Imputation of missing individuals for network inference from RNA sequencing data

Alyssa Imbert
Supervised by Nathalie Villa-Vialaneix and Nathalie Viguerie

Netbio

09/11/2017

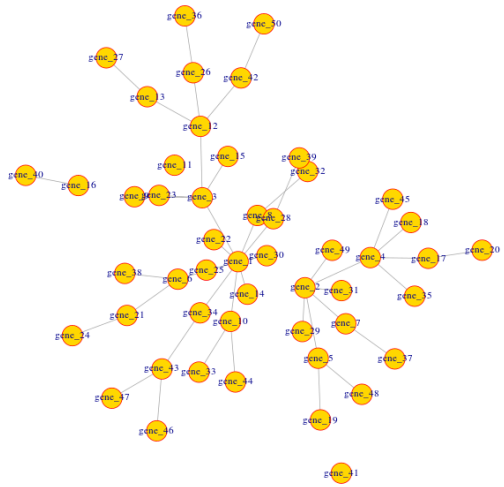
Table of contents

Principle

Network : visualization of interactions between genes

large RNA-seq expression data
$(n \ll p)$

$$\text{individuals } n\ (n \ll p) \quad \underbrace{\left\{ X = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & X_i^j & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \right.}_{\text{variables (gene expressions), } p}$$



Aim : obtain a network with

- node : gene ;
- edges : significant and direct co-expression between two genes

Graphical Gaussian Model (GGM)

framework : micro-array

- $(X_i)_{i=1,\dots,n}$ Gaussian random variables i.i.d. $(\mathcal{N}(0,\Sigma))$ $(j = 1,\dots,p)$
- Use of partial correlations : $\pi_{jj'} = \text{cor}(X^j, X^{j'} | X^k, k \neq j, j')$

$$j \text{ and } j' \text{ are linked} \Leftrightarrow \text{cor}(X^j, X^{j'} | (X^k)_{k \neq j,j'}) \neq 0$$

Graphical Gaussian Model (GGM)

framework : micro-array

- $(X_i)_{i=1,\dots,n}$ Gaussian random variables i.i.d. $(\mathcal{N}(0,\Sigma))$ $(j = 1,\dots,p)$
- Use of partial correlations : $\pi_{jj'} = \text{cor}(X^j, X^{j'} | X^k, k \neq j, j')$

$$j \text{ and } j' \text{ are linked} \Leftrightarrow \text{cor}(X^j, X^{j'} | (X^k)_{k \neq j,j'}) \neq 0$$

Various approaches to infer gene expression networks :

- *Schäfer and Strimmer (2005)*
  - with bootstrapping or shrinkage and a proposal for a Bayesian test for significance
- Sparse approaches :
  - *Meinshausen and Bühlmann (2006)*
  - *Friedman and al. (2008)*

Network inference and RNA-seq data

- **RNA-seq data** :
  - counts $\rightarrow$ discrete data ;
  - over-dispersed data (variance $>$ mean).
- **Network inference method** :
  - Transform data $\rightarrow$ approach gaussian distribution
    $\rightarrow$ GGM
  - Use appropriate models based on Poisson distribution
    - ⋆ Log-linear Poisson graphical model(llgm), *Allen et Liu (2012)* ;
    - ⋆ Hierarchical log-normal Poisson graphical model, *Gallopin & al. (2013)*.

Log-linear Poisson graphical model(llgm)

Allen G.I. et Liu Z., 2012

- Power transformation of the data : $x_{ij} \rightarrow x_{ij}^{\alpha}$, $\alpha \in]0, 1]$
- Let $z_j = (x_{1j}^{\alpha}, ..., x_{nj}^{\alpha})$ 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 ℓ_1 penalty to the log-likelihood with a regularization parameter λ
- choice of λ with a re-sampling procedure : criterion StARS [1]

1. Stability Approach to Regularization Selection *Liu H. et al., 2010*

StARS

Choice λ with StARS :

- creation of a vector Λ with decreasing values λ
- subsamples of X
- infer a network for each subsample and regularization parameter λ of vector Λ
- $\lambda_{opt} = argmin_{\lambda} \left\{ min_{0 \leq \rho \leq \lambda} \left[\sum_{j<k} 2\bar{A}_{jk}(\rho)(1 - \bar{A}_{jk}(\rho)) / \binom{p}{2} \right] \leq \beta \right\}$ where $\bar{A}_{jk}(\lambda) = \frac{1}{B} \sum_{b=1}^{B} A_{jk}^{(b)}$, $\beta = 0.05$ by default

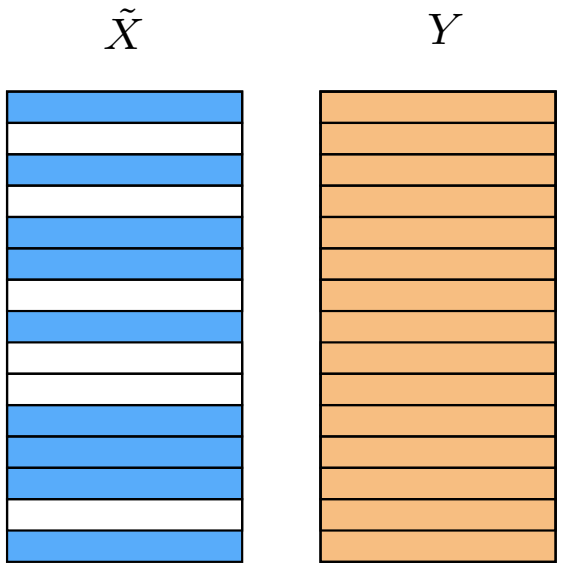
Table of contents

Motivation

- RNA-seq data : generally, few samples
  $\hookrightarrow$ infer network is difficult
- Network inference is sensitive to influential observations, *Bar-Hen A., 2016*.
- Aim : Find a solution to limit the loss of information
- Auxiliary data : bring information
  $\hookrightarrow$ use this supplementary information to improve network inference

Framework and notation

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



Problem

Search an imputation method which allows to :

- preserve the link between variables (genes)
  $\rightarrow$ impute missing individuals **entirely** = impute simultaneous all variables

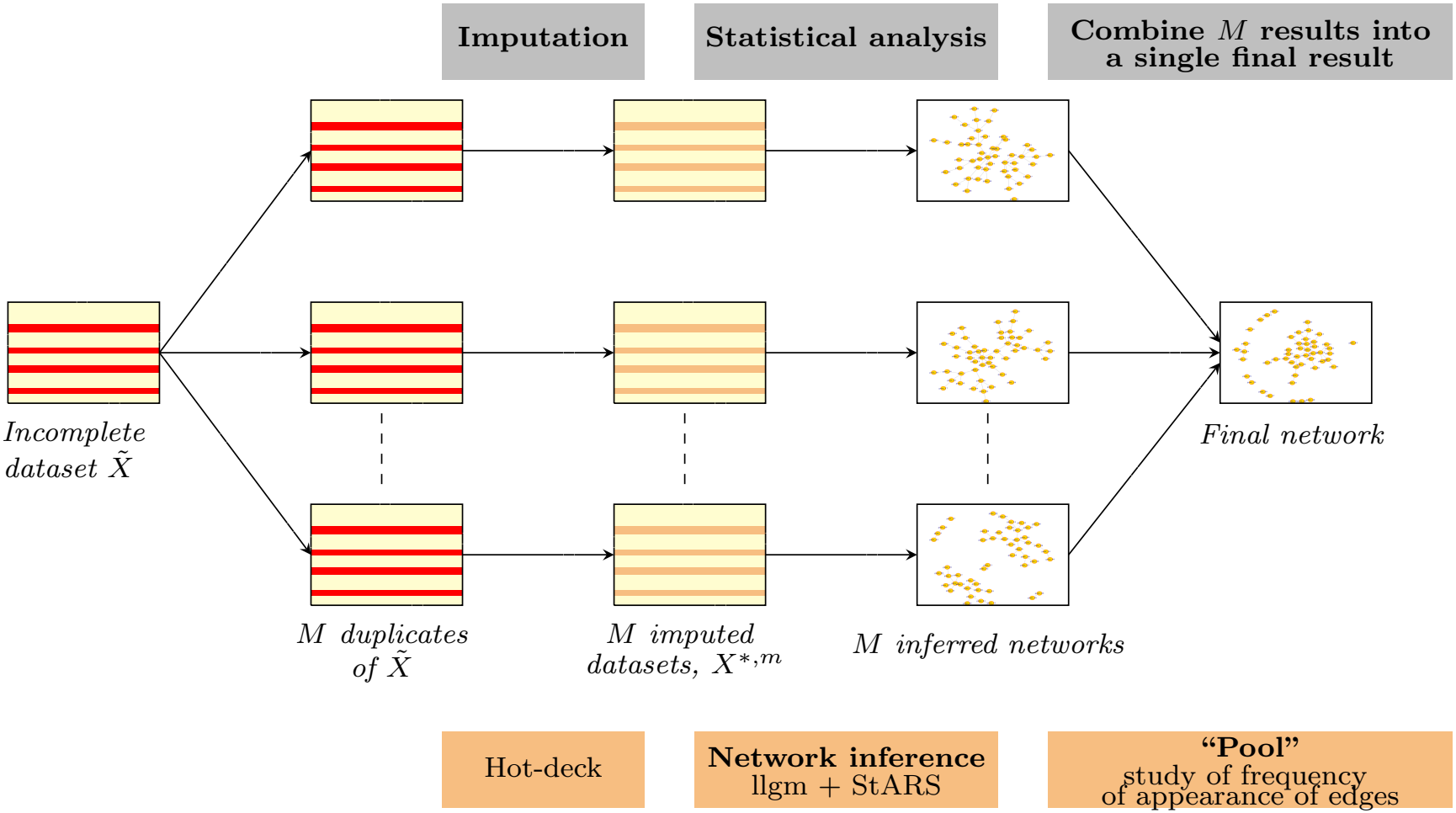
- Take into account uncertainty which is linked to imputation

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

Table of contents

Multiple hot-deck imputation (hd-MI)

General schema

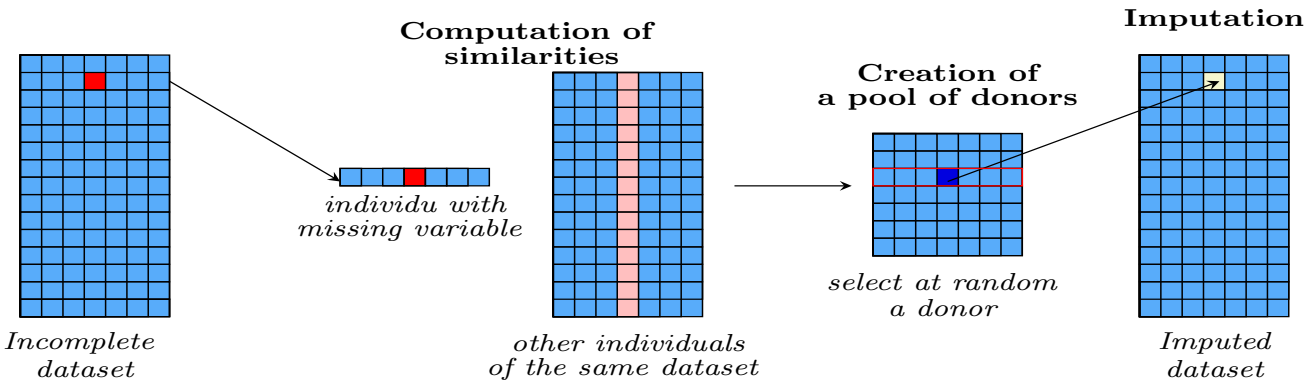


llgm = log-linear Poisson graphical model (*Allen et Liu, 2012*)

Hot-deck imputation [2]

A set of methods based on the concept of "donors"

- Definition

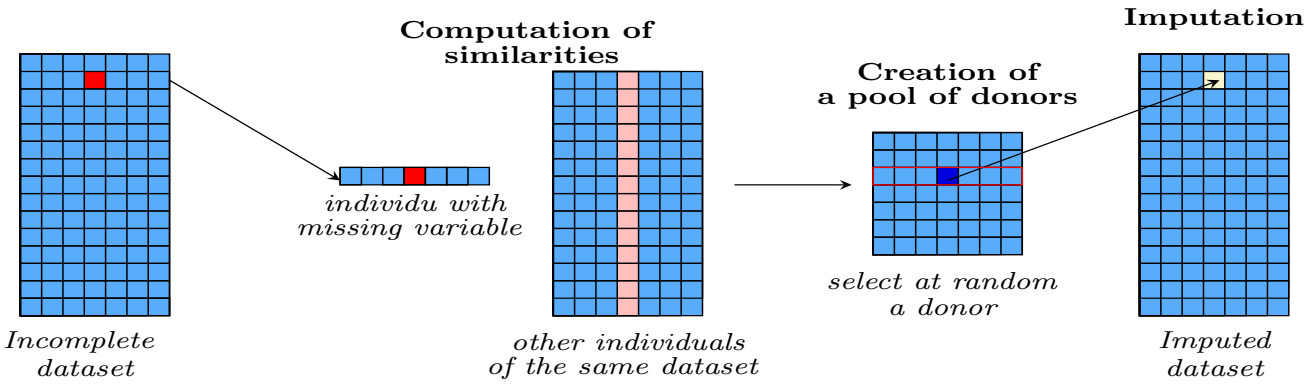


2. Revue : *Andridge R. R. et Little R. J. A., 2010*

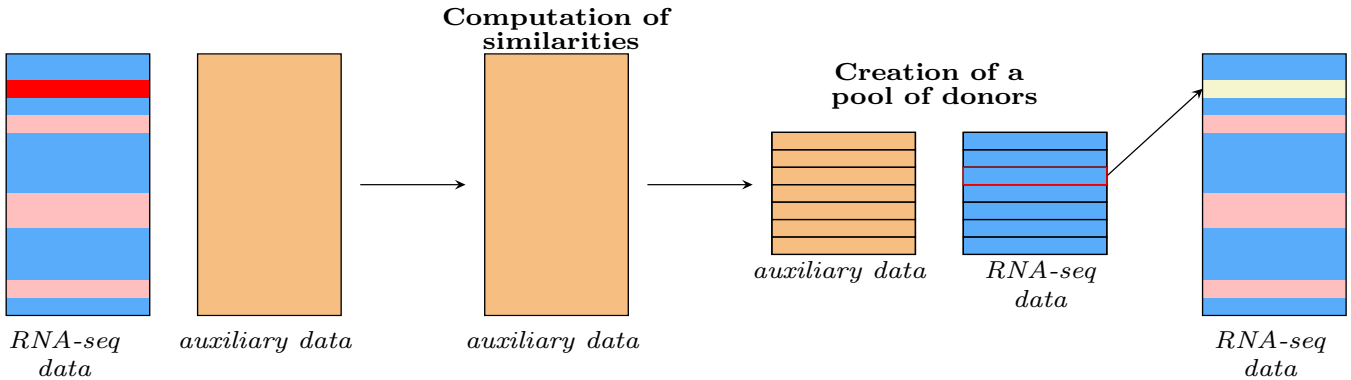
Hot-deck imputation [2]

A set of methods based on the concept of "donors"

- Definition



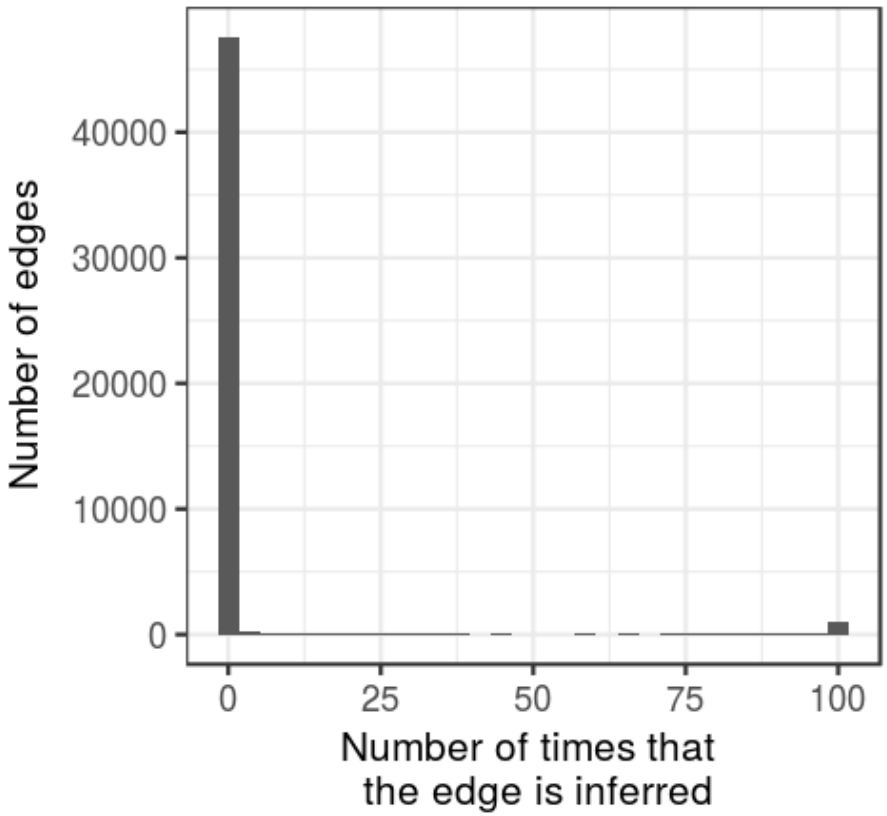
- Our case :



2. Revue : *Andridge R. R. et Little R. J. A., 2010*

Illustration : "Pool"

Frequency of appearance of edges
Example for $M = 100$ networks :



- study the number of times an edge is predicted among the M networks :

$$r(e) = \frac{\text{number of times the edge } e \text{is predicted}}{M}$$

- Choice a reliability threshold : r_0
- Final network composed of the edges e such that $r(e) \geq r_0$

Multiple hot-deck imputation

Test 2 approaches :

- with **an affinity score** [3] (R package hot.deck) :

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

where σ = fixed threshold and $\mathcal{D}(i) = \{j : \ s(i,j) = \max_{l\neq i} s(i,l)\}$

- with k **nearest neighbors** (knn), euclidean metric :

$$d(i,j) = \sum_{k=1}^{q}(y_{ik} - y_{jk})^2$$

3. *Cranmer S.J. and Gill J., 2012*

How choose the threshold σ ?

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-$\mathcal{D}(i)$:

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

where

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

Table of contents

Framework

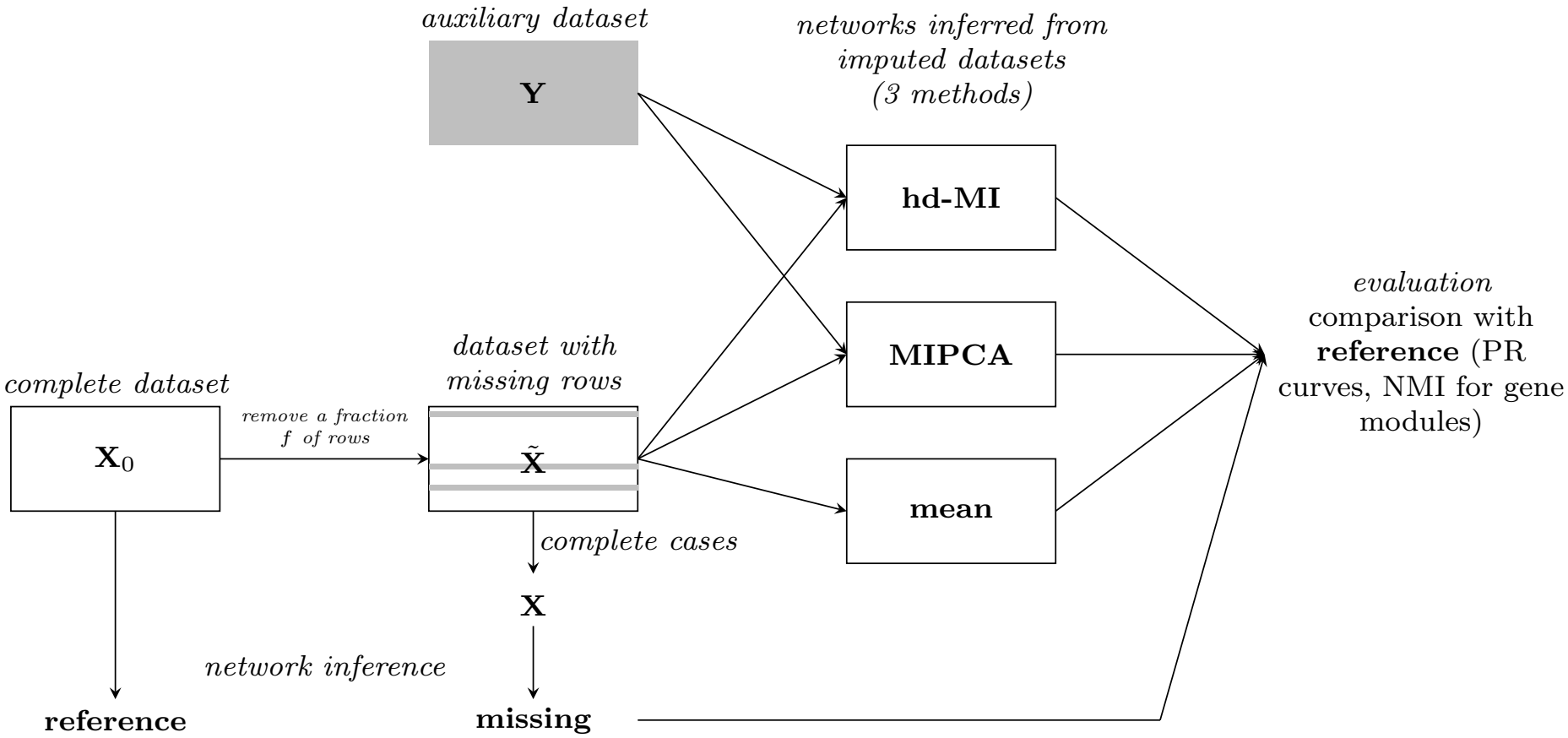
- Test on real datasets coming from 2 projects :
  - GTEx : Genotype-Tissue Expression [4],
  - DiOGenes [5],

- 3 imputation methods :
  - simple and naive method : imputation by mean
  - multiple imputation based on PCA : MIPCA, *Josse et al., 2011*
  - our method : hd-MI

- 10%, 20%, 30%, 40% missing individuals

4. *Lonsdale et al., 2013*
5. *Larsen et al., 2010*

Overview of the evaluation process



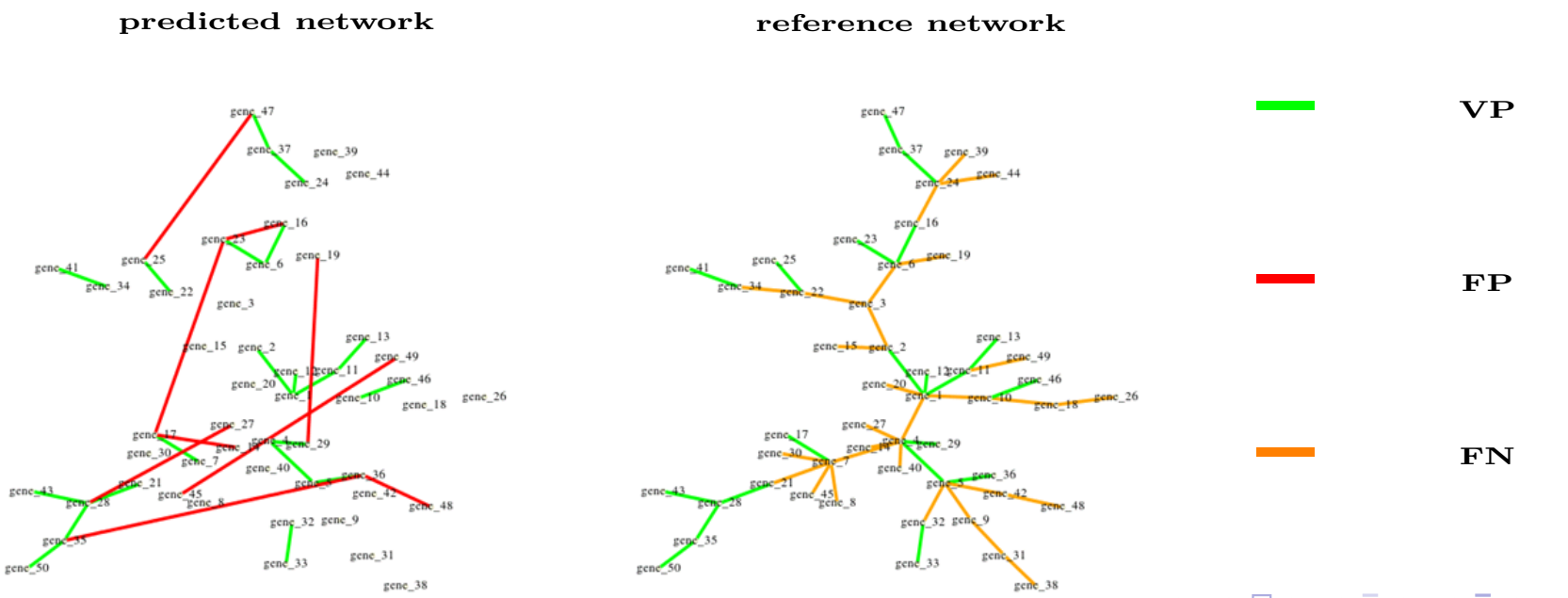
Precision/recall

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

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

- Rappel : $\mathrm{R} = VP/(VP + FN)$

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



Gene modules

- **Aim** : see if imputation preserve gene modules
- Search gene modules in the different networks : clustering of nodes
- Comparison with gene modules obtained with reference network : NMI [6]
  - NMI between $[0, 1]$
  - NMI $= 1$: modules between the 2 networks are the same
  - NMI $= 0$: modules between the 2 networks are independent

6. normalized mutual information measure, *Danon L. and al (2005)*

Table of contents

GTEx

Data presentation

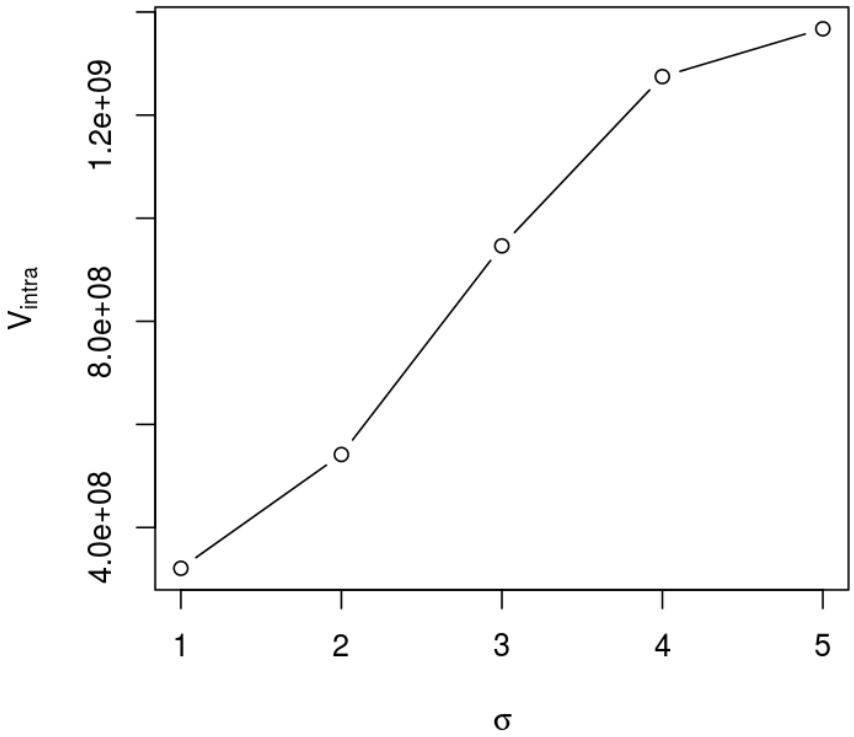
- RNA-seq data on more than 30 human tissues ;
- **Choice of 2 tissues** : 2 tissues whose expression profile are close [7]
  - X : lung,
  - Y : thyroid,
- normalization of RNA-seq data RNA-seq : TMM package edgeR ;
- description of datasets :
  - 320 samples for X,
  - 323 samples for Y,
- evaluation : keep only **221 common samples** ;
- **select the most variables genes** (higher variances) :
  - for X : p= 100,
  - for Y : q = 50,
  - 36 common genes.
- Results for **20%** missing observations

7. *Melé et al., 2015*

Choice of σ and distribution of appearance of edges

GTEx, 20% missing observations

Choice of σ

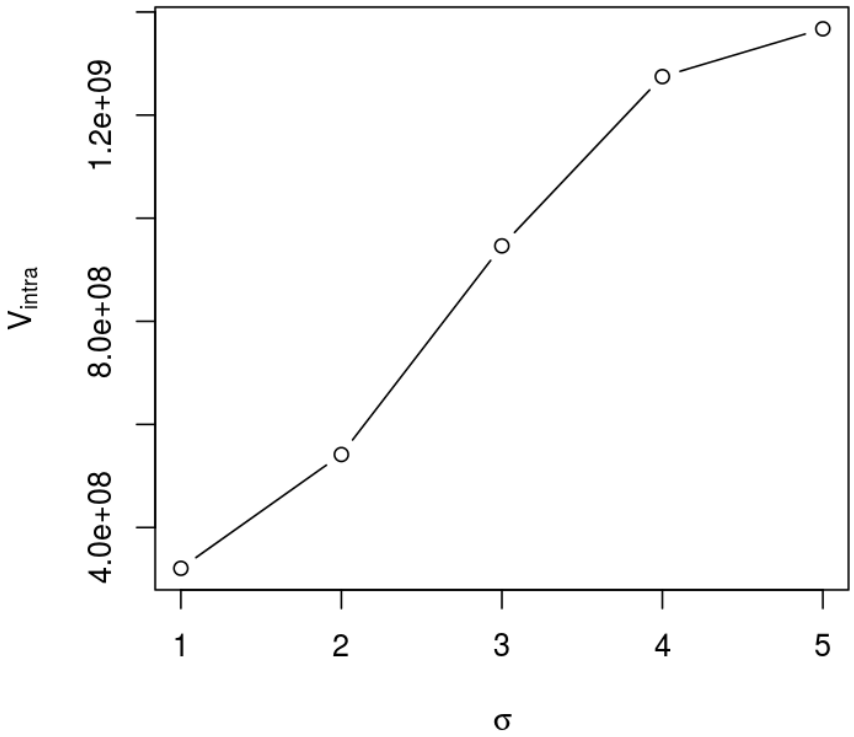


Choice : $\sigma = 2$

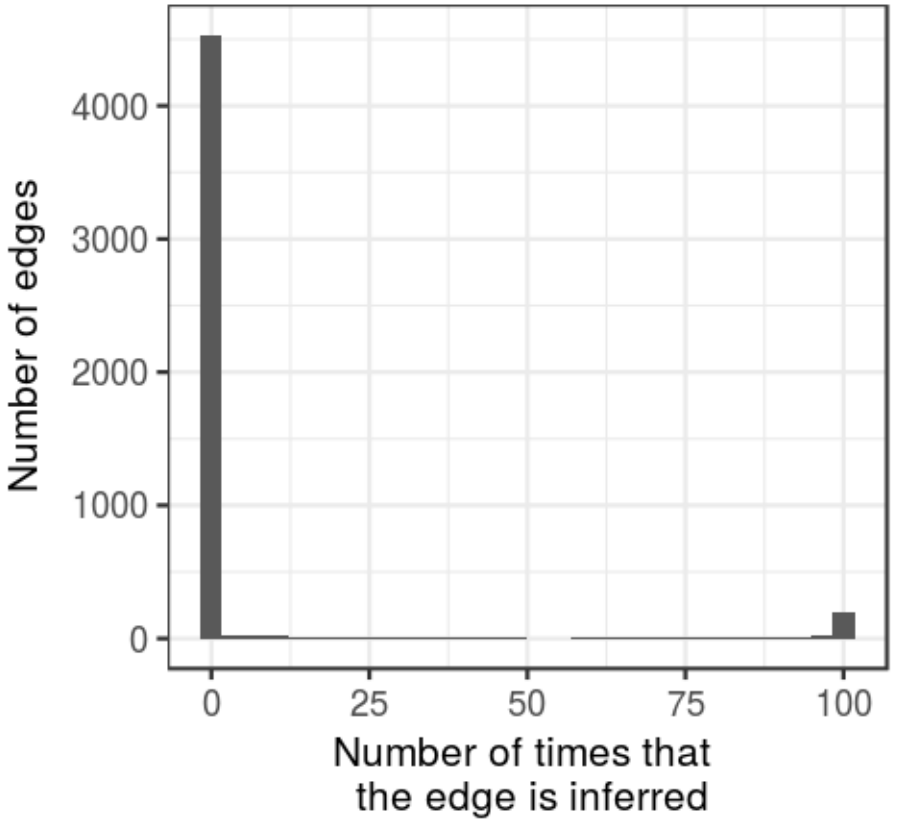
Choice of σ and distribution of appearance of edges

GTEx, 20% missing observations

Choice of σ



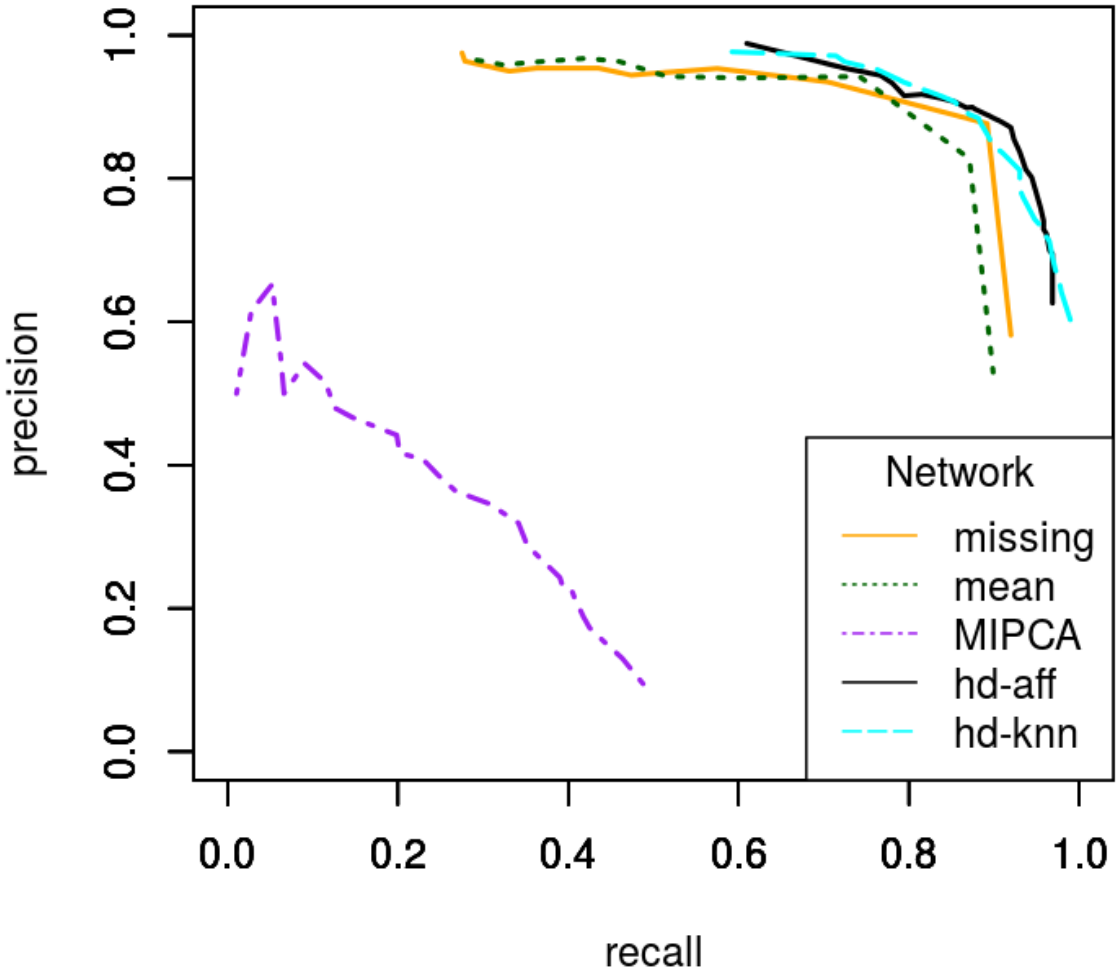
Distribution d'apparition d'une arête



Choice : $\sigma = 2$

Curve precision/recall

GTEx, 20% missing observations



Comparison of gene modules

GTEx, 20% missing observations

- Search of gene modules on the largest component :
  $\hookrightarrow$ function *spinglass_community()*

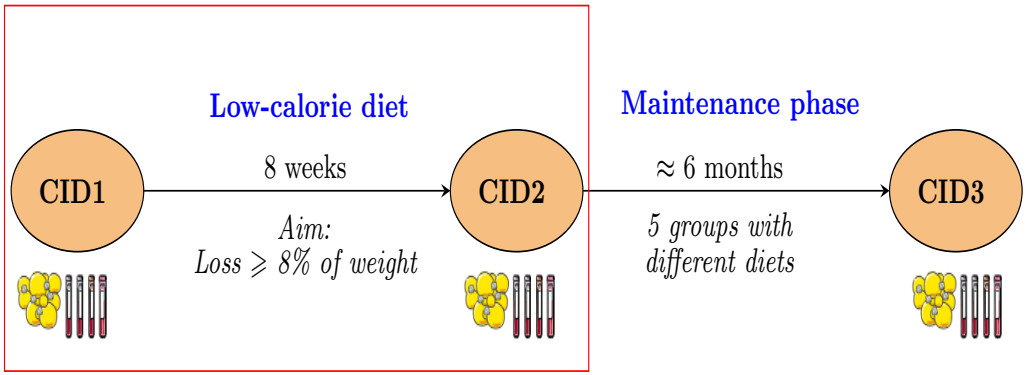
- comparison gene modules : NMI

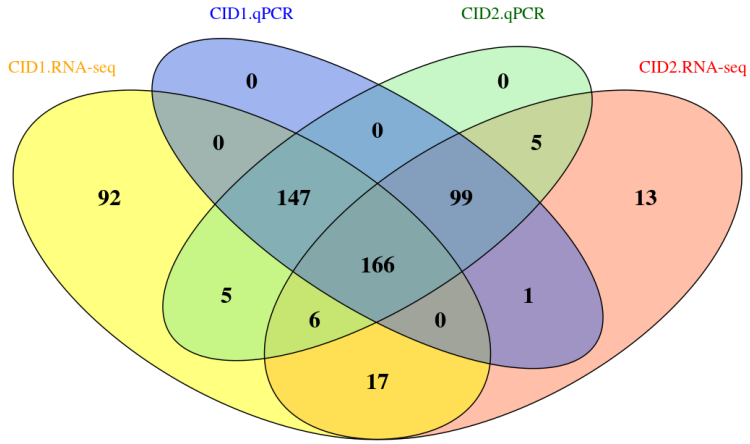
graph	reference	missing	mean	MIPCA	hd-aff	hd-knn
$\sharp$ modules	7	7	7	1	8	8
NMI		0.557	0.573	1 [8]	**0.667**	0.603

8. only 3 genes on the largest component

DiOGenes

Data presentation





RNA-seq :

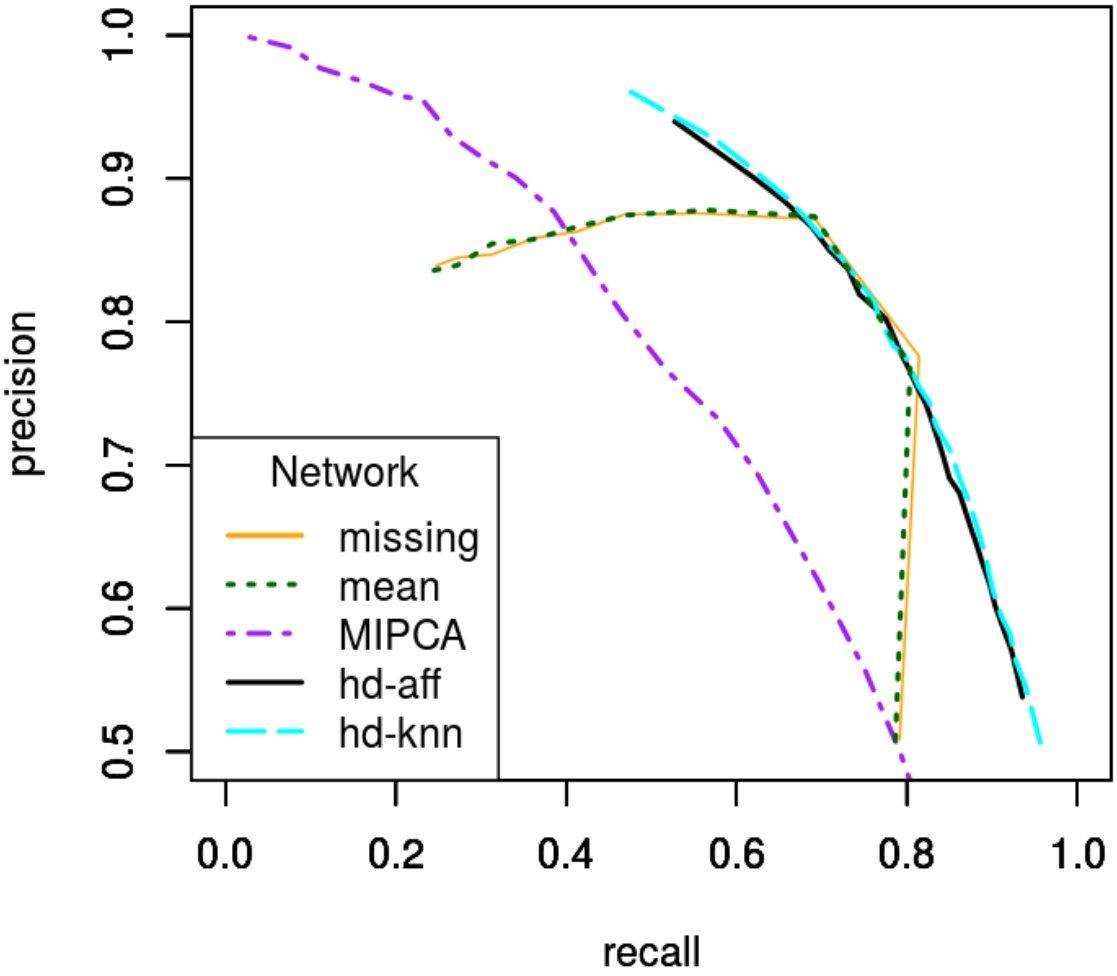
- 433 individuals in CID1,
- 307 individuals in CID2,
- **189 common individuals**,
- 317 genes

Auxiliary data : RT-qPCR :

- 166 individuals for CID1,
- 172 individuals for CID2,
- 284 genes.

Curves precision/recall, CID1

DiOGenes, 20% missing observations



Gene modules, CID1

DiOGenes, 20% missing observations

- Search of gene modules on the largest component :
  ↪ function *spinglass_community()*

- comparison gene modules : NMI

graph	reference	missing	mean	MIPCA	hd-aff	hd-knn
♯ modules	7	7	7	10	8	8
NMI		0.526	0.612	0.346	0.493	0.492
NMI with CID2	0.423	0.421	0.424	0.341	0.38	0.383

Conclusion

- For high precision, best recall with our method hd-MI
- less false positives with hd-MI
- GTEx : best nmi with hd-MI
  $\longrightarrow$ preserve gene modules
- beyond 30% missing individuals, results deteriorate :
  $\longrightarrow$ curve PR for hd-MI below missing PR curve

- Article : in revision
- R package : RNAseqNet

Thanks for your attention