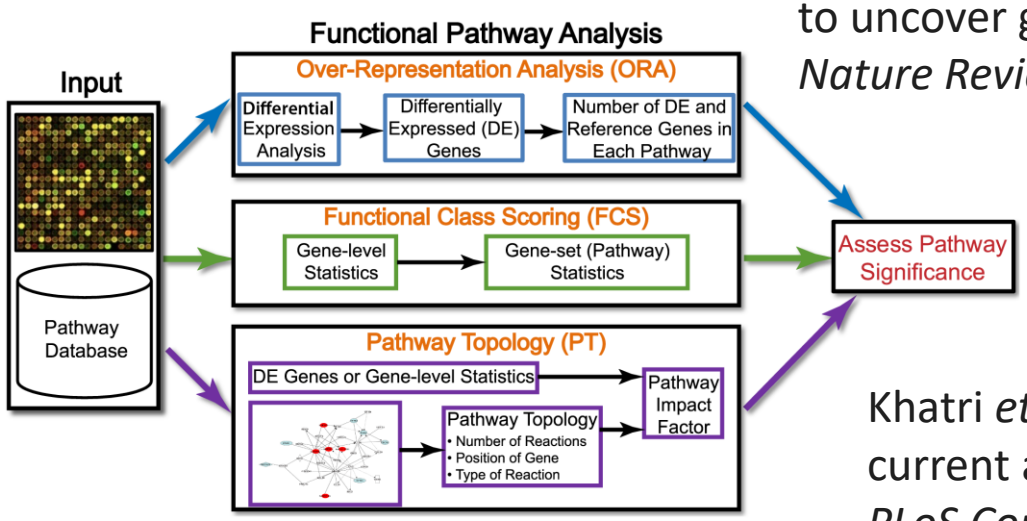
Data integration: approaches

- **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.

Data integration: (some) published approaches

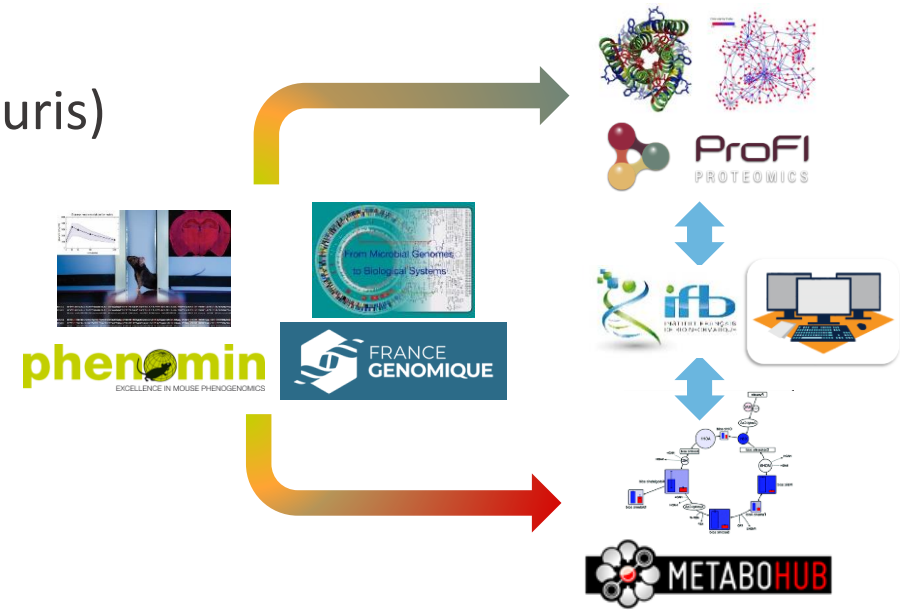
- **Biostatistics**
  - Fusion
    - low
    - middle
    - High
- **Bioinformatics**
  - Mapping
  - Enrichment
    - Molecule set
    - Topology-based

	Unsupervised	Supervised
low		
middle	O2PLS (Bylesjö et al., 2009) RGCCA (Budzinski et al., 2019)	Multiblocks PLS (Moyon *et al.*, 2012) PLS-DA + PCA (Blanchet *et al.*, 2016) Elastic Net + Elastic Net (Ghaemi *et al.*, 2018)
High		LDA, RF (Webb-Robertson *et al.*, 2016)
Molecule set		IMPaLA (Bahado-Singh *et al.*, 2017)
Topology-based		

ProMetIS: deep phenotyping of mouse models by proteomics and metabolomics

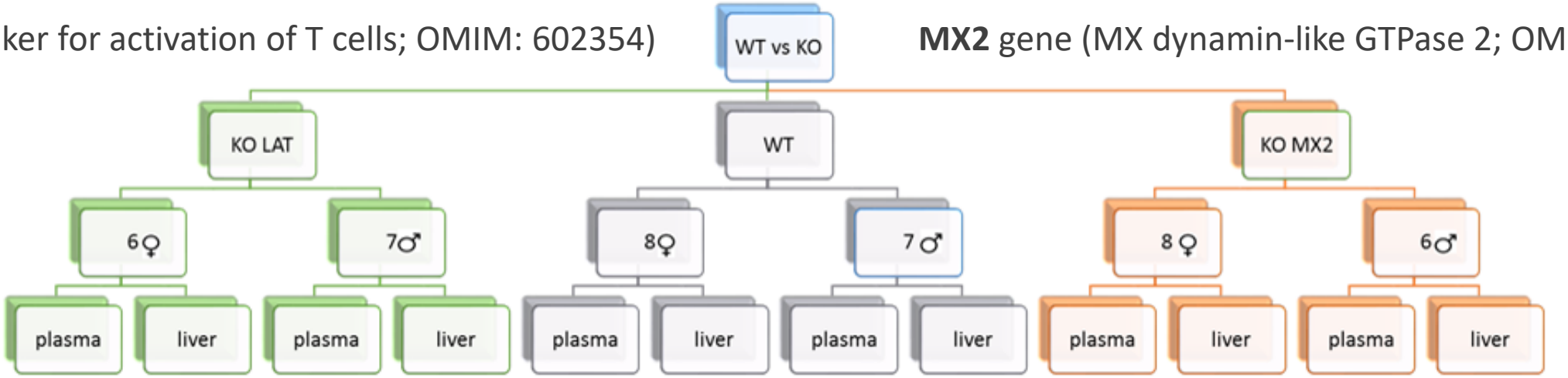
The ProMetIS project

- **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)



MX2 gene (MX dynamin-like GTPase 2; OMIM: 147890)

- **LAT involved in:**
  - T-cell receptor (TCR) signaling
  - Neurodevelopmental diseases

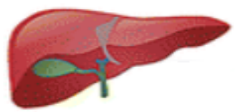
Roncagalli et al. (2010). LAT signaling pathology: an ”autoimmune” condition without T cell self-reactivity. *Trends in Immunology*, **31**:253–259.

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.

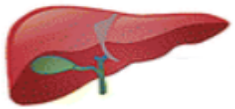
9 preclinical, proteomic and metabolomic datasets

Preclinical

Proteomics



Metabolomics



C18 pos Hypersil	C18 pos Hypersil	C18 pos Acquity
Hilic neg	Hilic neg	C18 neg Acquity

Post-processing

Preclinical

NAs < 20%
Variance > 0

213

Proteomics

log2 transformation

NA < 20% for at least 1 group

Filtering contaminants

Imputation

2,098 | 422

Metabolomics

Blank filtering

Pool dilution

Signal drift correction

no NA for at least 1 group
Variance > 0

Pool CV ≤ 0.3

Pool CV / sample CV ≤ 1

Chemical redundancy

log2 transformation

5,195	5,476	4,804
2,062	1,844	1,449

Single omics data analysis workflow

SciδophenIA

phenomis
Thévenot *et al.*

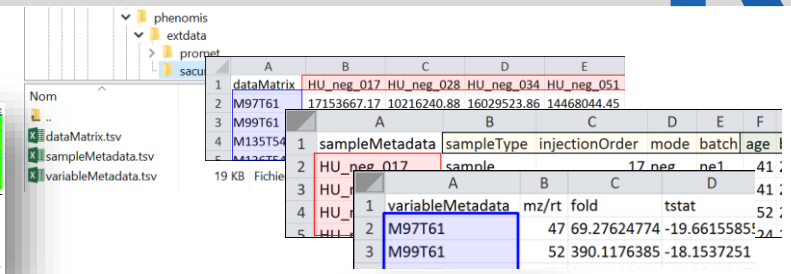
ropls
Thévenot *et al.*, 2015

biosigner
Rinaudo *et al.*, 2016

biodb
Roger *et al.*

Importing

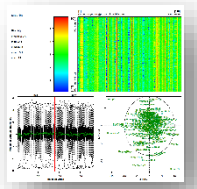
eSet <- **phenomis::reading**(dirC)



Post-processing

- Quality control

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



- Signal drift and batch effect correction

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

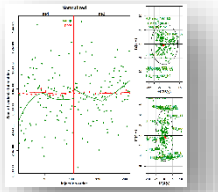
- Transformation

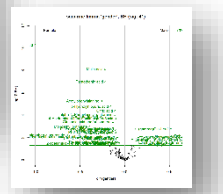
eSet <- **phenomis::transforming**(eSet, methodC = 'log2')

Statistics

- Univariate hypothesis testing

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



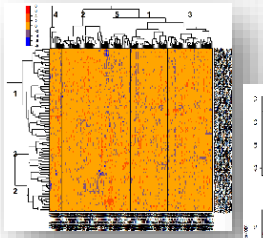


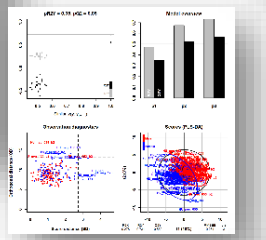
- PCA

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

- Heatmap

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



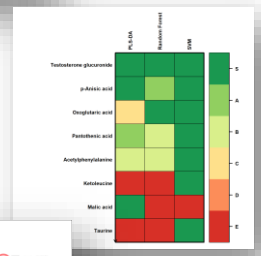


- (O)PLS(-DA)

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

- Feature selection

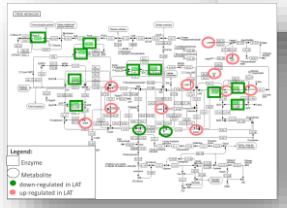
setBiosign <- **biosigner::biosign**(eSet, 'gender', seedI = 123)
eSet <- **biosigner::getEset**(setBiosign)



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)

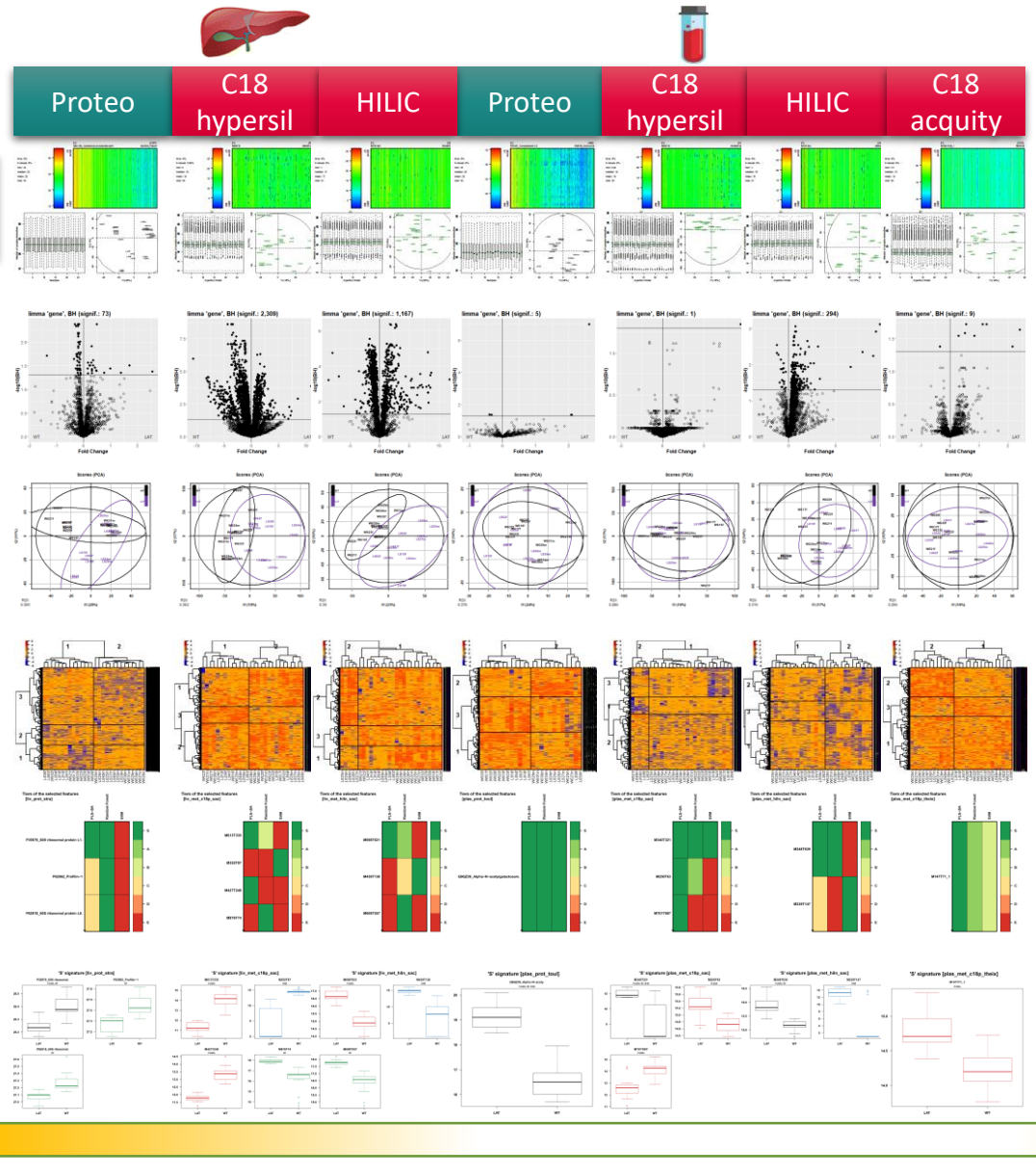
Importing

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

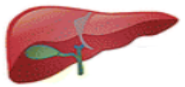
Statistics
- Univariate hypothesis testing
- PCA
- Heatmap
- (O)PLS(-DA)
- Feature selection

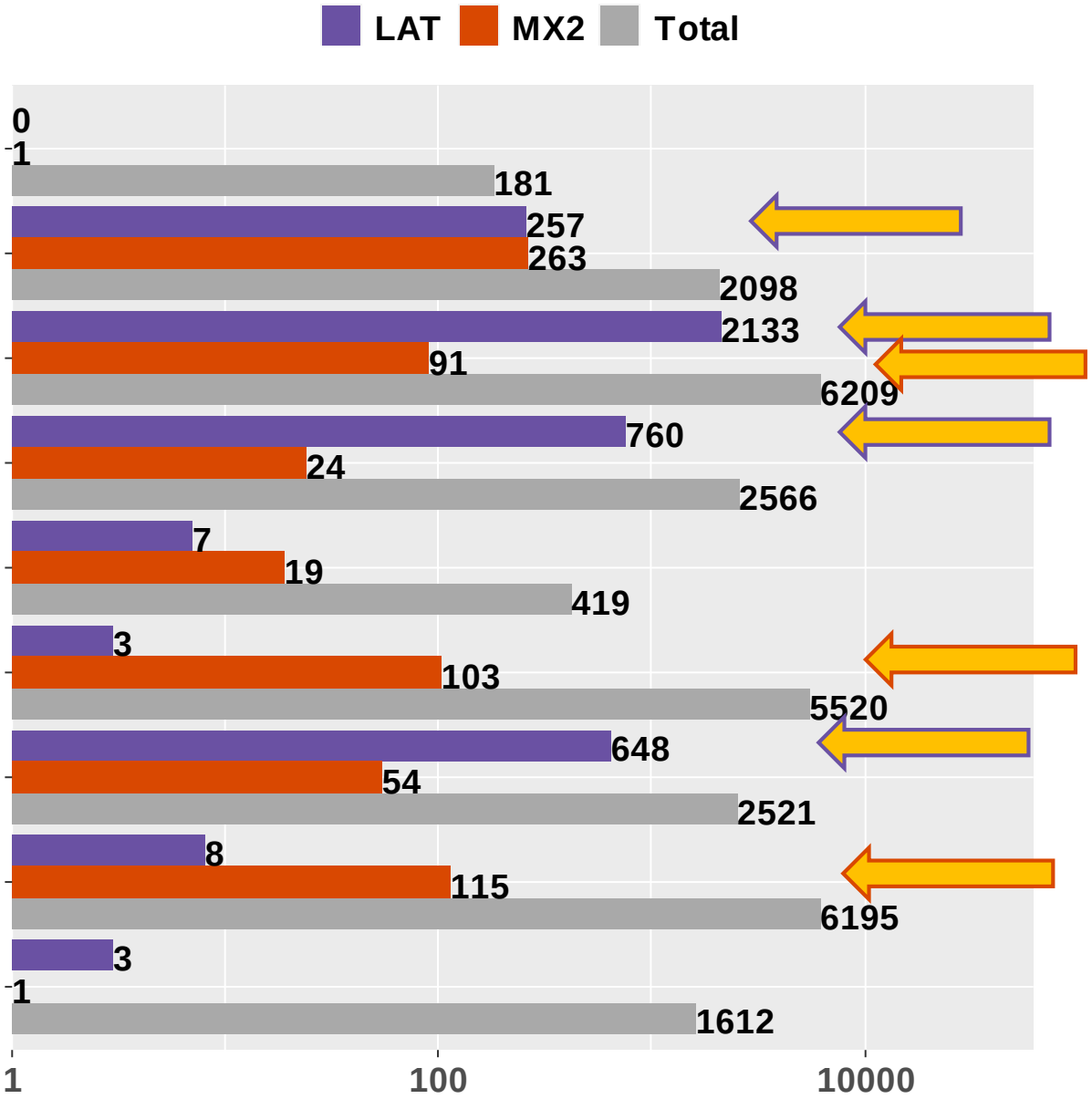
Annotation
- Chemical annotation

Exporting

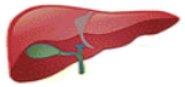


Significant features KO vs WT (limma)



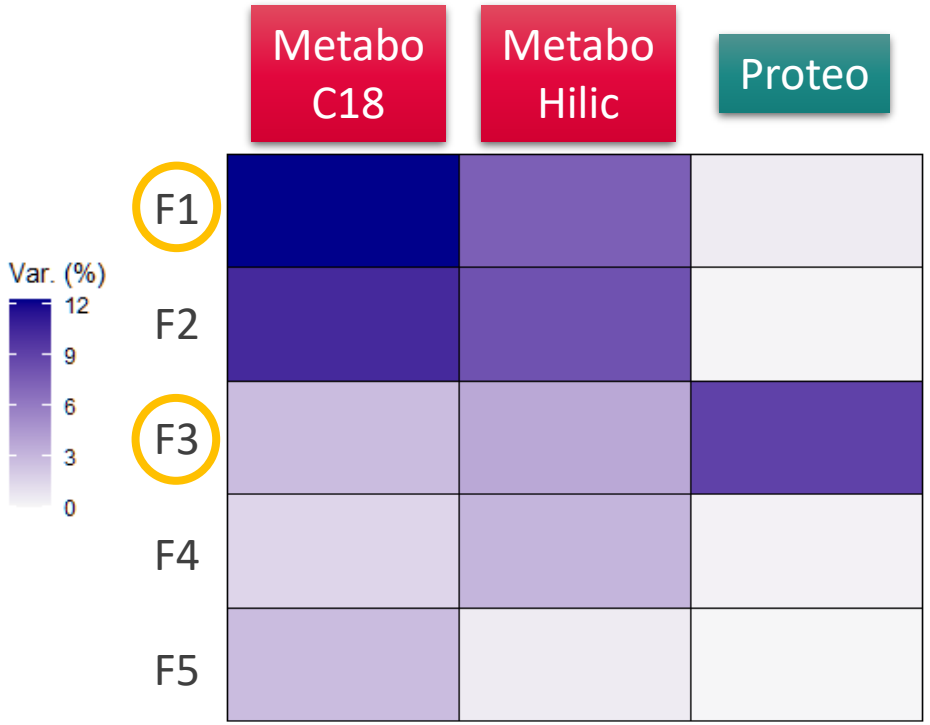


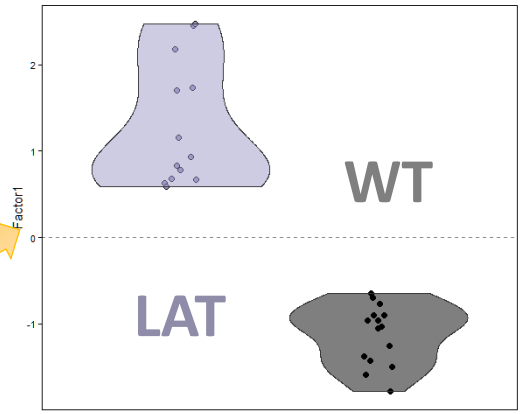
Multi-Omics Factor Analysis

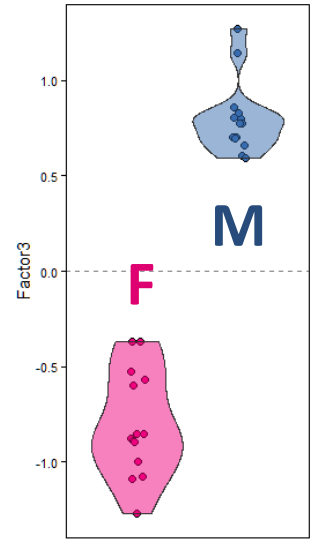


LAT vs WT

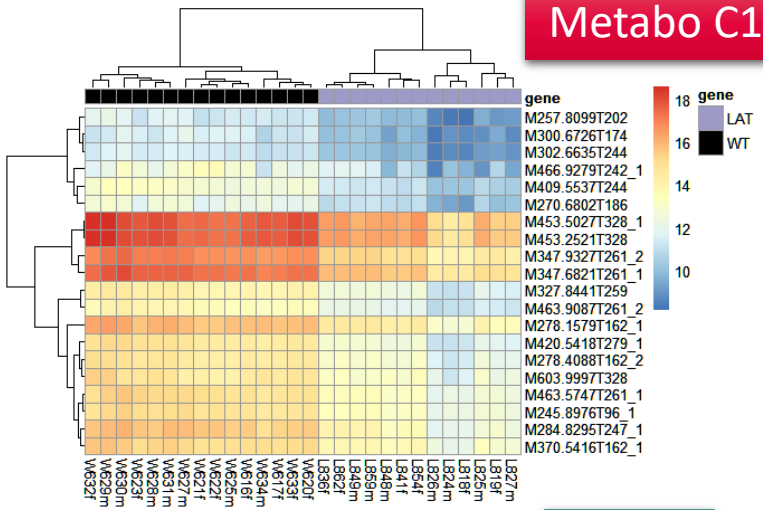
Argelaguet *et al.* (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, **14**.



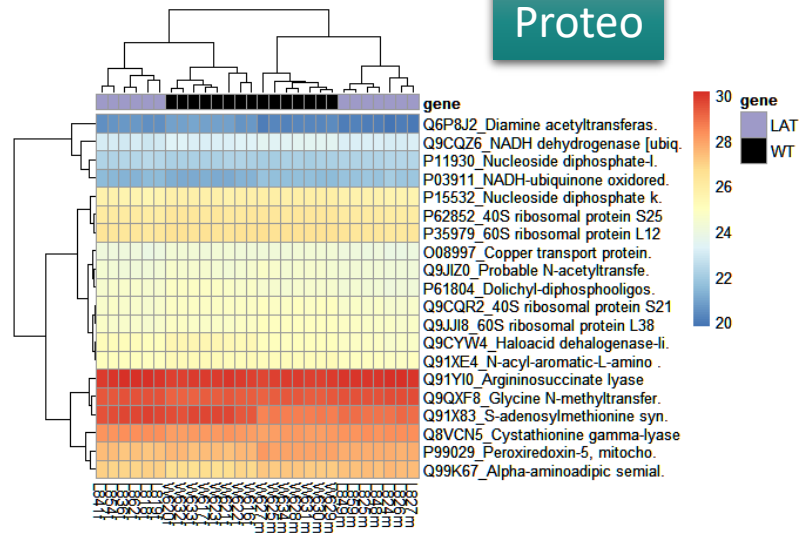




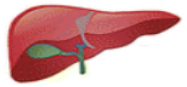
Metabo C18



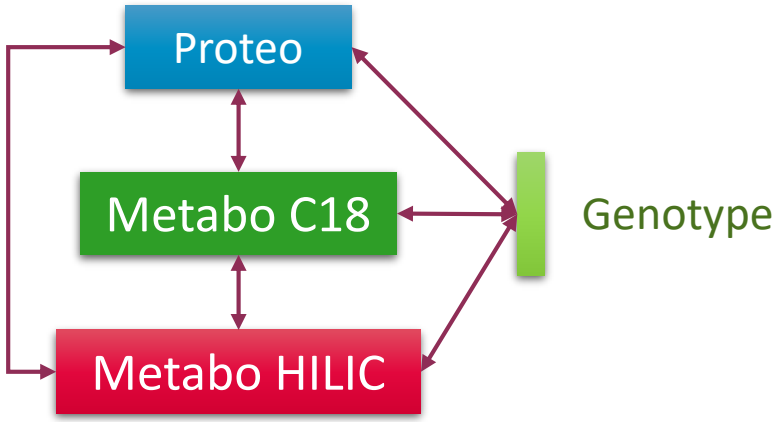
Proteo

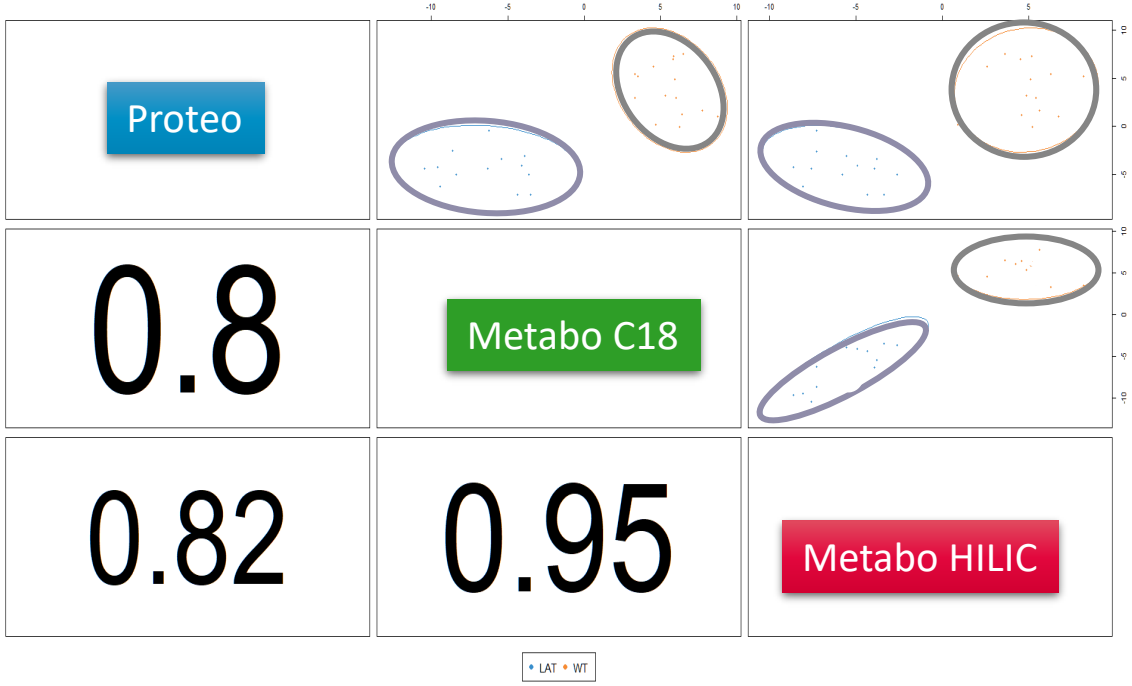


Sparse Generalized Canonical Correlation Analysis - Discriminant Analysis



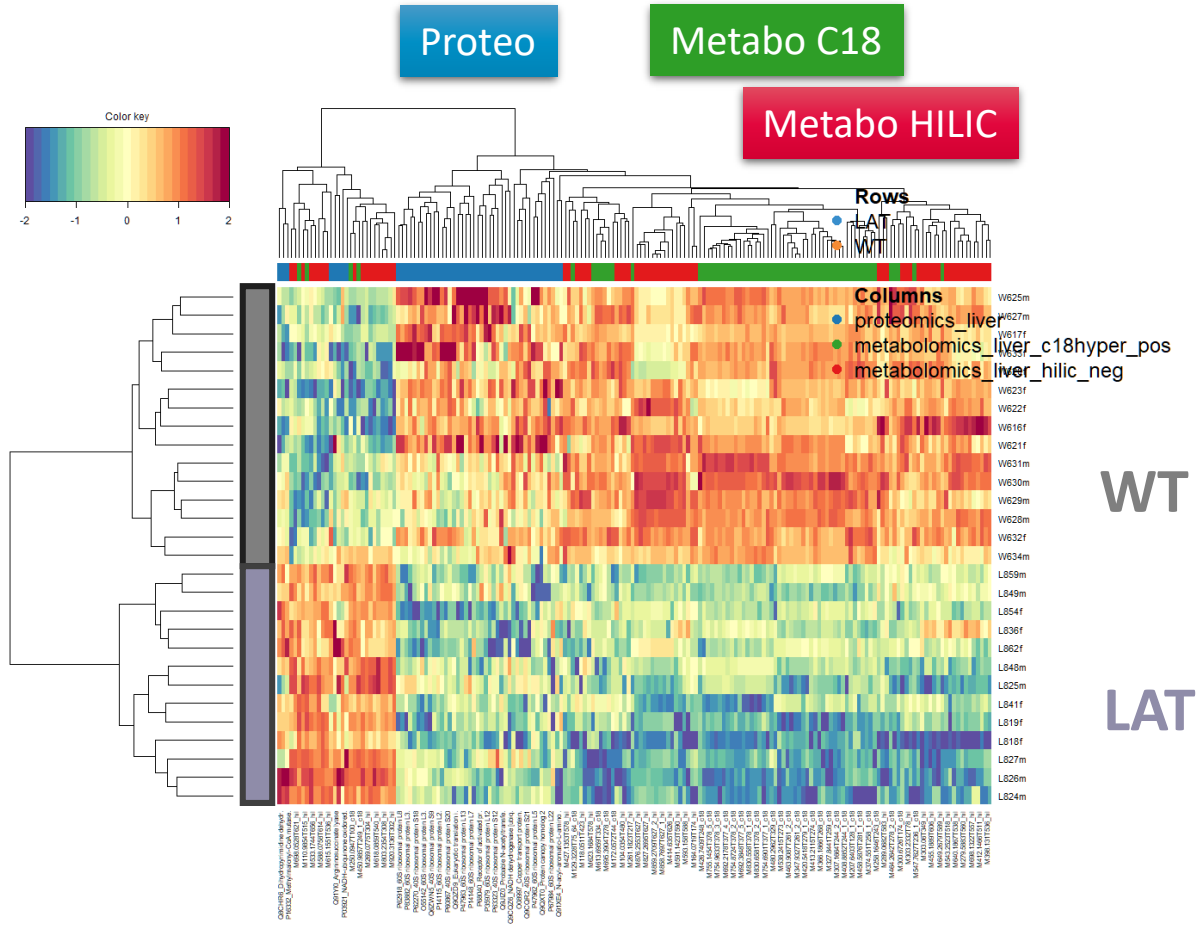
LAT vs WT



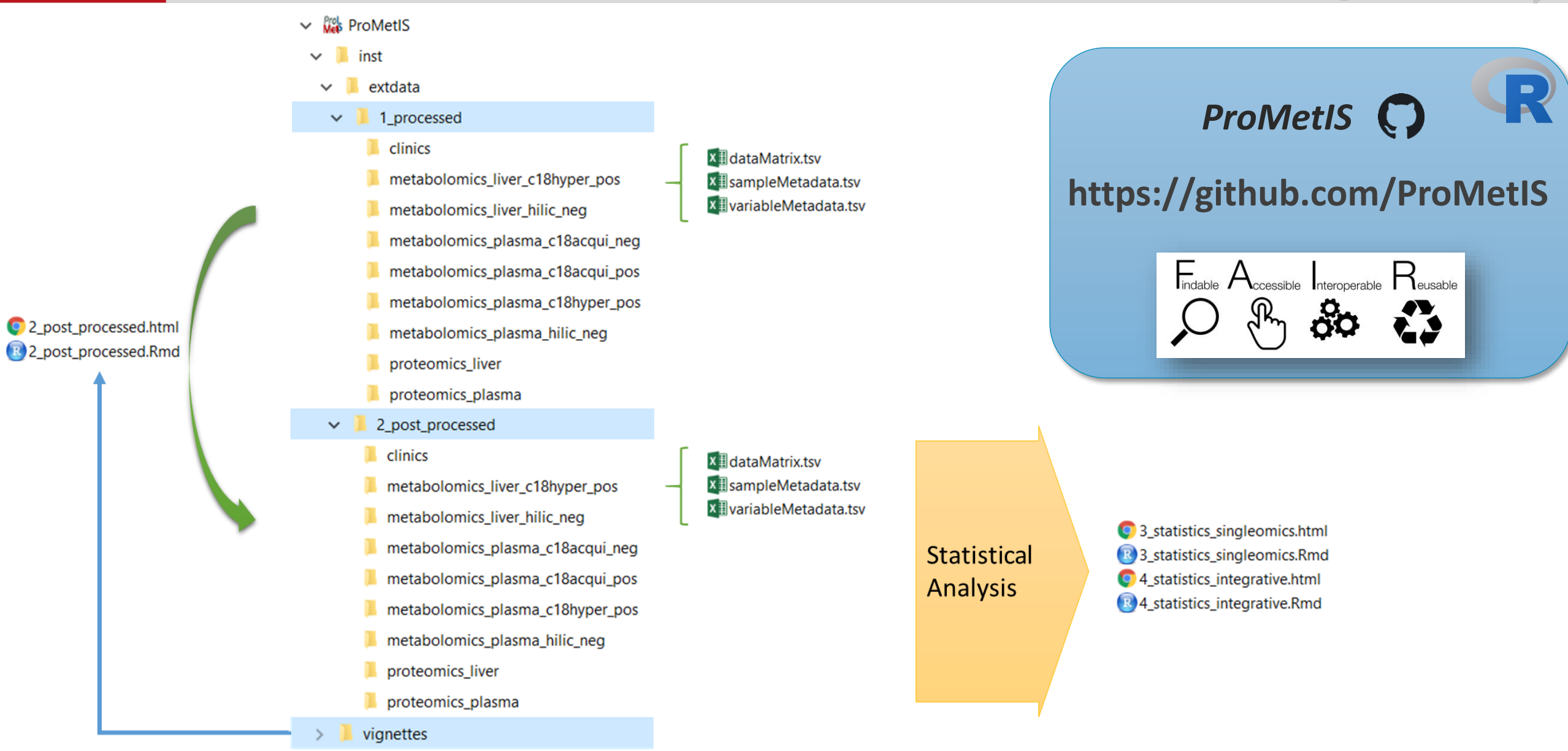


Tenenhaus *et al.* (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics*, **15**:569–583.

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



Datasets & code availability: the ProMetIS package



ProMetIS

https://github.com/ProMetIS

Findable Accessible Interoperable Reusable

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
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