

## Some examples of data integration

Alyssa Imbert

Biopuces, March 10, 2022

# Table of contents

## ① PhD: Integration of heterogeneous complex data from unbalanced datasets

- Multiple hot-deck imputation
- Association of clinical and transcriptomics variables



## ② Post-Doc: Metabolomics and proteomics data integration for deep phenotyping

- Presentation of ProMetIS project
- Intra-omics analysis
- Multi-omics analysis



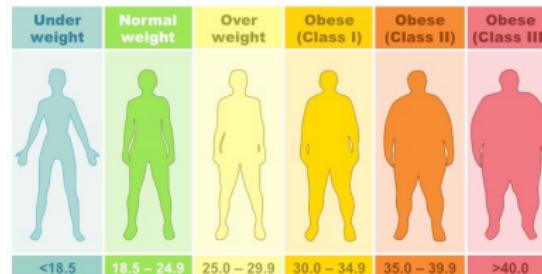
## Section 1

PhD: Integration of heterogeneous complex data from unbalanced datasets

# Obesity in few words

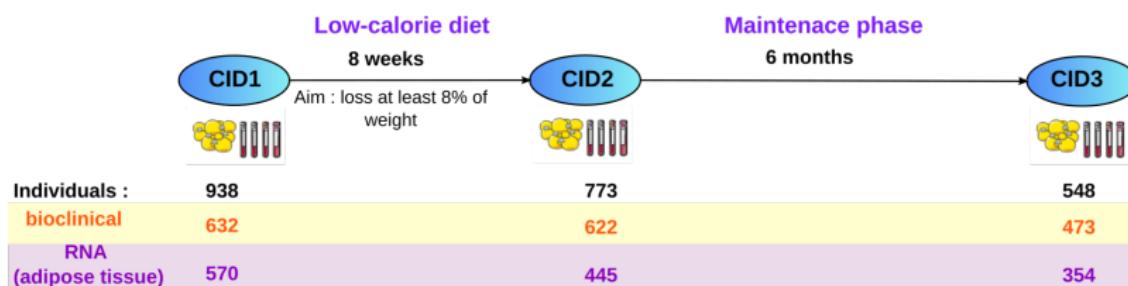
- **Obesity** : defined as abnormal or excessive fat accumulation that presents a risk to health
  - ↗ risk of cardiovascular diseases, type II diabetes, cancers, ...
- In 2016 (OMS) :
  - number of obesity cases x3 since 1975,
  - 39% of overweight adults, 13% obese
- BMI (Body Mass Index) : simpler way to assess obesity

$$\text{BMI} = \frac{\text{weight (kg)}}{\text{size}^2(\text{m}^2)}$$



(Source figure: [https://ib.bioninja.com.au/\\_Media/bmi-categories\\_med.jpeg](https://ib.bioninja.com.au/_Media/bmi-categories_med.jpeg))

# DiOGenes

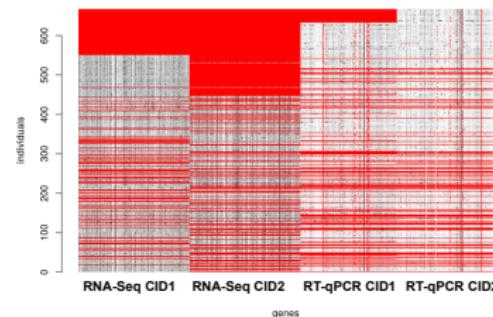
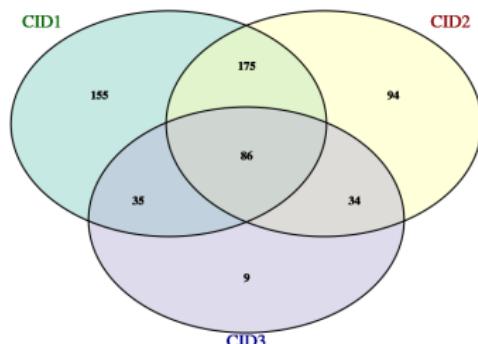


Each time step (CID: Clinical Investigation Day):

- Clinical data
- Transcriptomic data:
  - RT-qPCR
  - next-generation sequencing (NGS): RNA-Seq and QuantSeq

# Presentation of datasets

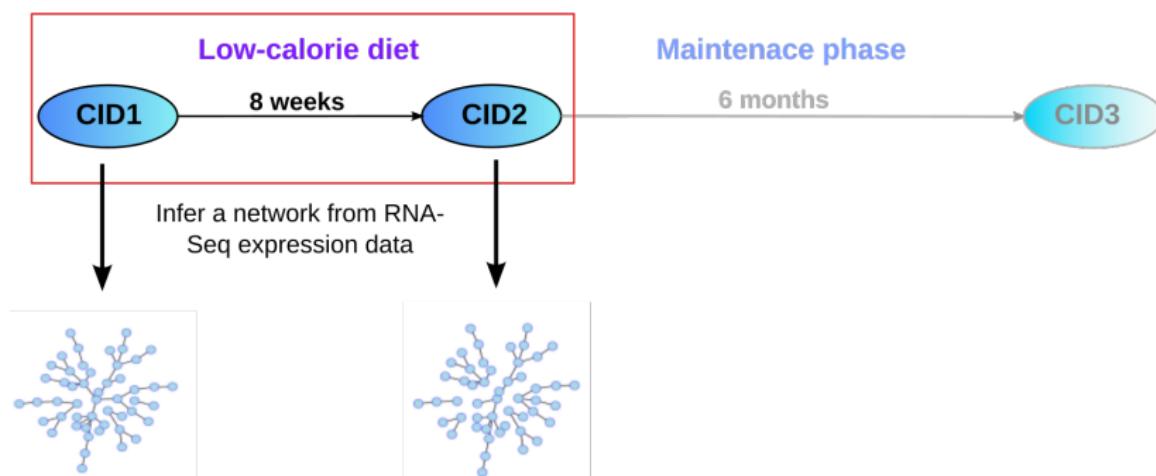
|            | clin. | RT-qPCR | RNA-Seq | QuantSeq |
|------------|-------|---------|---------|----------|
| Nb var.    | > 80  | 284     | 54 043  | 32 041   |
| Nb samples |       |         |         |          |
| CID1       | 632   | 495     | 451     | 416      |
| CID2       | 622   | 544     | 389     | 291      |
| CID3       | 473   | 371     | 164     | 211      |



Nb. individuals, RNA-Seq

# Visualization of the problem with DiOGenes

Aim: Study the impact of a low-calorie diet on gene regulation



- Choice of model for network inference?

# Network inference and RNA-seq data

- **RNA-seq data:**

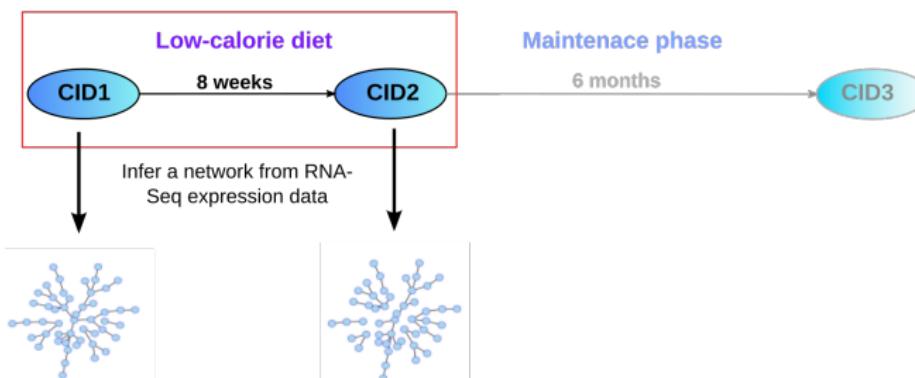
- counts → discrete data;
- over-dispersed data (variance > mean).

- **Network inference method:**

- Transform data → approach gaussian distribution  
→ Gaussian Graphical Model (GGM)
- Use appropriate models based on Poisson distribution
  - **Log-linear Poisson graphical model (llgm)**  
*[Allen and Liu, 2012]*; [Method](#)
  - hierarchical log-normal Poisson graphical model  
*[Gallopin et al., 2013]*.
  - poisson log-normal model:  
*[Choi et al., 2017, Chiquet et al., 2019]*

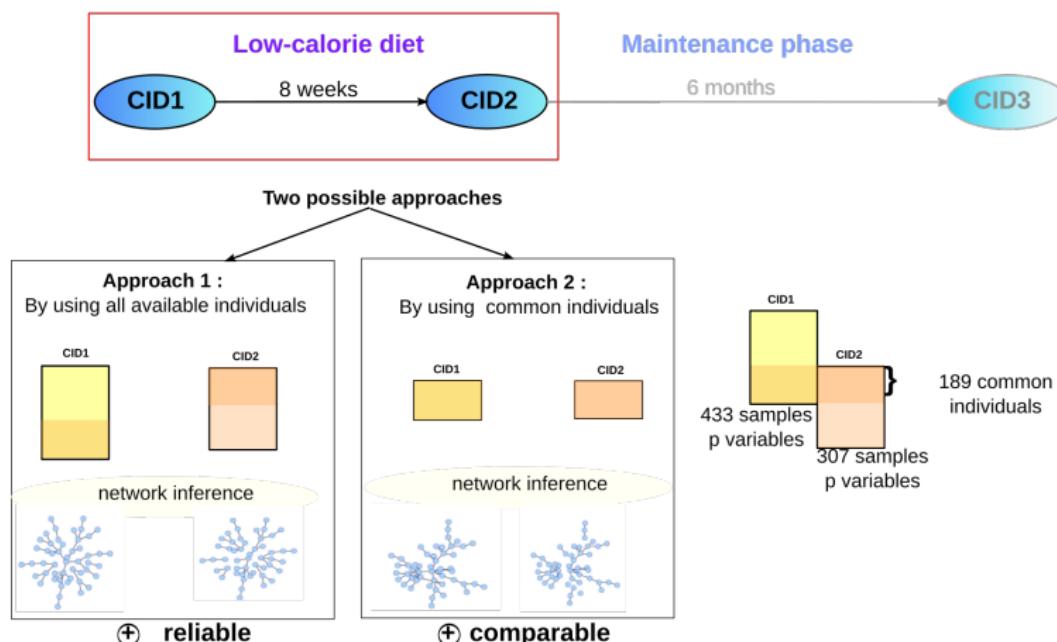
# Visualization of the problem with DiOGenes

Aim: Study the impact of a low-calorie diet on gene regulation



- Choice of model for network inference: **log-linear Poisson graphical model** (llgm)
- Which individuals are used to infer the network?

# Choice of individuals



**Proposal :** increase the quality of network inference by imputing missing individuals

# Problem

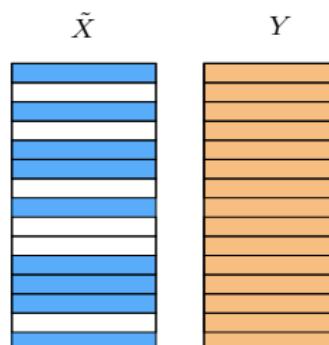
Search an imputation method which allows to:

- preserve the link between variables (genes)  
→ impute missing individuals **entirely** = impute simultaneously all variables
- Take into account uncertainty which are linked to imputation

**Aim:** improve the quality of inference by using external information (important  $n$  very small)

## Framework and notation

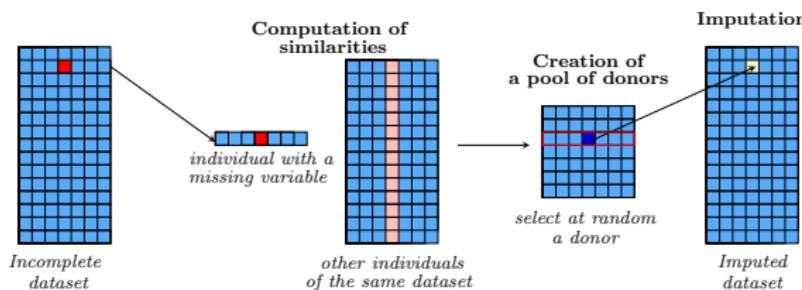
- Matrix  $\tilde{X}$  of size  $n_1 \times p$  → expression measures of interest (RNA-seq);
- matrix  $Y$  of size  $n \times q$  → metabolome, phenotypic data, qPCR expression, ...;
- $n_1$  samples (individuals) in common between  $\tilde{X}$  and  $Y$ ;
- presence of missing data → experimental reasons
- missing data supposed to be MAR (Missing At Random).



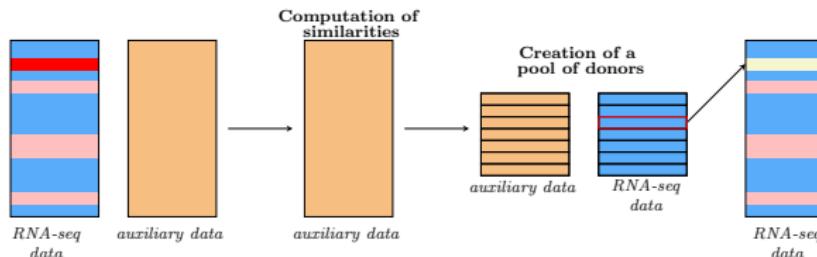
# Hot-deck imputation

A set of methods based on the concept of donors [Andridge and Little, 2010]

## Definition

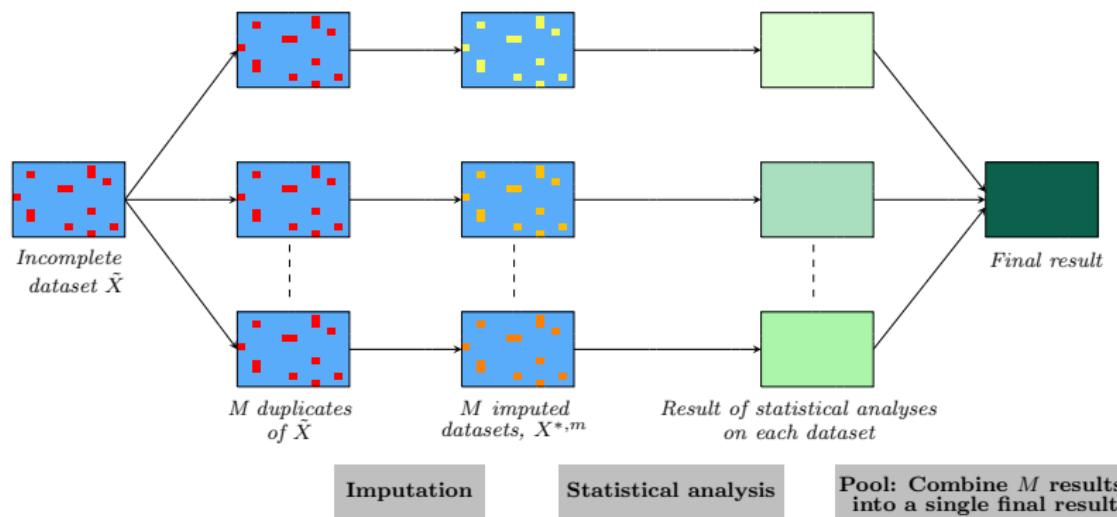


In our case:



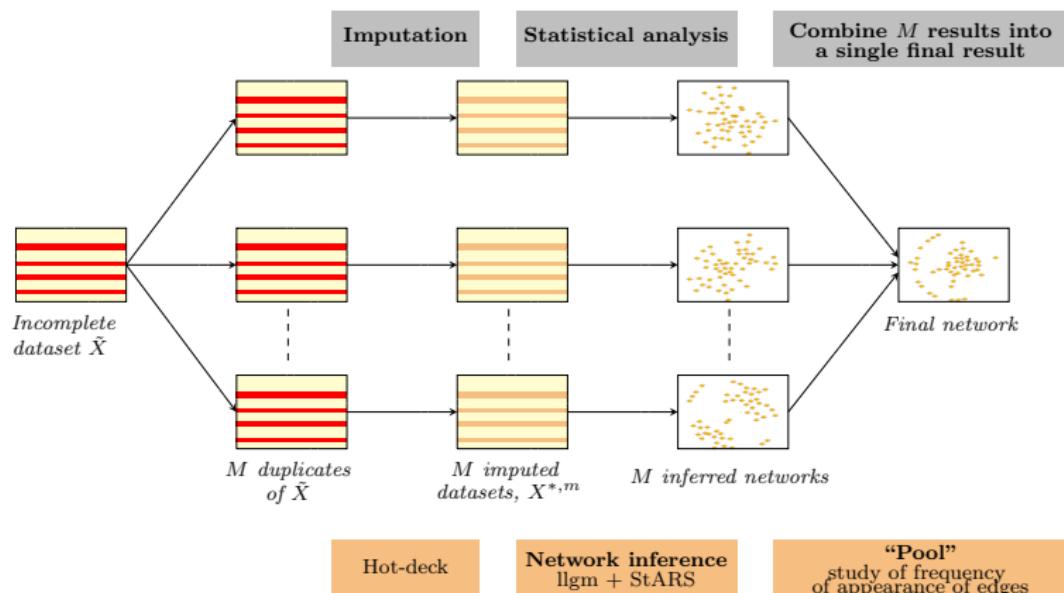
# Multiple imputation

A way to take into account uncertainty which are linked to imputation



[Rubin D., 1976, Rubin D., 2012]

# Multiple hot-deck imputation



# Multiple hot-deck imputation (hd-MI)

Similarity

Test different approaches:

- with **an affinity score** [Cranmer and Gill, 2012]:  
R package `hot.deck`

$$s(i,j) = \frac{1}{q} \sum_{k=1}^q \mathbb{I}_{\{|y_{ik} - y_{jk}| < \sigma\}}$$

where  $\sigma$  = fixed threshold and

$$\mathcal{D}(i) = \{j : s(i,j) = \max_{l \neq i} s(i,l)\}$$
 choice of sigma

- other approaches:
  - scaled affinity score (unit variance)
  - $k$  **nearest neighbors** ( $k$ -NN), Euclidean metric
  - $k$ -NN, Mahalanobis metric
  - $k$ -NN, CCA approach: most similar neighbor (MSN)  
[Crookston and Finley, 2008]  
→ sparse CCA +  $k$ -NN

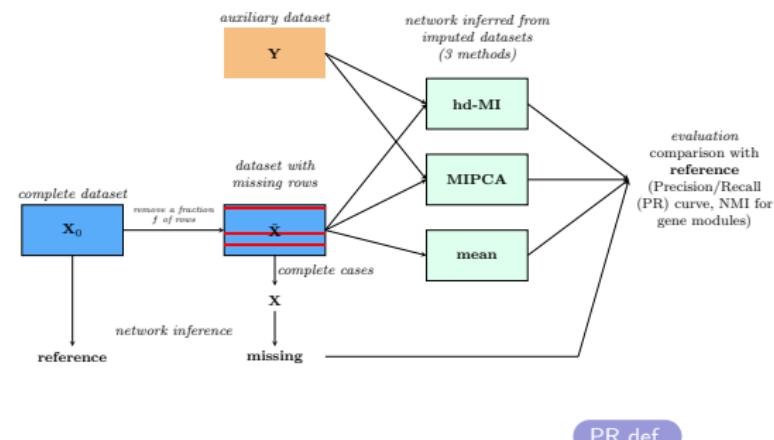
# Evaluation process, framework

- Test on real dataset,  
2 projects:

- GTEx
- DiOGenes

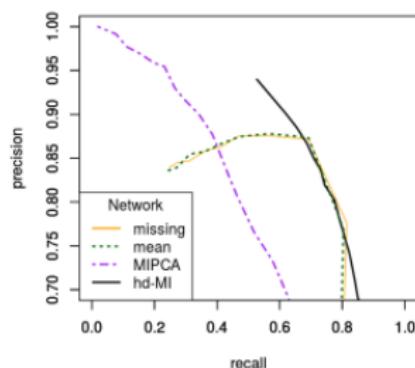
- 3 imputation methods:
  - mean
  - MIPCA<sup>1</sup>
  - our method:  
hd-MI

- 10%, 20%, 30%, 40% missing individuals
- $M = 100$

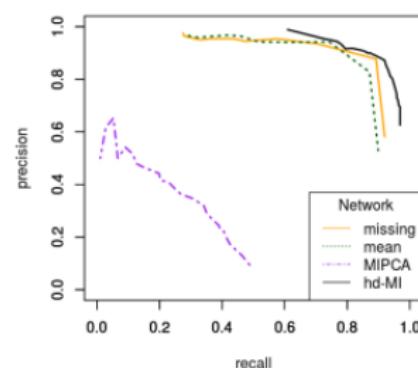


<sup>1</sup> MIPCA: Multiple Imputation PCA [Josse et al., 2011]

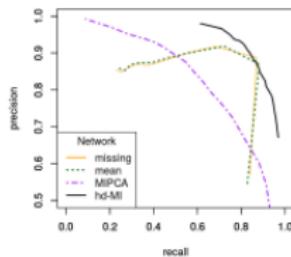
# Some precision/recall curve



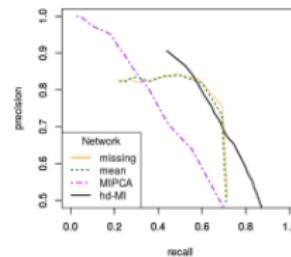
(a) DiOGenes - 20%



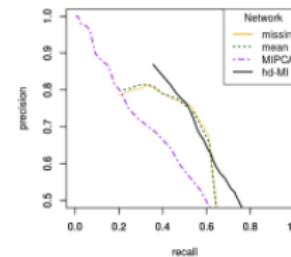
(b) GTEx - 20%



(c) DiOGenes - 10%



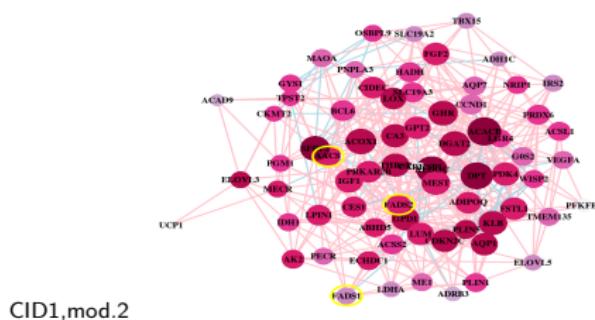
(d) DiOGenes - 30%



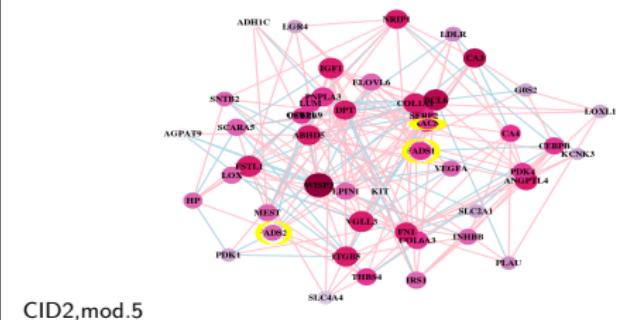
(e) DiOGenes - 40%

## Application on DiOGenes dataset

Persistence of the links between *FADS1*, *FADS2* et *AACS* (found linked here and in previous networks)

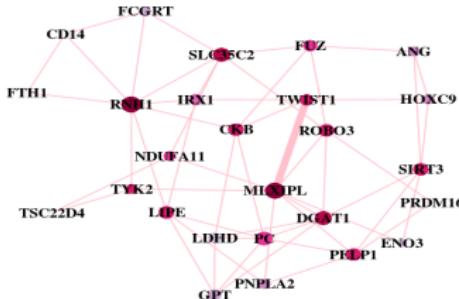


CID1,mod.2

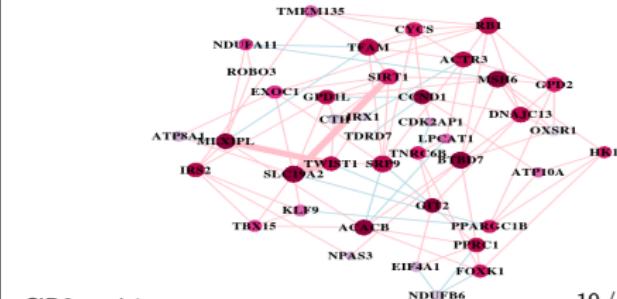


CID2,mod.5

**New links:** enlightened adipose tissue *SLC19A2* as novel partner in glucose homeostasis, besides *TWIST1* and *MLX1PL*



CID1,mod.1



CID2,mod.1

# Conclusion

- Importance of the choice of the matrix Y (auxiliary dataset)
- For high precision(i.e. less FP) , best recall (i.e. less FN) with our method hd-MI
- beyond 30% of missing individuals: results deteriorate  
*rightarrow* curve PR for hd-MI below missing PR curve
- R package: **RNAseqNet** (CRAN)

Imbert A. et al. (2018), *Multiple hot-deck imputation for network inference from RNA sequencing data*. *Bioinformatics* 34(10):1726–1732.  
(<https://doi.org/10.1093/bioinformatics/btx819>)

- Review on missing data

Imbert A. et Vialaneix N. (2018), *Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques : une revue des approches existantes.*, Journal de la Société Française de Statistique.

## To go further

- Network inferred by using only gene expression
- other types of available data
- to get an overview of the whole system: use different type of data (e.g. transcriptomics, clinical)
- **Problem:** multiple sources, heterogeneous, large size
- need to use **integrative methods**

# Biological question

**Question:** What changes in gene expression are associated with a change in one of the clinical variables of interest?

## Datasets

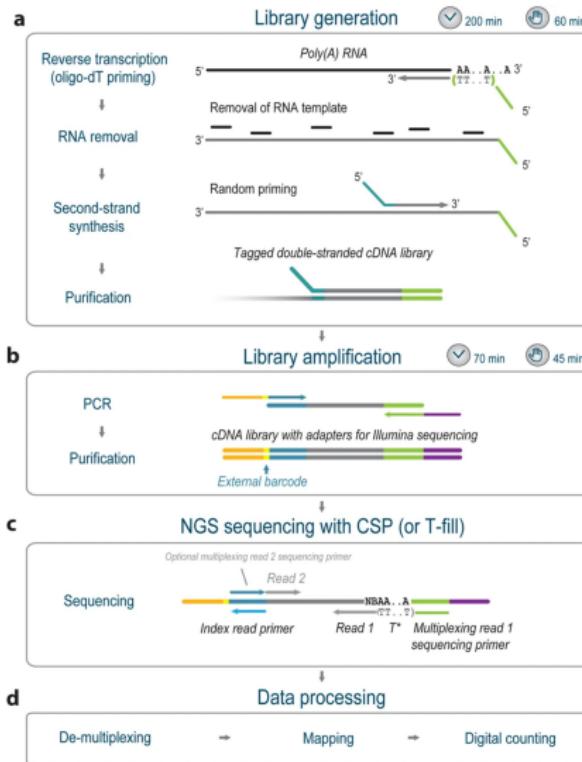
- Gene expression → QuantSeq
- a dozen selected clinical variables

## Aim:

- Analyze QuantSeq data
- infer a network with genes and clinical variables

# QuantSeq

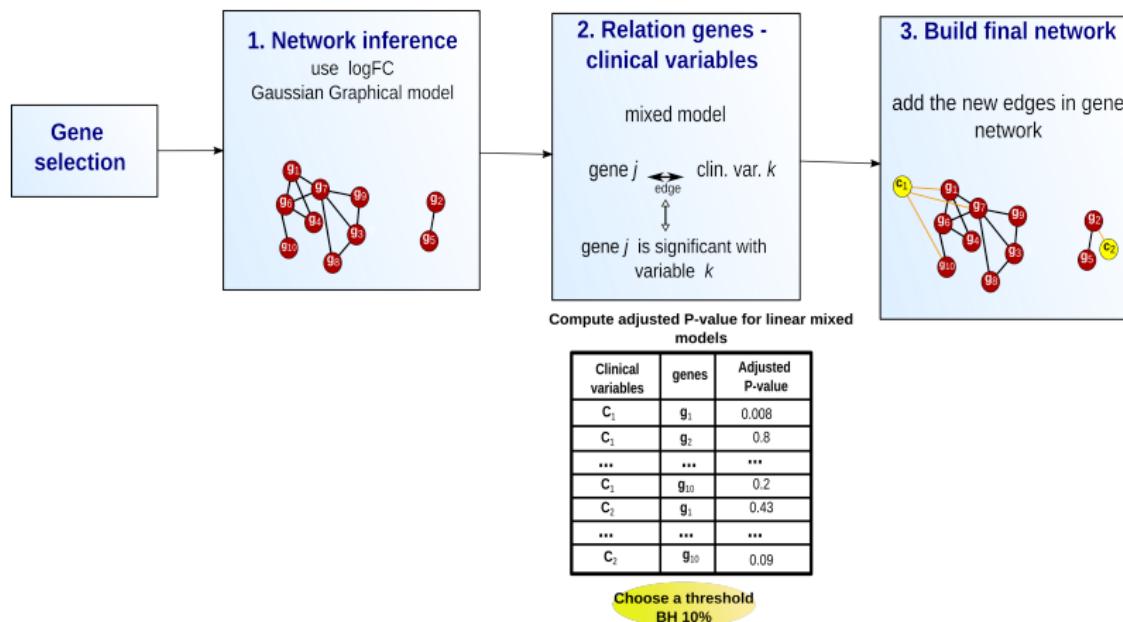
[Moll P. et al, 2014]



- ⊕: more tolerant of poor RNA quality, faster, less expensive
- ⊖: no search for isoforms

# An approach based on network inference

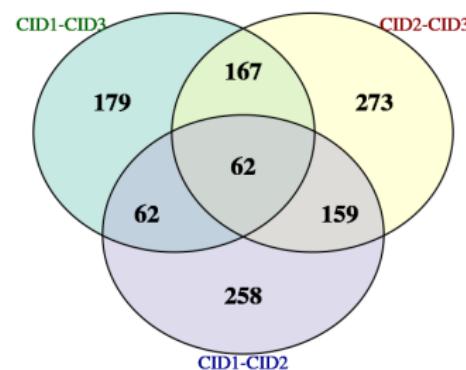
For each contrast



# 1. Gene selection

## 3 "thresholds":

- deletion of poorly expressed genes (genes with too many null counts, or missing logFC): arbitrary threshold: 25%
- Differentially expressed genes: adjusted pvalue (BH) < 5%
- sufficiently regulated expression:  $|FC| > 1.3$

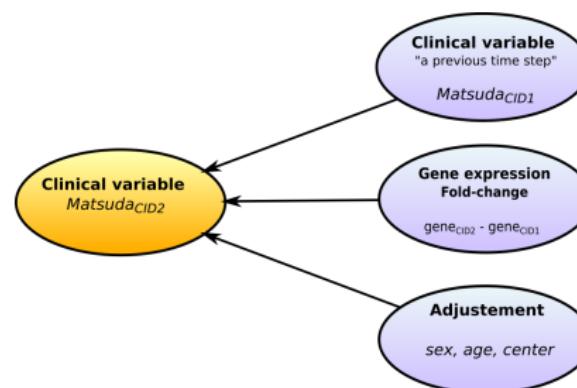


|           | Nb obs. | Nb genes |
|-----------|---------|----------|
| CID1/CID2 | 183     | 541      |
| CID2/CID3 | 122     | 661      |
| CID1/CID3 | 139     | 470      |

## 2. How estimate links between genes and clinical variables?

### Use mixed linear models

Example of model (contrast CID1/CID2)

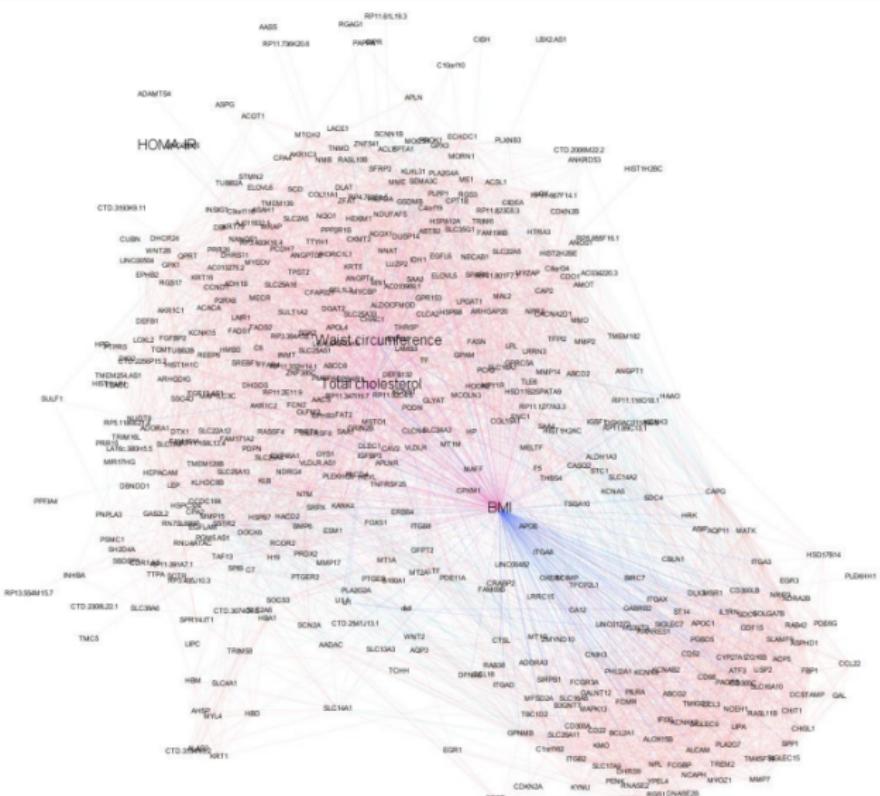


$$\text{Matsuda}_{CID2} = \text{Matsuda}_{CID1} + \log\text{FC}_{DEG} + \text{sex} + \text{age} + \text{center}$$

One model per selected genes + correction for multiple test

R package: nlme

### 3. Final network



## Some biological results

- Found 5 modules (loss-calorie diet phase), 3 included at least one bio-clinical variable
- Change in BMI connected with changes in mRNA level of genes with inflammatory response signature  
→ change in BMI negatively associated to changes in expression of genes encoding secreted protein (*GDF15*, *CCL3* and *SPP1*)
- network analyses identified a novel AT feature with *GDF15* upregulated with calorie restriction induced weight loss, concomitantly to macrophage markers

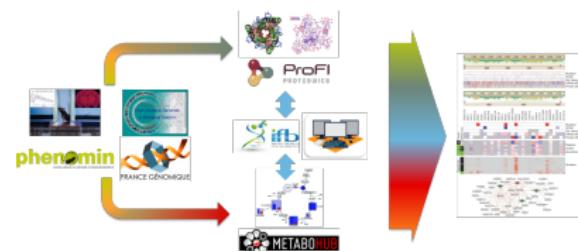
Imbert A. et al. (2022), Network analyses reveal negative link between changes in adipose tissue *GDF15* and BMI during dietary induced weight loss. *Journal of Clinical Endocrinology & Metabolism* (<https://doi.org/10.1210/clinem/dgab621>)

## Section 2

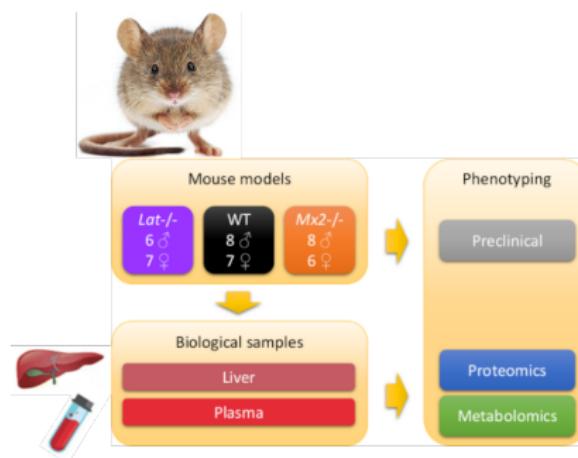
Post-Doc: Metabolomics and proteomics data  
integration for deep phenotyping

# ProMetIS project

- **Objective:** high-throughput integration of proteomics and metabolomics data
- **Case study:** molecular phenotyping of mouse models from the IMPC consortium
  - 2 K-O (LAT and MX2) and one control group (WT)
- **Partner infrastructures**
  - France Génomique
  - PHENOMIN (Institut Clinique de la souris)
  - ProFI proteomics
  - Metabohub
  - Institut Français de Bioinformatique



# Biological question: characterization of knock-out mice

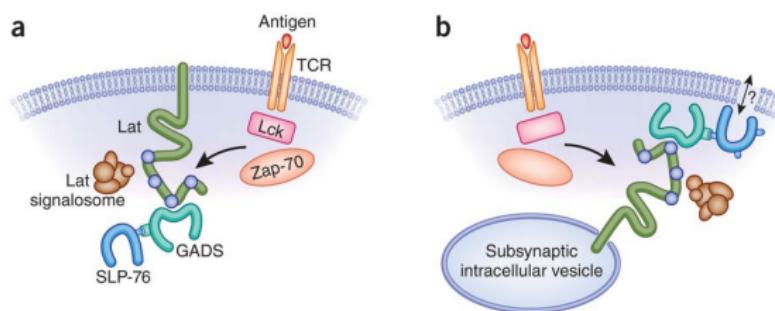


Imbert A. et al. (2021), ProMetIS: deep phenotyping of mouse models by combined proteomics and metabolomics analysis. *Scientific Data*

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

# LAT

- ▷ **LAT : linker for activation of T cells** involved in
  - T-cell receptor (TCR) signaling [*Loviglio et al., 2017*]
  - Neurodevelopmental diseases [*Roncagalli et al, 2010*]

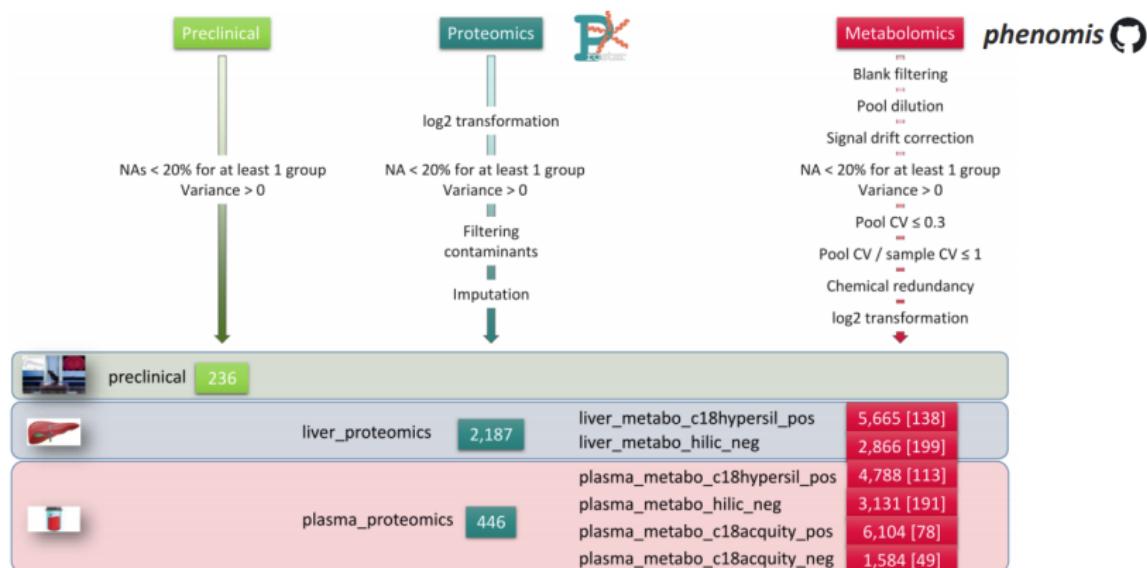


[*Malissen and Marguet, 2011*]

# Material and methods

|                               | Metabolomics   | Proteomics  |
|-------------------------------|--|---|
|                               |  METABO HUB |  ProFI<br>PROTEOMICS |
| Liquid Chromatography         | C18 and Zic-pHILIC   | Trapping + C18 separation   |
| Mass Spectrometry             | Exactive (Thermo)/Q-TOF Impact HD2 (Bruker)  | Q-Exactive Plus (Thermo)/ DDA Top 10 acquisition  |
| Data Processing               | XCMS (Workflow4Metabolomics)   | Mascot database searching Proline   |
| Annotation/<br>Identification | KEGG, HMDB, METLIN, In-house   | SwissProt   |

# Datasets: preclinical, proteomics and metabolomics

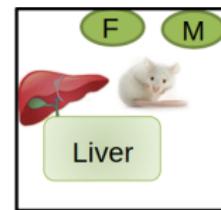


# Analysis plan

n = 28



Genotype  
LAT vs WT



## Intra-omics analysis

- Exploratory analysis (PCA)
- Differential analysis (linear model with limma)
- Multivariate modeling (PLS-DA)
- Feature selection (biosigner)

## Data integration

- Mapping and pathway analysis
- Multi-block approach

# Format: 3 tables

ExpressionSet, MultiDataSet

## ① dataMatrix.tsv:

- names of your samples in the first row
- name of your variables in the first column

## ② sampleMetadata.tsv:

- names of the factors about samples
- names of your samples which must exactly match those of dataMatrix

## ③ variableMetadata.tsv:

- names of the metadata (mz/rt, etc.)
- names of variables, which must exactly match those of dataMatrix

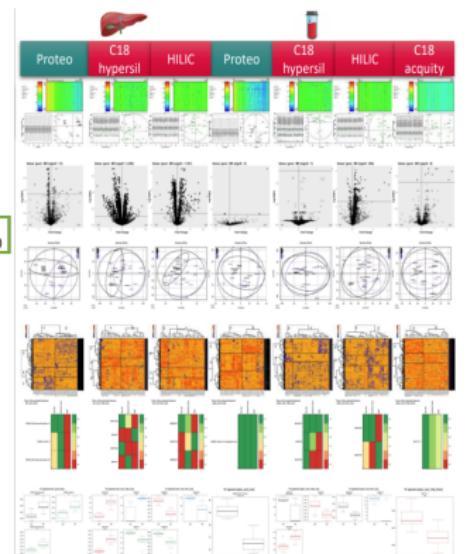
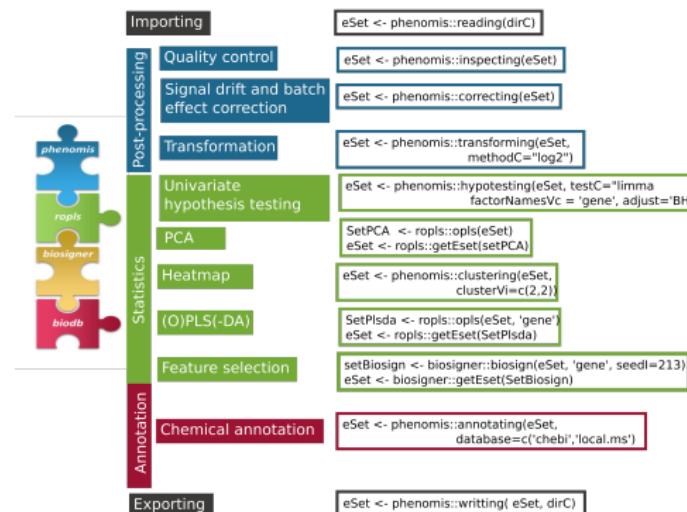
|   | A          | B           | C           | D           | E           | F | G | H |
|---|------------|-------------|-------------|-------------|-------------|---|---|---|
| 1 | dataMatrix | HU_neg_017  | HU_neg_028  | HU_neg_034  | HU_neg_051  |   |   |   |
| 2 | M97T61     | 17153667.17 | 10216240.88 | 16029523.86 | 14468044.45 |   |   |   |
| 3 | M99T61     | 795428.1989 | 400570.6324 | 831219.0107 | 671471.606  |   |   |   |
| 4 | M135T54    | 7057880.716 | 11926973.53 | 9514452.963 | 6990900.537 |   |   |   |
| 5 | AA136T54   | 331030.3107 | 332801.8693 | 313006.3006 | 332423.1003 |   |   |   |

|   | A              | B          | C              | D    | E     | F   | G   | H       |
|---|----------------|------------|----------------|------|-------|-----|-----|---------|
| 1 | sampleMetadata | sampleType | injectionOrder | mode | batch | age | bmi | gender  |
| 2 | HU_neg_017     | sample     |                | 17   | neg   | ne1 | 41  | 23.03 M |
| 3 | HU_neg_028     | sample     |                | 23   | neg   | ne1 | 41  | 23.92 F |
| 4 | HU_neg_034     | sample     |                | 26   | neg   | ne1 | 52  | 23.37 M |
| 5 | HU_neg_051     | sample     |                | 45   | neg   | ne1 | 24  | 23.22 F |

|   | A                | B     | C           | D            | E      | F           | G           | H       |
|---|------------------|-------|-------------|--------------|--------|-------------|-------------|---------|
| 1 | variableMetadata | mz/rt | fold        | tstat        | pvalue | mzmed       | mzmin       | mzmax   |
| 2 | M97T61           | 47    | 69.27624774 | -19.66155855 | 0      | 96.95889309 | 96.9544608  | 96.9605 |
| 3 | M99T61           | 52    | 390.1176385 | -18.1537251  | 0      | 98.955561   | 98.9554026  | 98.9556 |
| 4 | M135T54          | 179   | 394.008022  | -18.58129475 | 0      | 135.0296344 | 135.0295548 | 135.02  |
| 5 | M136T54          | 183   | inf         | -17.61021775 | 0      | 136.029175  | 136.0284993 | 136.03  |
| 6 | M187T53          | 487   | 1345.318461 | -19.79392715 | 0      | 187.0373874 | 187.0373051 | 187.03  |
| 7 | AA136T54         | 656   | inf         | -18.3639838  | 0      | 182.0344606 | 180.0343633 | 180.03  |

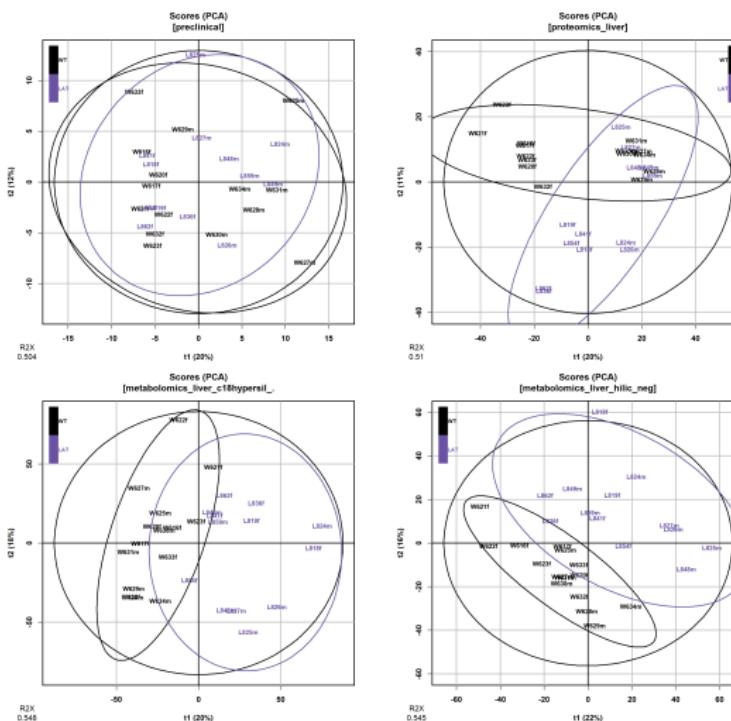
# Workflow

## Presentation of the R package phenomis



<https://github.com/SciDoPhenIA/phenomis>

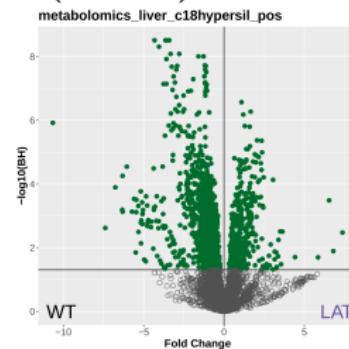
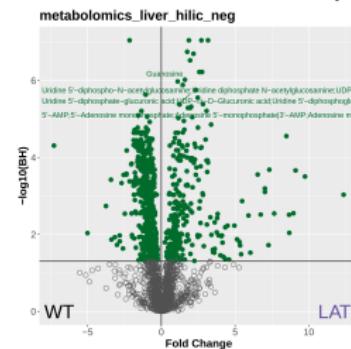
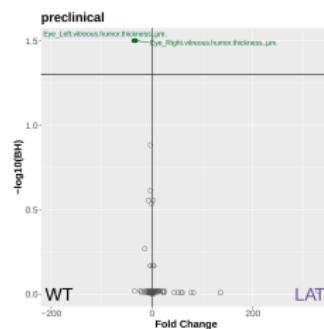
# PCA, liver, colored by gene



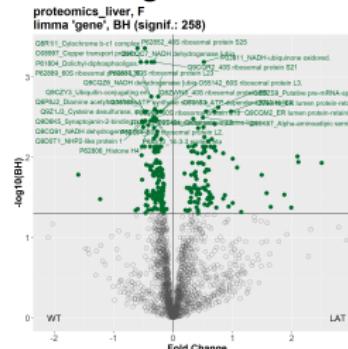
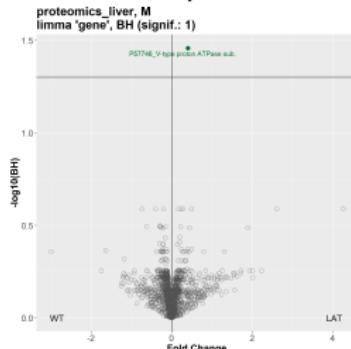
PCA colored by sex

# Differential analysis, liver

Model:  $\sim \text{gene} + \text{sex} + \text{gene:sex} \rightarrow \text{correction multiple test (FDR 5\%)}$



Proteomics: separate sex and model:  $\sim \text{gene} \rightarrow \text{correction multiple test (FDR 5\%)}$

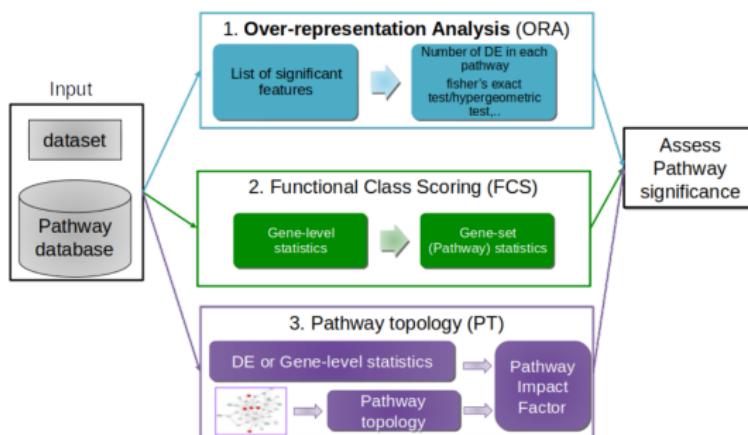


## Number of significant features, liver

| Dataset               | Number of significant features   | Number of features |
|-----------------------|----------------------------------|--------------------|
| Proteomics            | Significant interaction gene:sex | 2098               |
|                       | Female: 258 and Male: 1          |                    |
| Metabo c18+           | 1608                             | 5665               |
| Metabo hilic -        | 826                              | 2866               |
| Annotated metabolites |                                  |                    |
| Met c18 +             | 41                               | 138                |
| Met hil-              | 61                               | 199                |

# Pathway analysis and mapping

- Enrichment analysis (using proteomic data)

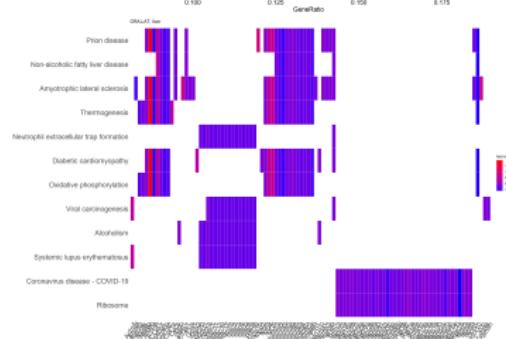
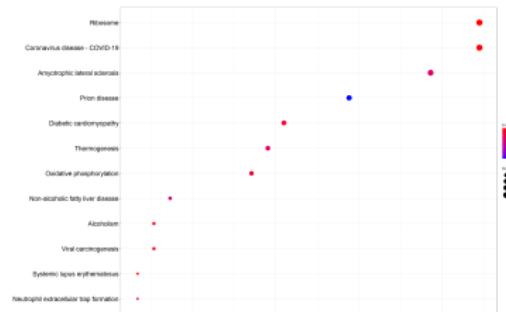


[Khatri et al., 2012]

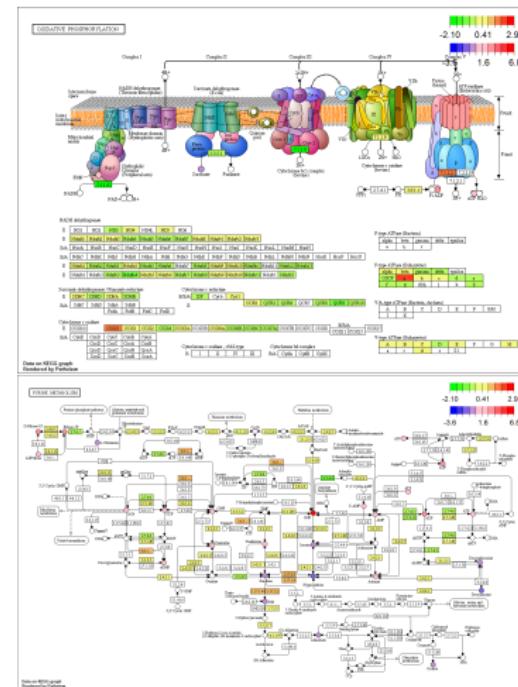
- Use databases that include both proteins (genes) and metabolites: KEGG
- Mapping proteins and metabolites → enriched pathways

# Enrichment analysis

## ORA analysis, Proteomics, KEGG:

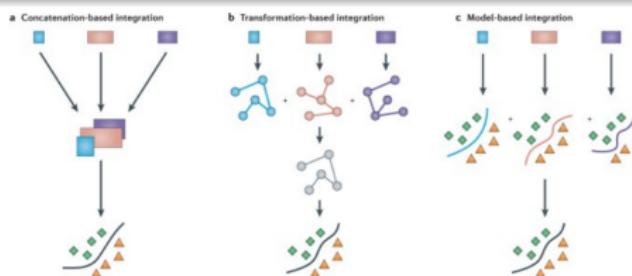


R package: clusterProfiler



R package: pathview

# Data integration



[Ritchie et al, 2015]

[Picard et al, 2021]:

- **Early integration:** concatenation-based
- **Mixed integration:** transformation-based (Kernel learning, graph)
- **Intermediate integration:** jointly integrating the multi-omics datasets without needing prior transformation and without relying on a simple concatenation (rGCCA, joint NMF, iCluster, MOFA, ...)
- **Late integration:** model-based
- **Hierarchical integration:** inclusion of the prior knowledge of regulatory relationships between the different layers

▷ <https://github.com/mikelove/awesome-multi-omics>

# Multi-block analysis



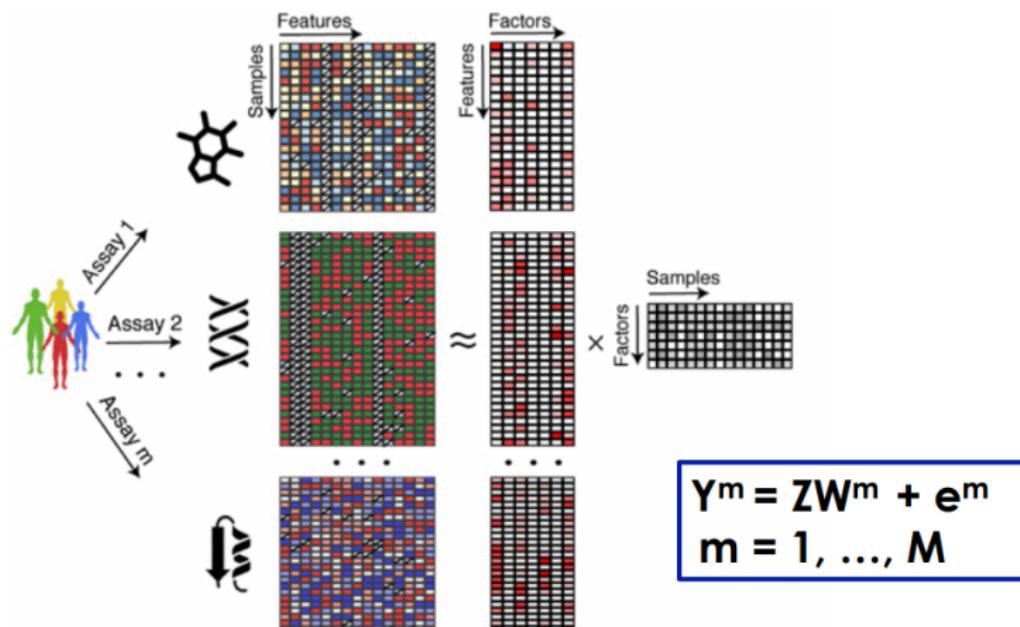
|  |   |  |
|--|---|--|
| <ul style="list-style-type: none"><li>• Univariate/bivariate</li></ul>                                 | Correlation, statistic test (t test, ANOVA, etc.) |  |
| <ul style="list-style-type: none"><li>• Unsupervised multivariate analysis</li></ul>                   | PCA   |  |
| <ul style="list-style-type: none"><li>• Supervised multivariate analysis</li></ul>                     | PLS, PLS-DA                                       |  |
| <ul style="list-style-type: none"><li>• Integration with 2 datasets (quantitative variables)</li></ul> | PLS, CCA, rCCA, sPLS                              |  |
| <ul style="list-style-type: none"><li>• Multi-block approach</li></ul>                                 | rGCCA, sGCCA<br>MOFA, MCIA                        |  |

Source: <http://mixomics.org/>, presentation

# Unsupervised approach: MOFA

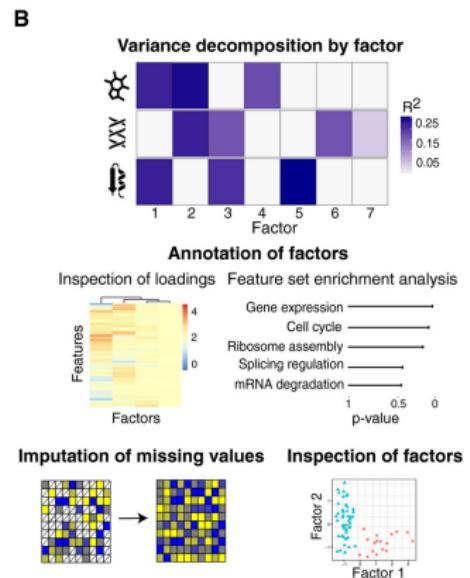
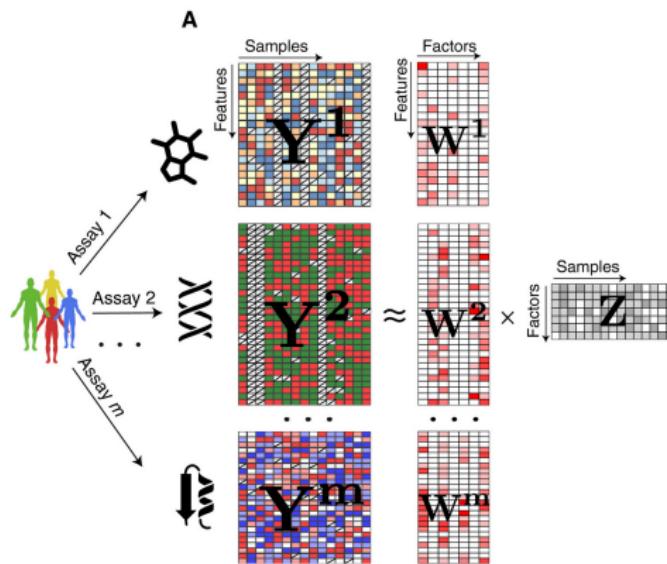
[Argelaguet R et al, 2018, Argelaguet R et al, 2020]

## MOFA model

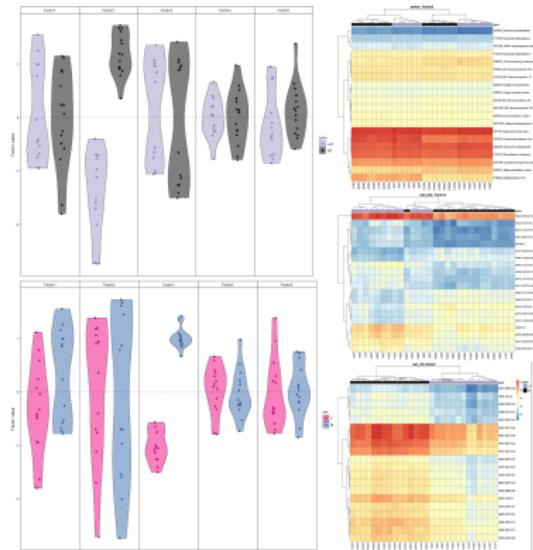
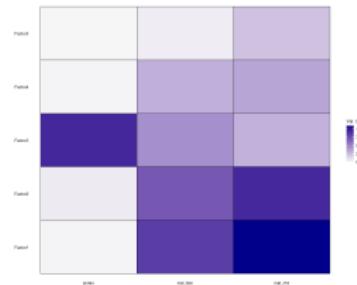
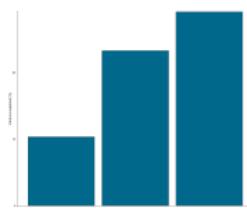


# Unsupervised approach: MOFA

[Argelaguet R et al, 2018, Argelaguet R et al, 2020]



# MOFA, results



## warning:

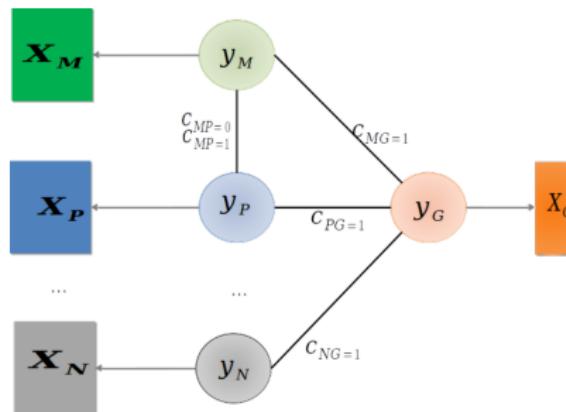
- size of the blocks → impact Illustration
- no orthogonality constraints: check that the Factors are largely uncorrelated

R package: MOFA2

# Supervised approach

Regularized Generalized Canonical Correlation Analysis (RGCCA),  
[Tenenhaus & Tenenhaus, 2011]

Define links between blocks:



## Aim:

- block components explain well their own block
- Block components are as correlated as possible for connected blocks.

# RGCCA/sGGCA

[Tenenhaus & Tenenhaus, 2011, Tenenhaus et al, 2014]

## RGCCA: optimization problem

$$\begin{aligned} \max_{w_1, \dots, w_J} & \sum_{j,k}^J c_{jk} g(\text{cov}(X_j w_j, X_k w_k)) \\ \text{s.t. } & (1 - \tau_j) \text{var}(X_j w_j) + \tau_j \|w_j\|_2^2 = 1, j = 1, \dots, J \end{aligned}$$

- $c_{jk} = 1$  if  $X_j \leftrightarrow X_k$ , 0 otherwise
- $g =$  any convex function
- $0 \leq \tau \leq 1$  continuum between correlation and covariance

## sGGCA: add a L1-penalty, $\tau_j = 1$

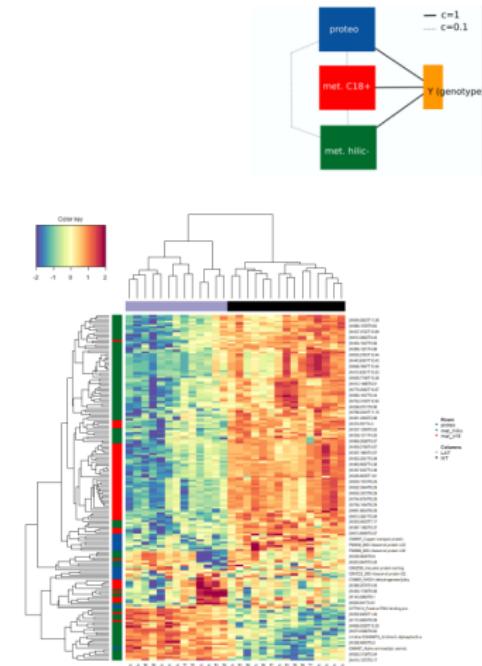
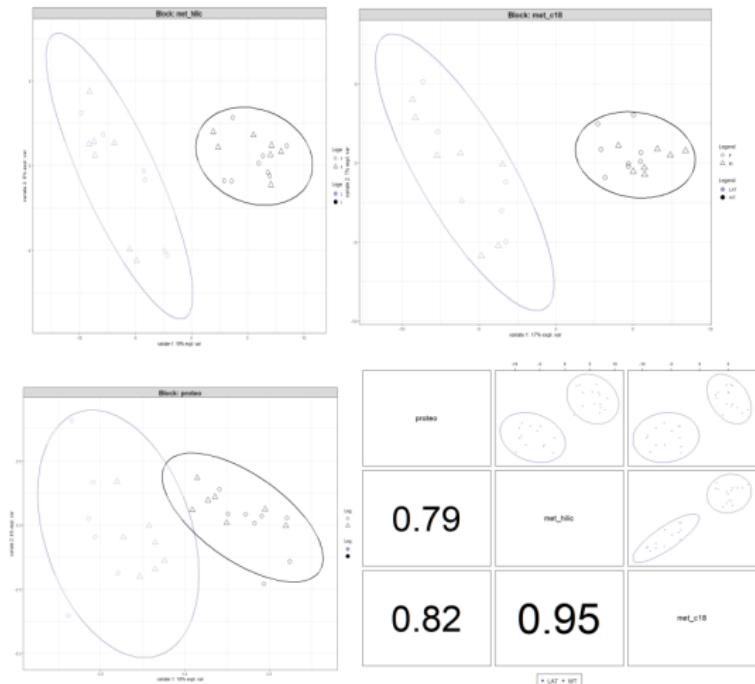
$$\begin{aligned} \max_{w_1, \dots, w_J} & \sum_{j,k}^J c_{jk} g(\text{cov}(X_j w_j, X_k w_k)) \\ \text{s.t. } & \|w_j\|_2 = 1 \text{ and } \|w_j\|_1 \leq s_j, j = 1, \dots, J \end{aligned}$$

where  $s_j$  is a user defined positive constant that determines the amount of sparsity for  $w_j$

R package: RGCCA and mixOmics (method DIABLO)

# sGCCA results

R package: mixOmics, DIABLO



See results for sgcca with only annotated features

# Thanks for your attention



Nathalie Vialaneix  
 Inserm  
Nathalie Viguerie



Armand Valsesia  
Jörg Hager et toute  
son équipe



Caroline Le Gall  
Pierre-Antoine Gourraud

## ProMetIS consortium



Etienne Thévenot  
Natacha Lenuzza  
Pierrick Roger  
François Fenaille  
Florence Castelli  
Christophe Junot  
Estelle Pujos-Guillot  
Marion Brandolini-Bunion  
Franck Giacomoni  
Fabien Jourdan  
  
Arthur Tenenhaus



Christine Carapito  
Magali Rompais  
Myriam Ferro  
Virginie Brun  
Christophe Bruley  
Jérôme Garin  
Anne Gonzalez-de-Peredo  
Emmanuelle Mouton  
Yves Vandenbrouck  
Thomas Burger



Tania Sorg  
Mohammed Selloum  
Laurent Vasseur  
Yann Herault



Christophe Lechaplain  
Pierre Le Ber  
Marcel Salanoubat  
Alain Perret



Olivier Sand  
Claudine Médigue  
David Vallenet



Ludovic Cottret

## Section 3

### References

# References I

-  Allen G. and Liu Z. (2012).  
A log-linear graphical model for inferring genetic networks from high-throughput sequencing data.  
*In Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*.
-  Andridge R. and Little R. (2010).  
A review of hot deck imputation for survey non-response.  
*International Statistical Review*, 78(1):40–64.
-  Argegaalet R. Velten B., Arnol A., Dietrich S., Zenz T., Marioni J., Buettner, F., Huber W. and Stegle O. (2018).  
Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets  
*Molecular Systems Biology*, 14:e8124
-  Argegaalet R., Arnol D., Bredikhin D., Deloro Y., Velten B., Marioni J. and Stegle O. (2020).  
MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data.  
*Genome Biology*
-  Bar-Hen A. and Poggi J. (2016).  
Influence measures and stability for graphical models.  
*Journal of Multivariate Analysis*, 147:145–154.
-  Chiquet, J. Mariadassous, M. and Robin S. (2019).  
Variational inference for sparse network reconstruction from count data.  
*36th International Conference on Machine Learning*

## References II

-  Choi Y., Coram M., Peng J., and Tang H.(2017).  
A poisson log-normal model for constructing gene covariation network using rna-seq data.  
*Journal of Computational Biology*, 24(7):721–731.
-  Cranmer S. and Gill J. (2012).  
We have to be discrete about this: a non-parametric imputation technique for missing categorical data.  
*British Journal of Political Science*, 43:425–449.
-  Crookston N. and Finley A. (2008).  
yalmpute: an R package for kNN imputation.  
*Journal of Statistical Software*, 23:10.
-  Gallopin M., Rau A., and Jaffrézic F. (2013).  
A hierarchical Poisson log-normal model for network inference from RNA sequencing data.  
*PLoS ONE*, 8(10).
-  Imbert A., Valsesia A., Le Gall C., Armenise C., Lefebvre G., Gourraud P., Viguerie N., and Vialaneix N. (2018).  
Multiple hot-deck imputation for network inference from RNA sequencing data.  
*Bioinformatics*, 34(10):1726–1732.
-  Josse, J., Pagès, J., and Husson, F. (2011).  
Multiple imputation in principal component analysis.  
*Advances in Data Analysis and Classification*, 5(3):231–246.

# References III

-  Larsen T., Dalskov S., van Baak M., Jebb S., Kafatos A., Pfeiffer A., Martinez J., Handjieva-Darlenska T., Kunešová M., Holst C., Saris W., and Astrup A. (2010).  
The diet, obesity and genes (diogenes) dietary study in eight European countries - A comprehensive design for long-term intervention.  
*Obesity Reviews*, 11(1):76–91.
-  Liu H., Roeber K., and Wasserman L. (2010).  
Stability approach to regularization selection (StARS) for high dimensional graphical models.  
In *Proceedings of Neural Information Processing Systems (NIPS 2010)*, volume 23, pages 1432–1440, Vancouver, Canada.
-  Lonsdale J., Thomas J., Salvatore M. et al. (2013).  
The genotype-tissue expression (GTEx) project.  
*Nature Genetics*, 45:580–585.
-  Loviglio M.N., Arbogast T. et al. (2017)  
TThe Immune Signaling Adaptor LAT Contributes to the Neuroanatomical Phenotype of 16p11.2 BP2-BP3 CNVs  
*The American Journal of Human Genetics*, 101:564–577.
-  Khatri P., Sirota M. , Butte A.J. (2012).  
Ten years of pathway analysis: current approaches and outstanding challenges.  
*PLoS Computational Biology*, 8: e1002375.
-  Malissen B., Marguet D. (2011)  
La(s)t but not least.  
*Nature Immunology*, 12(7):592-593.

# References IV

-  Moll P., Ante M., Seitz A., and Reda T. (2014).  
QuantSeq 3' mRNA sequencing for RNA quantification.  
*Nature Methods*, 11(November) :25.
-  Picard M., Scott-Boyer M-P., Bodein A., Périn O., Droit A. (2021).  
Integration strategies of multi-omics data for machine learning analysis  
*Computational and Structural Biotechnology Journal*, 19, 3735-3747
-  Ritchie M.D., Holzinger E.R., Li R., Pendergrass S.A. and Kom D. (2015).  
Methods of integrating data to uncover genotype–phenotype interactions.  
*Nature Reviews Genetics*, 16:85–97.
-  Roncagalli R, Mingueneau M., Grégoire C., Malissen M.  
LAT signaling pathology: An "autoimmune" condition without T cell self-reactivity  
*Trends in immunology*, 31(7):253-259
-  Rubin D. (19676).  
Inference and missing data.  
*Biometrika*, 63(3) :581-592.
-  Rubin D. (2012).  
Multiple imputation after 18+ years.  
*Journal of the American Statistical Association*, 91(434) :473–489.

# References V

-  Tenenhaus A. & Tenenhaus M. (2011).  
Regularized Generalized Canonical Correlation Analysis.  
*Psychometrika*, 76(2) :257-284.
-  Tenenhaus A., Phillippe C., Guillemot V., Le Cao K-A., Grill J. and Frouin V. (2014).  
Variable selection for generalised canonical correlation analysis.  
*Biostatistics*, 15(3):569-83.
-  Zitnik M., Nguyen F., Wang B., Leskovec, J., Goldenberg A., Hoffman M.M. (2019).  
Machine learning for integrating data in biology and medicine:Principles, practice, and opportunities.  
*Information Fusion* 50: 71-91

# Log-linear Poisson graphical model(llgm)

[Allen and Liu, 2012]

- Power transformation of the data:  $x_{ij} \rightarrow x_{ij}^\alpha$ ,  $\alpha \in ]0, 1]$
- Let  $z_j = (x_{1j}^\alpha, \dots, x_{nj}^\alpha)$  be the transformed vector of expression values for gene  $j$

$$p(Z_{ij}|z_{i(-j)}) \sim \mathcal{P}(\mu_j) \text{ with } \log(\mu_j) = \sum_{j' \neq j} \beta_{jj'} \tilde{z}_{ij'}$$

where  $\tilde{z}$  corresponds to a standardization of the log-transformed data

- edge between genes  $j$  and  $j'$   $\Leftrightarrow \beta_{jj'} \beta_{j'j} \neq 0$
- sparse model  $\rightarrow$  add a  $\ell_1$  penalty to the log-likelihood with a regularization parameter  $\lambda$
- choice of  $\lambda$  with a re-sampling procedure: criterion

# StARS: Stability Approach to Regularization Selection

Choice  $\lambda$  with StARS:

- creation of a vector  $\Lambda$  with decreasing values  $\lambda$
- subsamples of  $X$
- infer a network for each subsample and regularization parameter  $\lambda$  of vector  $\Lambda$

Choice  $\lambda_{opt}$

$$\lambda_{opt} = \operatorname{argmin}_{\lambda} \left\{ \min_{0 \leq \rho \leq \lambda} \left[ \sum_{j < k} 2\bar{A}_{jk}(\rho)(1 - \bar{A}_{jk}(\rho)) / \binom{p}{2} \right] \leq \beta \right\}$$

where

$$\bar{A}_{jk}(\lambda) = \frac{1}{B} \sum_{b=1}^B A_{jk}^{(b)}, \beta = 0.05 \text{ by default}$$

## How choose the threshold $\sigma$ ?

Affinity score:  $s(i,j) = \frac{1}{q} \sum_{k=1}^q \mathbb{I}_{\{|y_{ik} - y_{jk}| < \sigma\}}$

Criterion: study of averaged inertia intra- $\mathcal{D}(i)$ :

$$V_{intra} = \frac{\sum_i \frac{\sum_{d: \text{donor of } i} (x_i - x_d)^2}{D_i}}{n}$$

where

- $n$ : number of missing individuals
- $D_i$ : number of donors for individual  $i$ .

[Back to similarity](#)

# Precision/recall

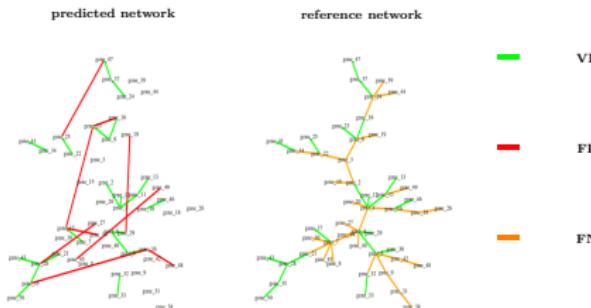
[Back to evaluation process](#)

- Precision:  $\text{Pr} = VP / (VP + FP)$

$$\frac{\text{number of } \mathbf{\text{predicted}} \text{ edges present in the reference network}}{\text{total number of predicted edges}}$$

- Recall:  $R = VP / (VP + FN)$

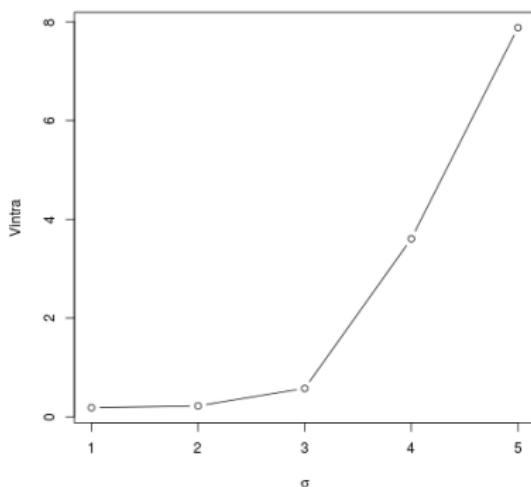
$$\frac{\text{number of } \mathbf{\text{predicted}} \text{ edges present in the reference network}}{\text{number of edges in the reference network}}$$



# Choice of $\sigma$ , distribution of appearance of edges

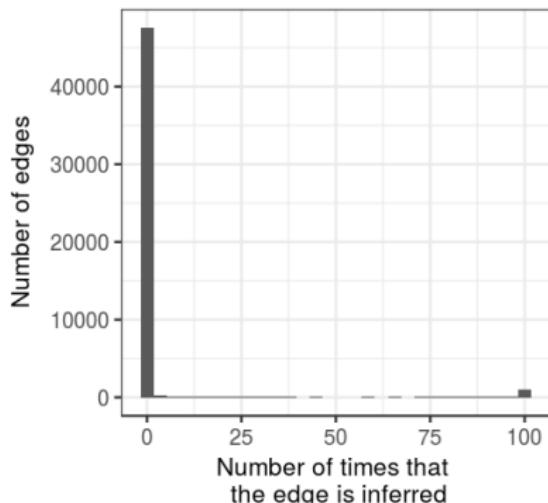
DiOGenes, CID1, 20% missing individuals

Choice of  $\sigma$

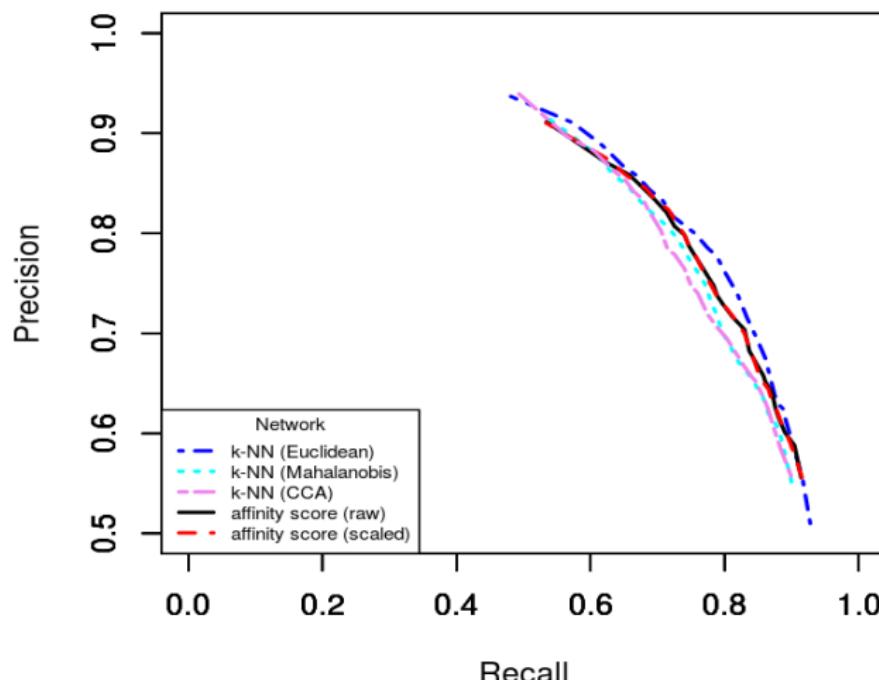


Choice:  $\sigma = 3$

Distribution of appearance of edges (among the  $M$  network)

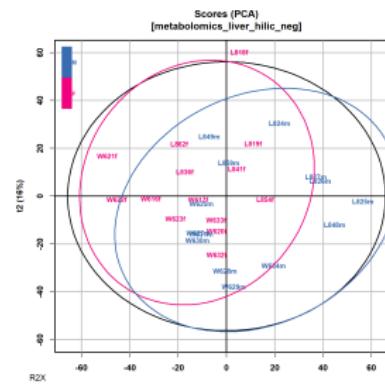
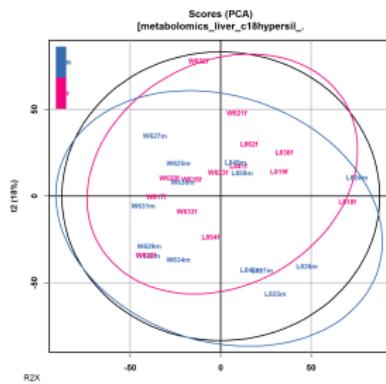
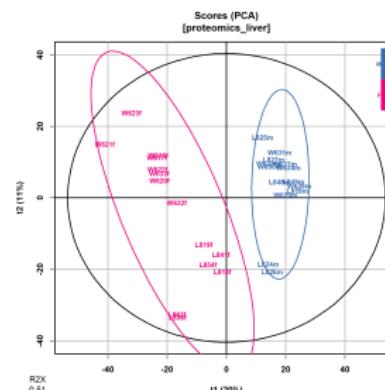
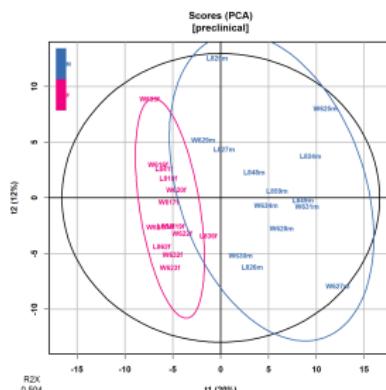


# Impact of the similarity chosen to create the pool of donors



# PCA, liver,colored by sex

[Back to PCA](#)



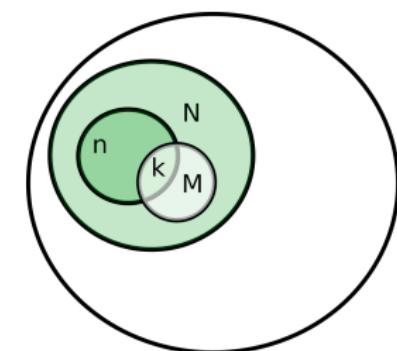
# ORA

**Null hypothesis:** Features in pathways are no more differentially expressed than those outside of pathway

**Proba. to observe at least  $k$  features of interest in a pathway by chance:**

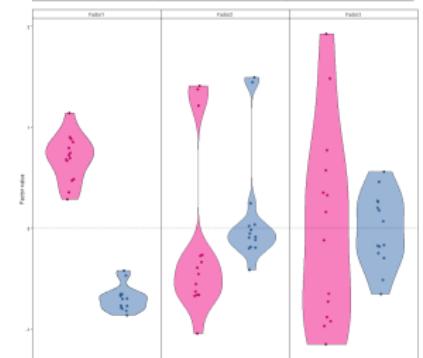
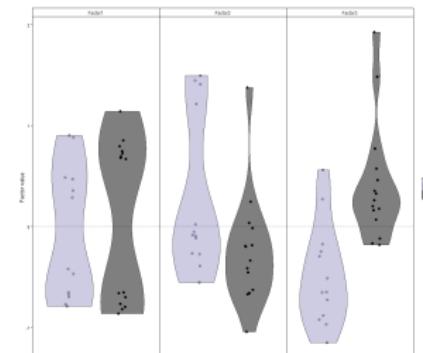
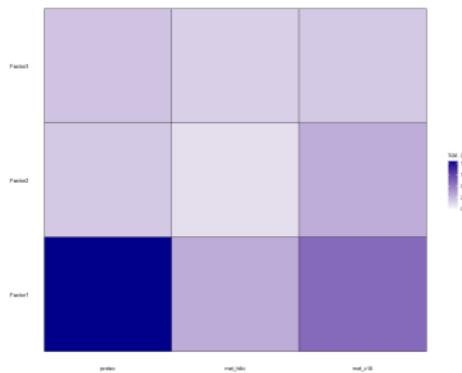
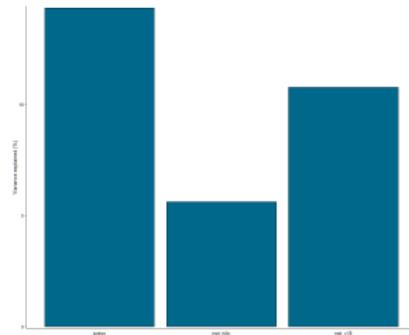
$$P(X \geq k) = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}}$$

- $N$ : size of background set
- $n$ : nb. of metabolites of interest
- $M$ : nb. of metabolites in the background set annotated to the  $i^{th}$  pathways
- $k$ : nb. of metabolites of interest which are annotated to the  $i^{th}$  pathways



Fisher's exact test or the test using hypergeometric distribution

# MOFA: size of block effect



Go to MOFA

# Multiple co-inertia analysis

MCIA is a multi-omics exploratory data analysis technique (*Meng et al. 2016*). The datasets are projected into the same dimensional space by defining both ‘global’ and ‘block-specific’ scores (and loadings), and maximizing the sum squared covariance between them (*Meng et al. 2014*).

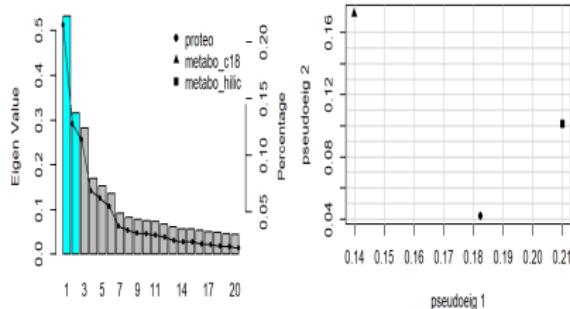
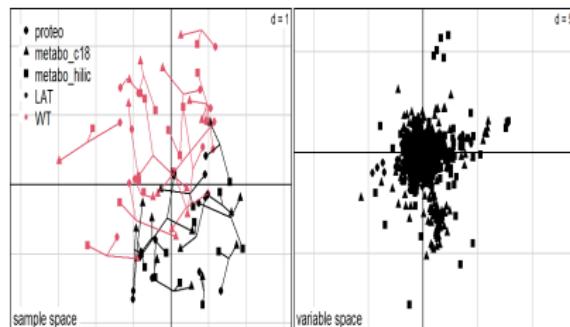
R package `omicade4`

[Back to MOFA](#)

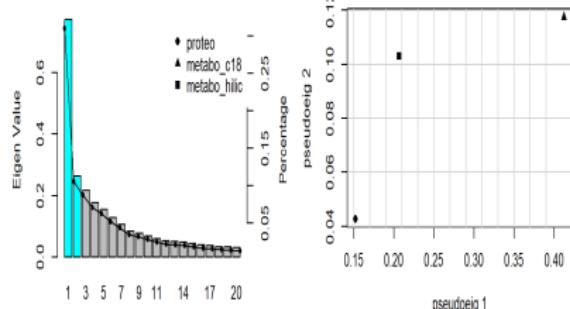
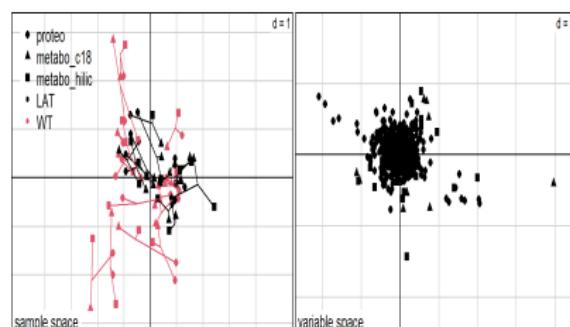
# Multiple co-inertia analysis

Colored by gene

All metabolites

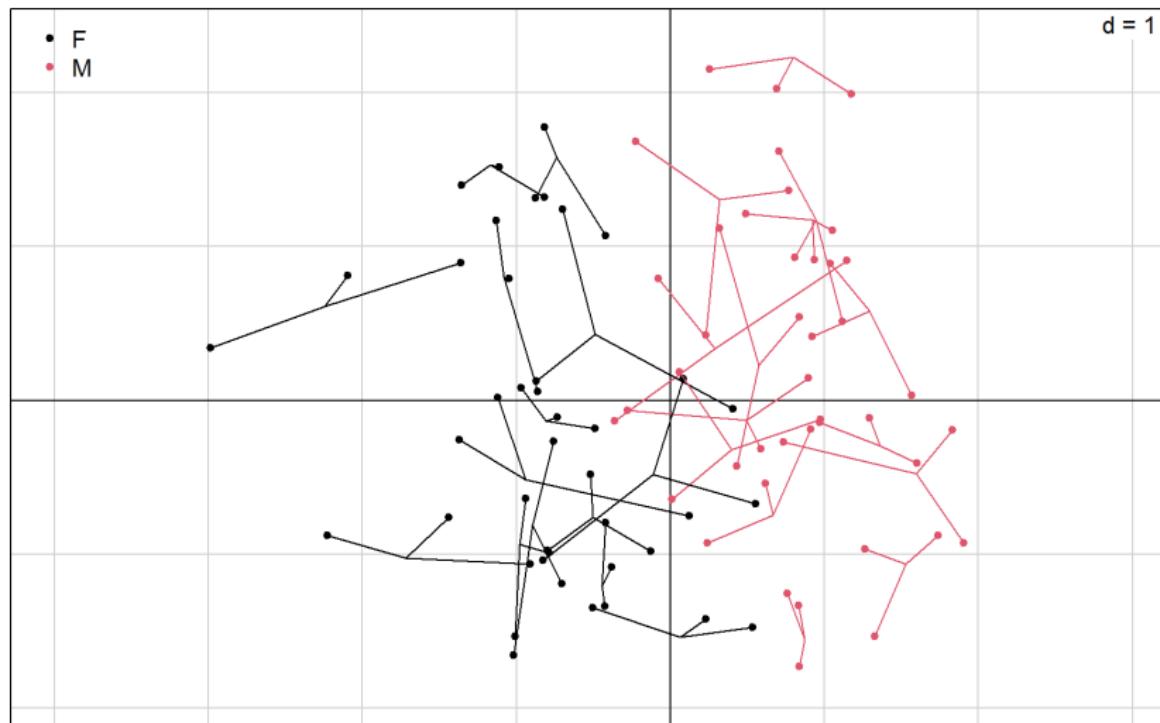


Only annotated metabolites



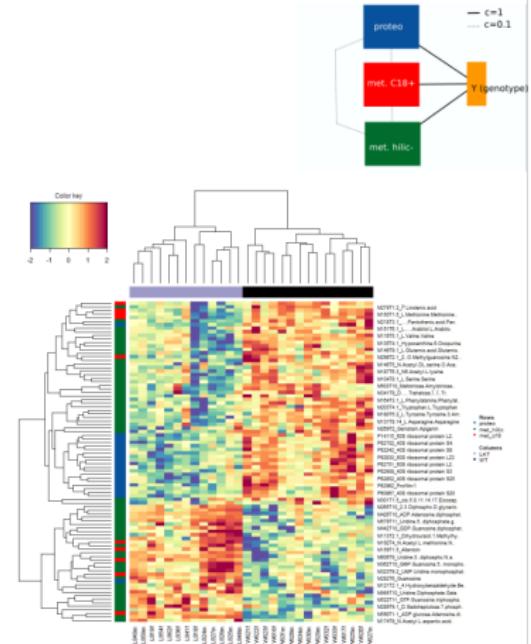
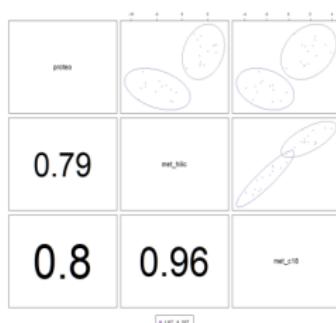
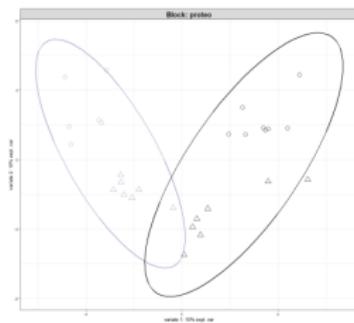
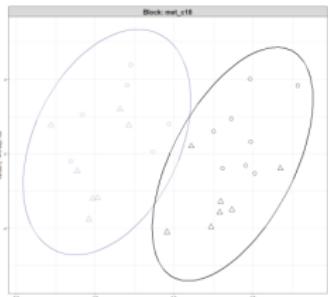
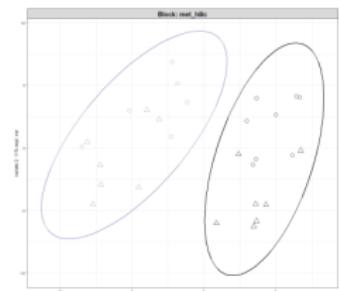
# Multiple co-inertia analysis

Colored by Sex, all metabolites



# SGCCA

Only annotated metabolites



Back to SGCCA