RNAseqNet : un package pour l'inférence de réseaux à partir de données RNA-Seq

Alyssa Imbert et Nathalie Villa-Vialaneix

Rencontres R 2018

05/07/2018

Sommaire

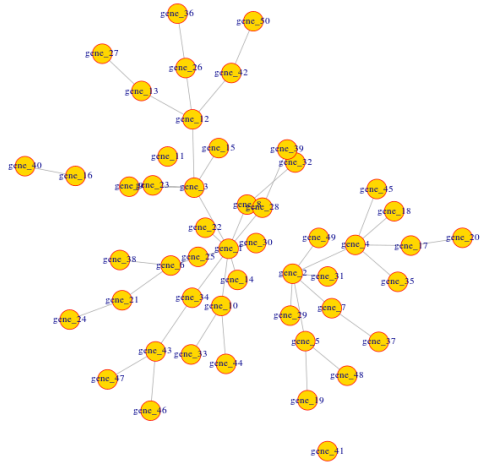
Sommaire

Principe

données d'expression RNA-Seq
$(n \ll p)$

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

Réseau : visualisation des dépendances conditionnelles entre les gènes



Obtenir un réseau :

- nœud : gène ;
- arête : lien direct et fort entre deux gènes

Inférence de réseau et données RNA-Seq

- **données RNA-Seq** :
  - comptages → données discrètes ;
  - données surdispersées (variance > moyenne).
- **Méthodes d'inférence de réseau** :
  - Transformer les données → approche basée sur des distributions gaussiennes
    → modèle graphique gaussien (GGM)
  - Modèles adaptés basés sur des distributions de Poisson
    - ⋆ **modèle log-linéaire de Poisson (llgm)** *[Allen and Liu, 2012]* ;
    - ⋆ modèle hiérarchique log-normal de Poisson *[Gallopin et al., 2013]*.

Modèle graphique log-linéaire de Poisson (llgm)

Allen and Liu, 2012

- Transformation puissance : $x_{ij} \rightarrow x_{ij}^{\alpha}$, $\alpha \in]0, 1]$ (package *PoiClaClu*)
- Soit $z_j = (x_{1j}^{\alpha}, ..., x_{nj}^{\alpha})$ les données transformées pour le gène j

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

- arête entre les gènes j et j' $\Leftrightarrow$ $\beta_{jj'}\beta_{j'j} \neq 0$
- modèle parcimonieux $\rightarrow$ ajout d'une pénalité ℓ_1 à la logvraisemblance avec le paramètre de régularisation λ
- choix de λ via une procédure de rééchantillonnage : critère StARS [1] *[Liu et al., 2010]*

1. Stability Approach to Regularization Selection

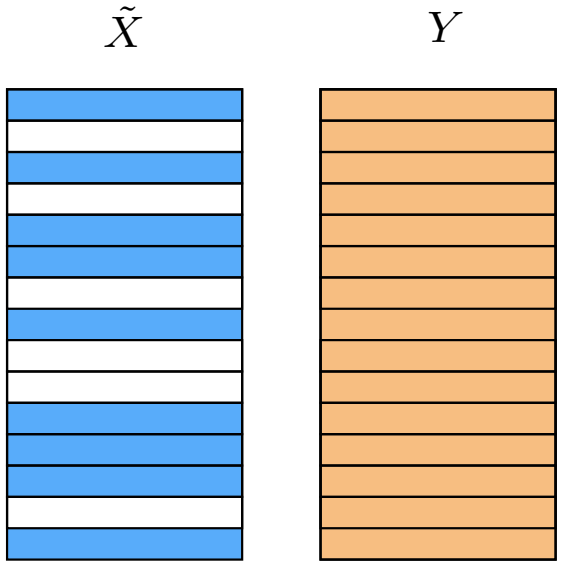
Sommaire

Motivation

- Données RNA-Seq : généralement peu d'échantillons
  $\hookrightarrow$ inférence de réseau difficile
- Inférence de réseau : sensible aux observations influentes
  [Bar-Hen and Poggi, 2016].
- Objectif : Trouver une solution pour limiter la perte d'information
- Données auxiliaires : apport d'information
  $\hookrightarrow$ utiliser cette information pour améliorer la qualité de l'inférence de réseau

Contexte et notations

- Matrice $\tilde{X}$ de taille $n_1 \times p \rightarrow$ données d'expression d'intérêt (RNA-Seq) ;
- matrice Y de taille $n \times q \rightarrow$ données phénotypiques, données d'expression qPCR,... ;
- n_1 échantillons (individus) communs à $\tilde{X}$ et Y ;
- présence de données manquantes $\longrightarrow$ raisons expérimentales



Problème

Chercher une méthode d'imputation qui permet :

- de préserver la structure de corrélation entre les variables
- prendre en compte l'incertitude liée à l'imputation

Objectif : améliorer la qualité de l'inférence de réseau en utilisant de l'information externe

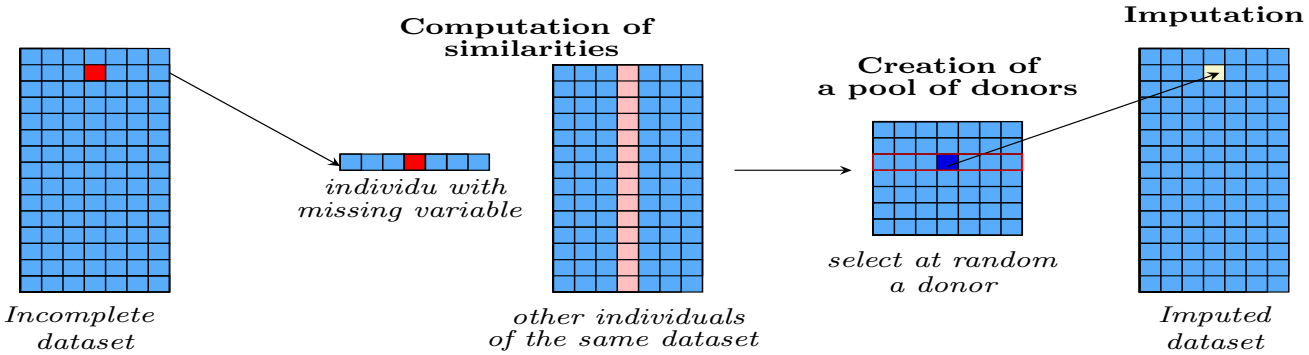
Sommaire

Imputation hot-deck

Un ensemble de méthodes basées sur le concept de "donneurs"

[Andridge and Little, 2010]

- Définition

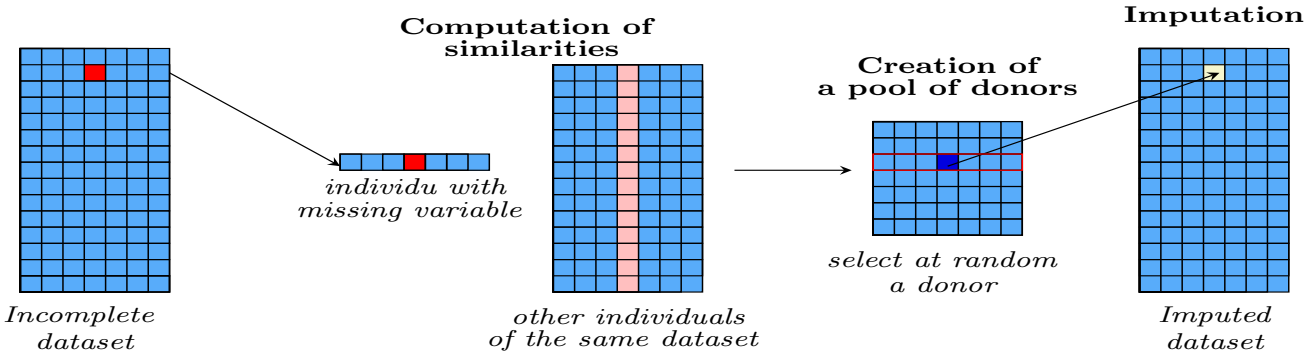


Imputation hot-deck

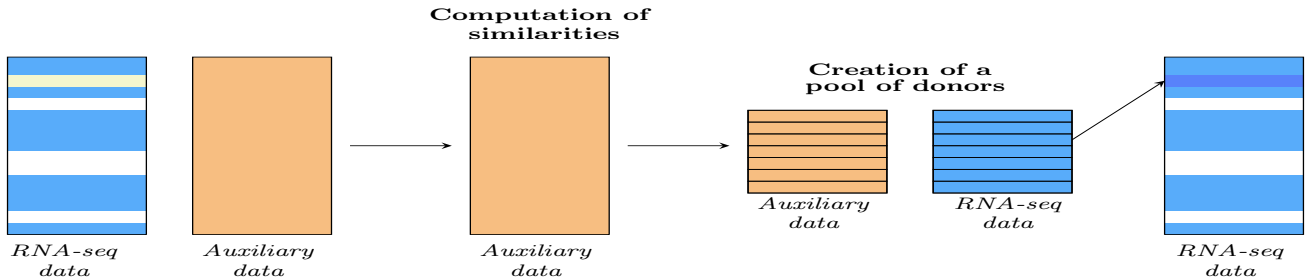
Un ensemble de méthodes basées sur le concept de "donneurs"

[Andridge and Little, 2010]

- Définition

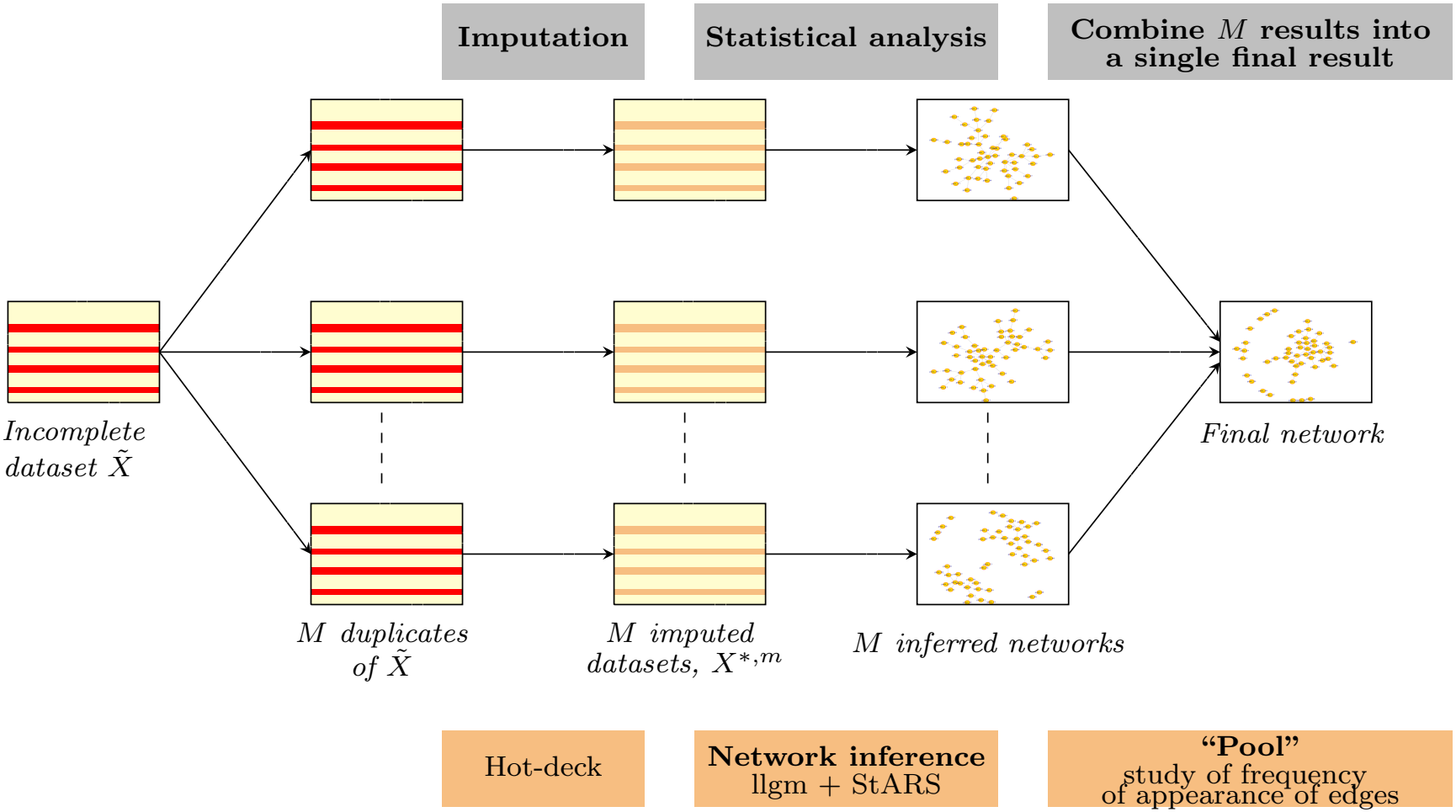


- Notre cas :



Imputation multiple hot-deck(hd-MI)

Schéma



llgm = modèle graphique log-linéaire Poisson *[Allen and Liu, 2012]*

Imputation multiple hot-deck

Tester différentes similarités :

- avec **un score d'affinité** *[Cranmer and Gill, 2012]* (package R *hot.deck*) :

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

où σ = seuil fixé et $\mathcal{D}(i) = \{j : \ s(i,j) = \max_{l\neq i} s(i,l)\}$

- avec différentes approches k **plus proches voisins**

Comment choisir le seuil σ ?

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

Critère : étude de l'inertie intra-$\mathcal{D}(i)$ moyenne :

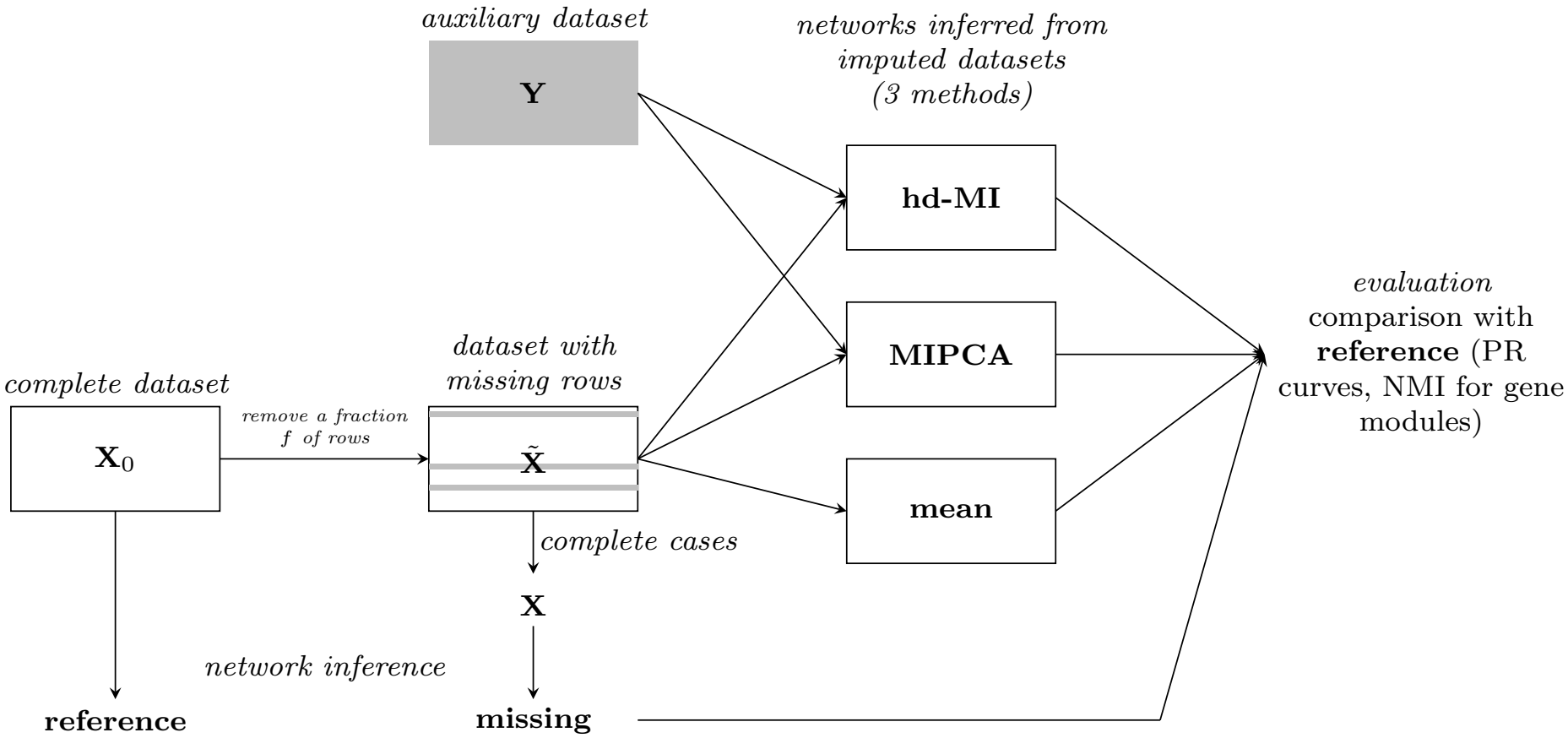
$$V_{intra} = \frac{\sum_i \frac{\sum_{\text{d : donneur de i}} ||x_i - x_d||^2}{D_i}}{n^*}$$

où

- n^* : nombre d'individus manquants
- D_i : nombre de donneurs pour l'individus i.

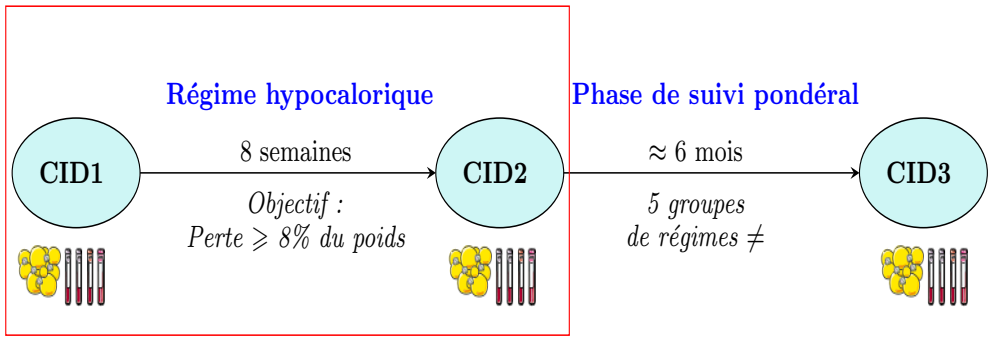
Sommaire

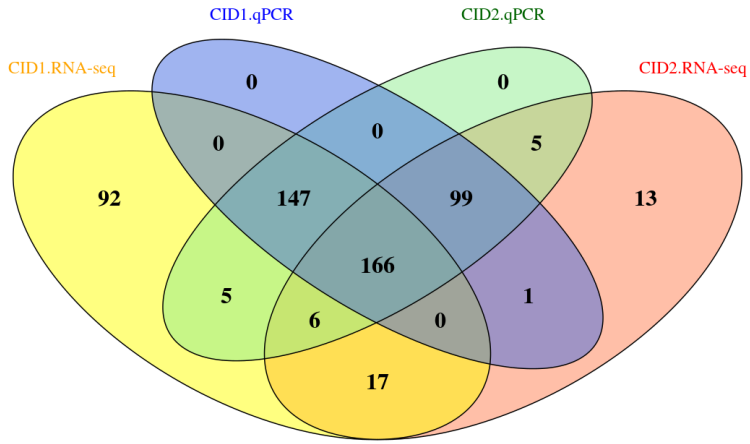
Processus d'évaluation



DiOGenes

Présentation des données





RNA-Seq :

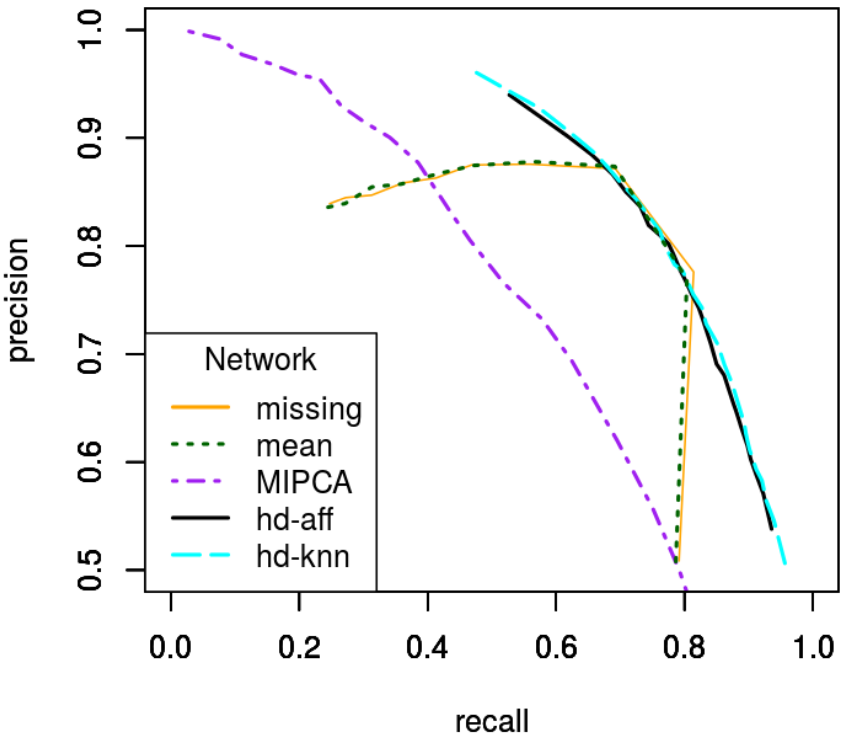
- 433 individus en CID1,
- 307 individus en CID2,
- **189 présents au deux temps**,
- 317 gènes

Données auxiliaires : RT-qPCR :

- 166 individus pour CID1,
- 172 individus pour CID2,
- 284 gènes.

Courbes précision/rappel, CID1

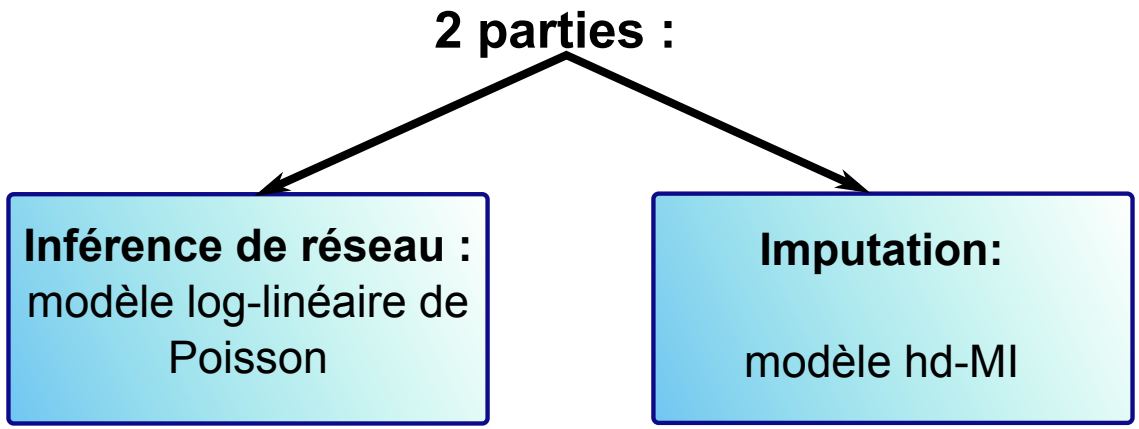
DiOGenes, 20% d'observations manquantes



- Précision élevée $\longrightarrow$ meilleur rappel avec hd-MI
- moins de faux positifs avec hd-MI
- au delà de 30% d'individus manquants $\rightarrow$ résultats détériorés
- **Article** : Imbert et al. (2018), *Multiple hot-deck imputation for network inference from RNA sequencing data*

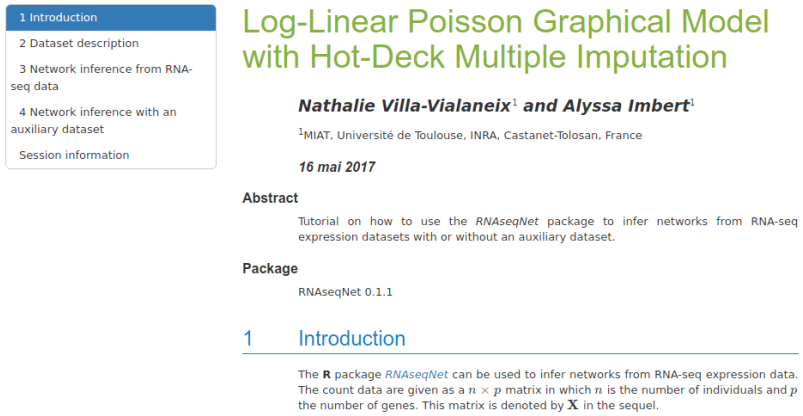
Sommaire

Présentation du package RNAseqNet



Pour tester les fonctions :

- 2 jeux issus du projet GTEx ;
- données d'expression RNA-Seq ;
- proviennent de 2 tissus différents ;
- données normalisées.



Log-Linear Poisson Graphical Model with Hot-Deck Multiple Imputation

Nathalie Villa-Vialaneix[1] *and Alyssa Imbert*[1]

[1]MIAT, Université de Toulouse, INRA, Castanet-Tolosan, France

16 mai 2017

Abstract

Tutorial on how to use the *RNAseqNet* package to infer networks from RNA-seq expression datasets with or without an auxiliary dataset.

Package

RNAseqNet 0.1.1

1 Introduction

The **R** package *RNAseqNet* can be used to infer networks from RNA-seq expression data. The count data are given as a $n \times p$ matrix in which n is the number of individuals and p the number of genes. This matrix is denoted by $\mathbf{X}$ in the sequel.

Inférence de réseau

1. Création d'un vecteur *lambdas*
2. Inférence de réseau :

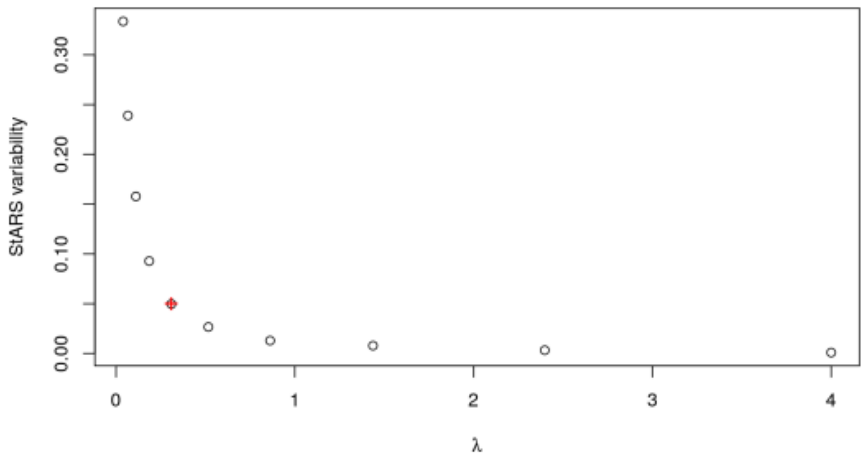
$$\text{gr} \leftarrow \textbf{GLMnetwork}(X, \text{lambdas})$$

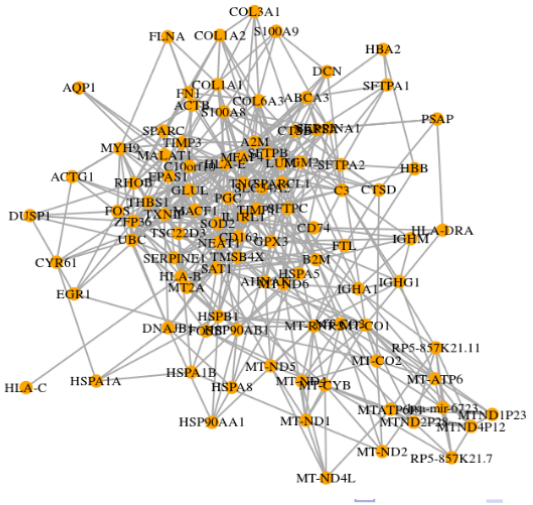
3. Choix du lambda "optimal" (critère StARS) :

$$\text{stability} \leftarrow \textbf{stabilitySelection}(X, \text{lambdas}, B = 50)$$

4. Réseau sélectionné :

$$\textbf{GLMnetToGraph}(\ \text{gr\$path[[stability\$best]]})$$



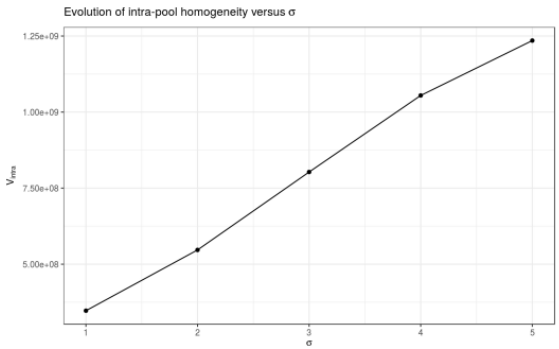


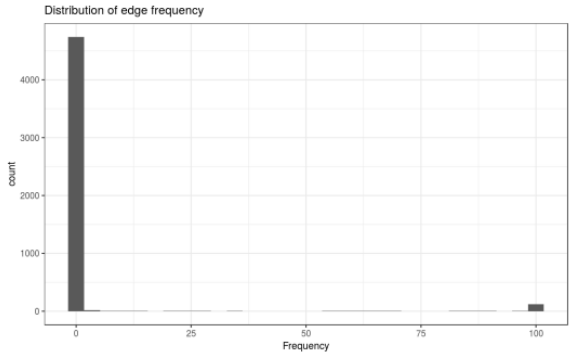
Imputation multiple hd-MI

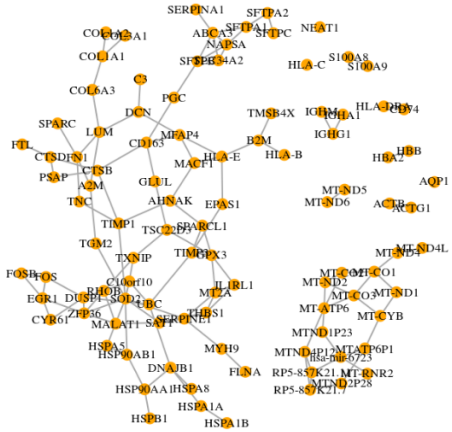
1. Choix de sigma : **chooseSigma**(X, Y, sigmalist)
2. Effectuer hd-MI et inférer un réseau pour chaque jeu imputé :

$$X_{hdmi} \leftarrow \textbf{imputedGLMnetwork}(X,\ Y,\ \text{sigma, lambdas},\ m = 100,\ B = 20)$$

3. Réseau final : **GLMnetToGraph**(X_{hdmi}, threshold $= 0.9$)







Différentes classes disponibles

Fonction R	Classe de la sortie	Fonctions associées
GLMnetwork	*GLMpath*	print, summary
stabilityselection	*stars*	print, summary, plot
imputedGLMnetwork	*HDpath*	print, summary, plot
imputeHD	*HDImputed*	print, summary

Merci de votre attention

Quelques références

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.

Bar-Hen, A. and Poggi, J. (2016).
Influence measures and stability for graphical models.
Journal of Multivariate Analysis, 147 :145–154.

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.

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).

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.

StARS

Choix de λ avec StARS :

- création d'un vecteur Λ avec des valeurs décroissantes de λ
- sous-échantillons de X
- inférer un réseau sur chaque sous-échantillon et élément du vecteur Λ

Choix de λ_{opt}

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

où
$\bar{A}_{jk}(\lambda) = \frac{1}{B} \sum_{b=1}^{B} A_{jk}^{(b)}$, $\beta = 0.05$ par défaut

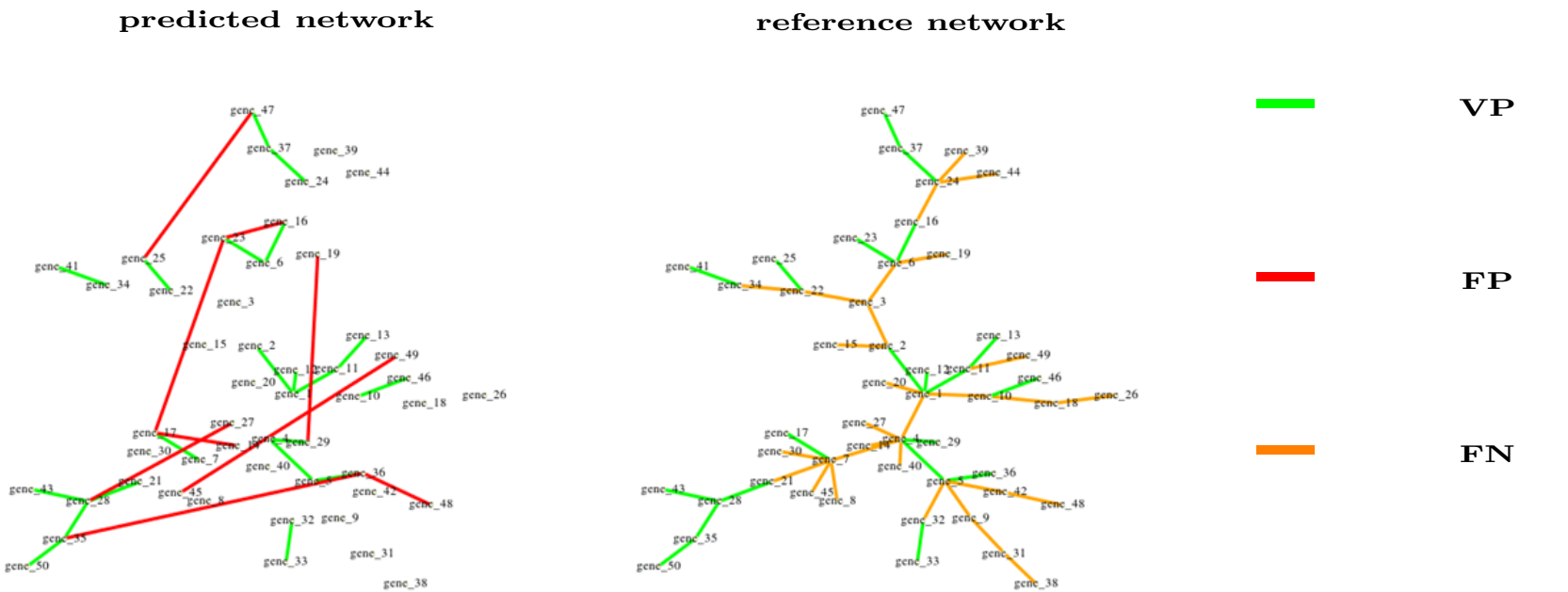
Précision/rappel

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

$$\frac{\text{nombre d'arêtes } \textbf{prédites} \text{ présentes dans le réseau de référence}}{\text{nombre total d'arêtes prédites}}$$

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

$$\frac{\text{nombre d'arêtes } \textbf{prédites} \text{ présentes dans le réseau de référence}}{\text{nombre d'arêtes dans le réseau de référence}}$$



Modules de gènes

- **Objectif** : voir si l'imputation préserve les modules de gènes
- Chercher les modules de gènes dans les différents réseaux : classification de nœuds
- Comparaison avec les modules obtenus pour le réseau de référence : critère NMI [2]
  - NMI compris entre $[0, 1]$
  - NMI $= 1$: modules entre les 2 réseaux sont identiques
  - NMI $= 0$: modules entre les 2 réseaux sont indépendants

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

Modules de gènes, CID1

DiOGenes, 20% d'individus manquants

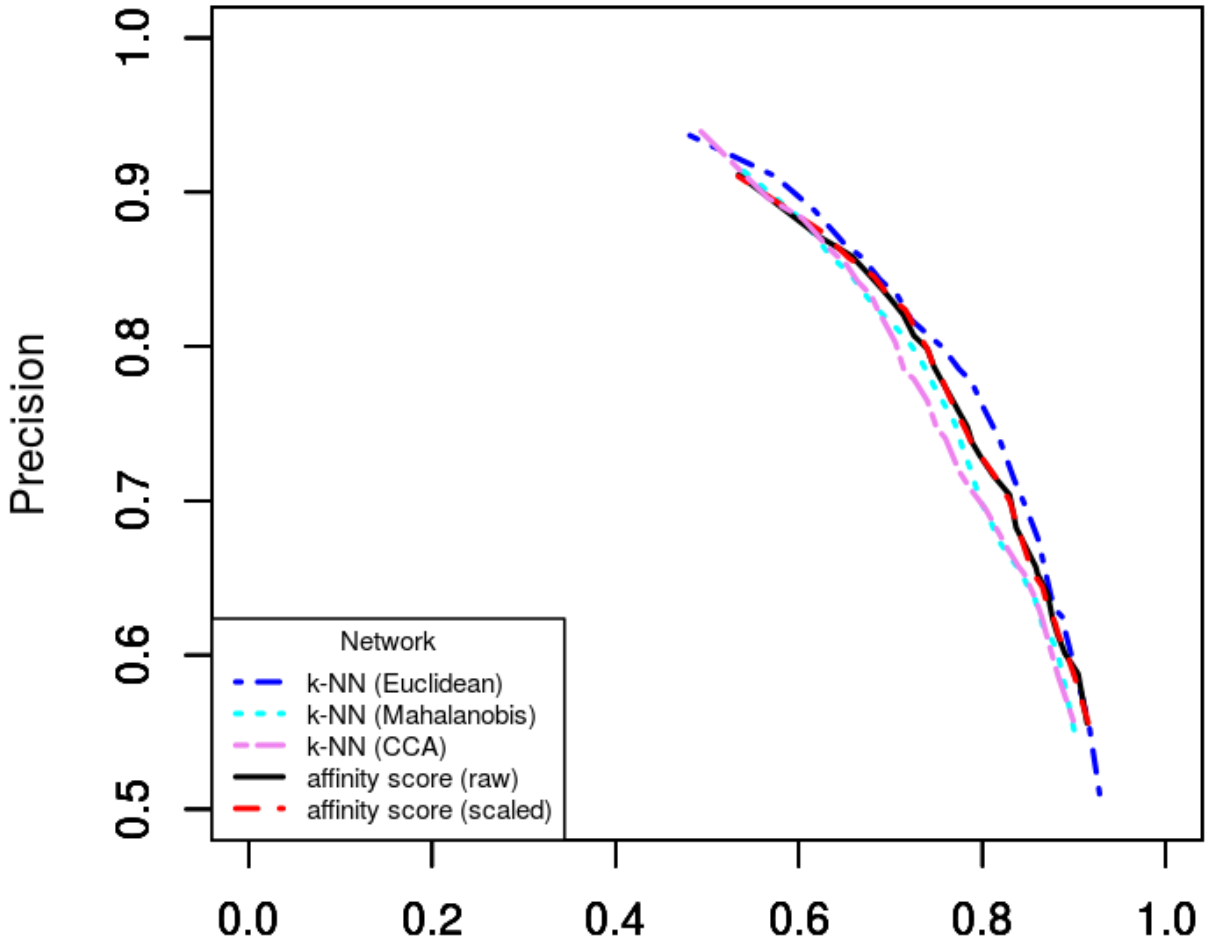
- Chercher les modules de gènes sur la plus large composante connexe :
  $\hookrightarrow$ fonction *spinglass_community()*

- comparaison des modules de gènes : NMI

réseau	reference	missing	mean	MIPCA	hd-MI
♯ modules	7	7	7	10	8
NMI		0.526	0.612	0.346	0.493
NMI avec CID2	0.423	0.421	0.424	0.341	0.38

Tester différentes similarités

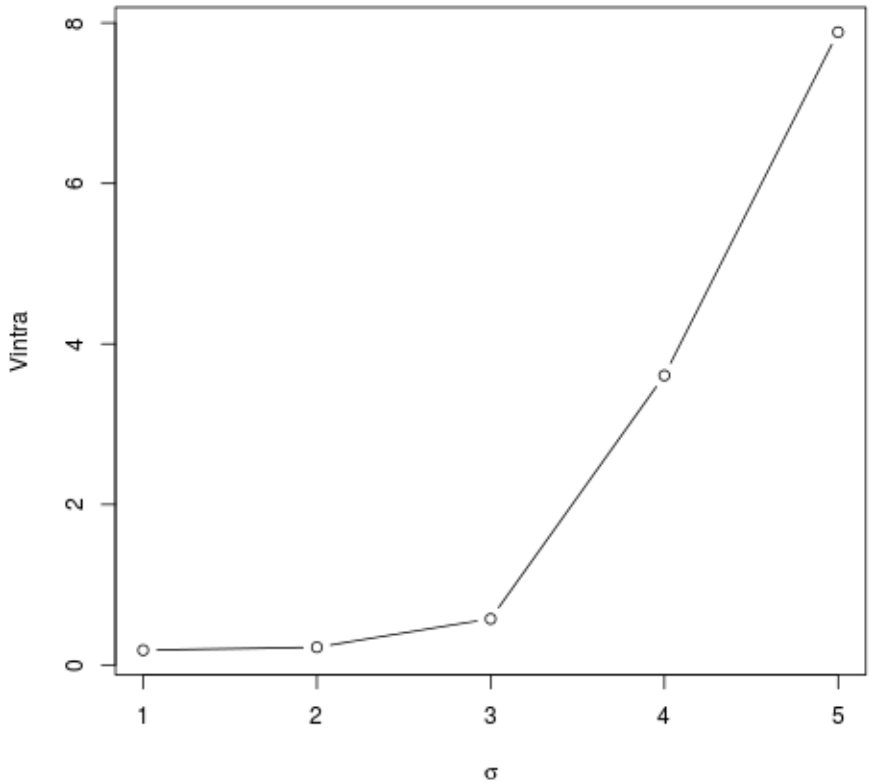
Courbe Précision/Rappel, DiOGenes, CID1, 20% d'individus manquants



Choix σ + distribution des arêtes

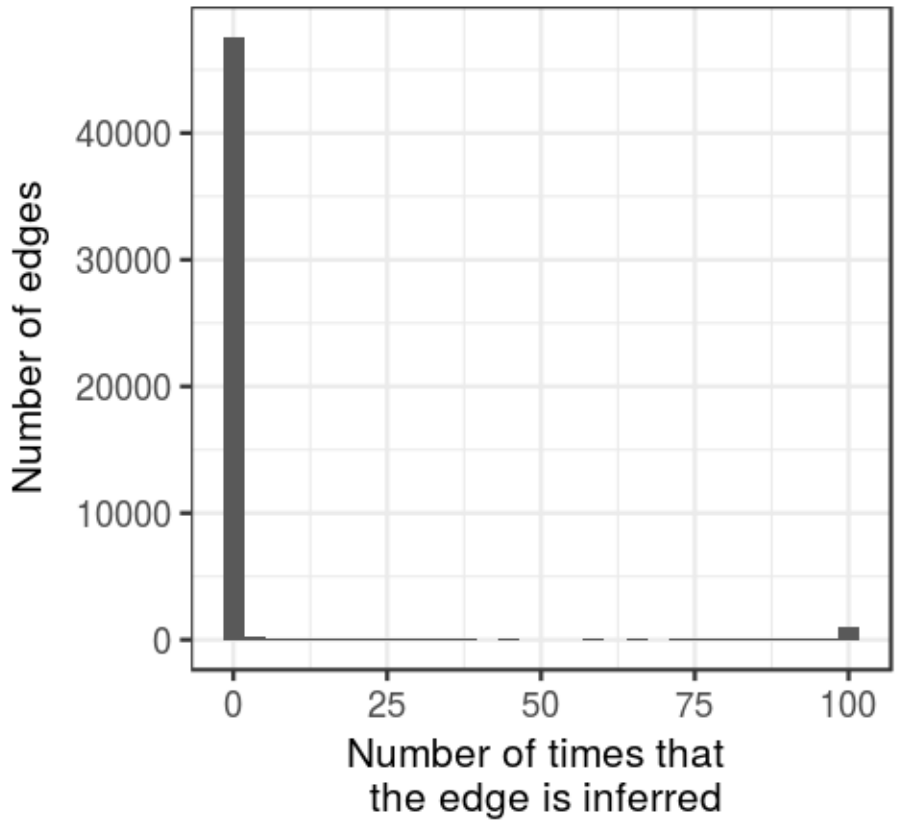
Illustration avec DiOGenes, CID1

Choix de la valeur de σ :



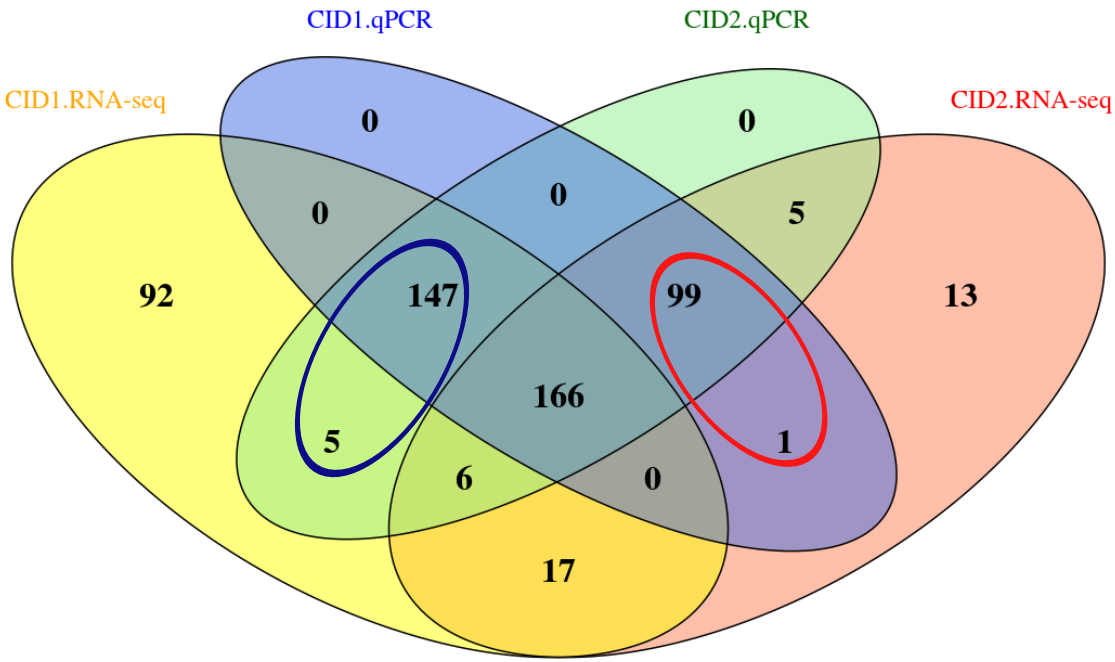
Choix $\sigma = 3$

Distribution d'une arête dans les M (100) réseaux inférés à partir des jeux imputés



Application Hd-MI, DiOGenes

Nombre d'individus à imputer pour chaque pas de temps



Nombre d'individus par CID (RNA-Seq/RT-qPCR)

Nombre d'individus imputés par **hd-MI** dans le cas de DiOgenes

- CID1 : 100 (♯ individus manquants en RNA-Seq à CID21 mais présent en RT-qPCR à CID1 et en RNA-Seq en CID2),
- CID2 : 152 (manquants en RNA-Seq CID2, présents en RT-qPCR à CID2 et RNA-seq CID1).