Inférence de réseaux de gènes à partir de données de séquençage haut-débit RNA-seq

Mélina Gallopin, Andréa Rau, Gilles Celeux, Florence Jaffrézic

Séminaire NETBIO

Lundi 19 novembre 2012

Paris

Questions

Inférence de réseaux de gènes: contexte

- *Données microarrays* ($\sim$ **1995**) Méthodes bien développées
- *Données de comptage RNA-seq* (**2008**) Peu de méthodes existent

Objectifs

1. Elaboration d'un modèle adapté à l'inférence à partir de données de comptage RNA-seq
2. Comparaison des différentes méthodes d'inférence de réseaux à partir de données RNA-seq

Faut-il privilégier:

- les modèles habituels appliqués aux données transformées?
- des modèles adaptés ne nécessitant pas de transformation préalable des données?

Données utilisées

$\Rightarrow$ Matrice des données $\mathbf{y}$ de dimension $(n \times p)$, n échantillons, p gènes

$\Rightarrow$ i indice les échantillons (en ligne)

$\Rightarrow$ j indice les gènes (en colonne)

	gène0	gène1	gène2	gène3	gène4	gène5	gène6	gène7	gène8
E1	2250	120	12394	1170	31	1313	692	857	160
E2	2406	146	13387	1271	37	1485	665	1074	170
E3	1937	130	10977	935	30	1079	463	833	123
E4	1781	126	9464	833	20	1022	388	970	190
E5	3344	439	15404	2303	46	2585	1387	2041	339
E6	2531	185	11749	1364	32	1518	680	1349	231
E7	3332	345	16680	2555	44	2600	1234	1930	317
E8	1883	115	11555	961	19	970	337	1079	143
E9	3572	402	16850	2940	50	2584	1213	2065	392
E10	1989	147	9540	1156	29	1169	400	1196	246
E11	3269	385	12173	2079	41	2515	1054	1900	355

- Données de comptage
  discrètes et positives
- Variabilité inter-échantillons importante
  variance empirique $\geq$ moyenne empirique
- $p >> n$

Méthodes d'inférence utilisées

1. **Modèle graphique gaussien** (Friedman, Hastie et Tibshirani 2007)
   après transformation des données $\mathbf{y} \rightarrow \log(\mathbf{y} + 1)$
2. **Modèle de Poisson log-linéaire** (Allen et Liu 2012)
   après transformation des données $\mathbf{y} \rightarrow \mathbf{y}^{\alpha}$ pour un $\alpha \in]0; 1]$
3. **Modèle hiérarchique Poisson log-normale**
   aucune transformation des données nécessaire

Le modèle graphique gaussien

Le modèle graphique gaussien: Définition

$\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})$ vecteur d'expression des p gènes pour un échantillon i modélisé par le vecteur aléatoire $\mathbf{Y}_i$.

Propriétés

Hypothèse $\mathbf{Y}_i \sim \mathbf{N}_p(\mu, \mathbf{\Sigma})$

Arcs du réseaux $\Leftrightarrow$ Coefficients non nuls de la matrice Σ^{-1}

Matrice de variance-covariance empirique

$$\mathbf{S} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y}_i - \bar{y})(\mathbf{y}_i - \bar{y})^T$$

Log-vraisemblance pénalisée du modèle

$$l(\mathbf{\Sigma}^{-1})_{penL_1} = log[det(\mathbf{\Sigma}^{-1})] - trace(\mathbf{S}\mathbf{\Sigma}^{-1}) - \lambda \|\mathbf{\Sigma}^{-1}\|_1$$

Le modèle graphique gaussien: Estimation et Sélection

Estimation Estimation de la matrice $\mathbf{\Sigma}^{-1}$ par l'algorithme glasso (Friedman, Hastie et Tibshirani 2007)

Sélection Choix du paramètre de régularisation λ par validation croisée

Adaptation aux données RNA-seq

- Données RNA-seq = données de comptage
- Transformation préalable nécessaire: $\mathbf{y} \rightarrow \log(\mathbf{y} + 1)$

Le modèle graphique de Poisson log-linéaire (Allen et Liu, 2012)

Le modèle graphique de Poisson

- Modèle bien adapté aux données discrètes (Allen et Liu, 2012)

$\mathbf{y}_j = (y_{1j}, \ldots, y_{nj})$ vecteur d'expression du gène j pour les n échantillons.
$\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})$ vecteur d'expression des p gènes pour un échantillon i

Définition

Hyp1 On modélise ces p valeurs d'expression par une loi de Poisson multivariée $\mathcal{P}(\mu_1), \ldots, \mathcal{P}(\mu_p)$.

Hyp2 On suppose: $p(\mathbf{y}_j | \mathbf{y}_{-j} = \mathbf{y}_{-j}) \sim \mathcal{P}(\mu_j)$ avec $\log(\mu_j) = \sum_{j' \neq j} \beta_{jj'} \tilde{\mathbf{y}}_{j'}$

Hyp3 $\mathbf{Y}_j \perp \mathbf{Y}_{j'} | \mathbf{Y}_{-(j \cup j')} \Leftrightarrow (\beta_{jj'} = 0 \text{ et } \beta_{j'j} = 0)$

- Méthode locale:
  une régression par gène $\Leftrightarrow$ inférence des voisins de ce gène

Le modèle de Poisson log-linéaire: Estimation

Estimation pour un gène

Log-vraisemblance pénalisée pour le gène j

$$\mathbf{y}_j \sum_{j' \neq j} \beta_{jj'} \tilde{\mathbf{y}}_{j'} - \exp \sum_{j' \neq j} \beta_{jj'} \tilde{\mathbf{y}}_{j'} - \lambda \|\boldsymbol{\beta}\|_{\ell_1}$$

Estimation de $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{j(j-1)}, \beta_{j(j+1)}, \ldots, \beta_{jp})$
par l'algorithme de *coordinate descent*
$\Rightarrow$ package R *glmnet* (Friedman, Hastie et Tibshirani, 2007)

Adaptation au cas de données à forte variance

Données RNA-seq réelles: Pour chaque gène,
variance empirique $\geq$ moyenne empirique

Modélisation de Poisson: Si $\mathbf{Y}_j \sim \mathcal{P}(\mu)$ alors $E(\mathbf{Y}) = Var(\mathbf{Y}) = \mu$

Solutions proposées

1. Transformation $\mathbf{y} \rightarrow \mathbf{y}^\alpha$ avec $\alpha \in]0; 1]$ maximisant le critère d'adéquation des données transformées $\mathbf{y}^\alpha$ avec la loi Poisson
   - proposée par Witten et Liu (2011)
   - implémentée dans le package *PoiClaClu*
2. Adaptation du modèle log-linéaire de Poisson pour prendre en compte la forte dispersion $\Rightarrow$ modèle de Poisson hiérarchique log-normale

Le modèle graphique hiérarchique Poisson log-normale

Modèle de Poisson hiérarchique log-normale: Définition

Idée: modéliser la *sur-dispersion* des données dans le modèle

Stratégie: Remplacer la loi de Poisson par une loi de Poisson hiérarchique log-normale

Modèle: $y_{i1}, \ldots, y_{ip} \sim \mathcal{P}(\mu_{i1}), \ldots, \mathcal{P}(\mu_{ip})$
$\log(\mu_{ij}) = \sum_{j' \neq j} \beta_{jj'} \tilde{y}_{ij'} + \epsilon_{ij}$ avec $\boldsymbol{\epsilon}_j \sim \mathbf{N}_n(0, \sigma_j^2 I_n)$

- Paramètres à estimer pour le gène j
  $(\boldsymbol{\beta}_j, \sigma_j^2) = (\beta_{j1}, \ldots, \beta_{j(j-1)}, \beta_{j(j+1)}, \ldots, \beta_{jp})$
- Même hypothèse d'indépendance conditionnelle
  $\mathbf{Y}_j \perp \mathbf{Y}_{j'} | \mathbf{Y}_{-(j \cup j')} \Leftrightarrow (\beta_{jj'} = 0 \text{ et } \beta_{j'j} = 0)$

Modèle de Poisson hiérarchique log-normale: Estimation

- Vraisemblance du modèle pour le gène j

$$\int_{\mathbb{R}} \{\prod_{i=1}^{n} [\exp\left(-\mu_{ij} + y_{ij} \log\left(\mu_{ij}\right) - \log\left(y_{ij}!\right)\right] \frac{1}{(2\pi)^{n/2}} \exp(-\frac{1}{2}||\epsilon||_2^2)\} d\epsilon$$

- Ajout de la pénalité $Q_\lambda(\beta_j, \sigma_j^2) = -2L_{(\beta_j, \sigma_j^2)}(y_{1j}, \ldots, y_{nj}) + \lambda||\beta_j||_1$

Solution au problème d'estimation des paramètres de (β_j, σ_j^2)

Glmmlasso: An Algorithm for High-Dimensional Generalized Linear Mixed Models Using ℓ_1-Penalization, Schelldorfer et Buhlmann (2011)

Etape 1 Approximation de Laplace pour obtenir une expression analytique de la log-vraisemblance

Etape 2 Approximation quadratique de la dérivée seconde de la fonction obtenue à l'étape 1 puis algorithme de *coordinate descent* pour estimer les paramètres

Modèle de Poisson hiérarchique log-normale: Sélection

Rappel: Méthode d'inférence locale

- Pour un gène $\Rightarrow$ une régression pour inférer les voisins du gène
- Nécessite autant de régressions de Poisson qu'il y a de gènes dans le réseau

Sélection du paramètre de régularisation λ en deux étapes

1) Pour la régression du gène j

Sélection de λ_j le critère BIC: $\lambda_j = \arg\max BIC$

avec $BIC = -2_{(\beta_j, \sigma_j^2)} + \log(n)[card(\beta_{jk} \neq 0) + 1]$

2) Pour l'ensemble des régressions

$$\lambda_{optimal} = \frac{\sum_{j=1}^{p} \lambda_j}{p}$$

Simulation de données de Poisson multivariées

Performance des modèles

Simulation d'une loi de Poisson multivariée

- Méthode décrite par Karlis dans *Multivariate Poisson Regression with covariance structure*

Simulation d'une matrice de données $(n \times 3)$

Etape 1: Simulation de $p + \binom{p}{2} = 6$ variables de Poisson indépendantes

$$(X_1, X_2, X_3, X_{12}, X_{13}, X_{23}) \sim (\mathcal{P}(\mu_1), \mathcal{P}(\mu_2), \mathcal{P}(\mu_3), \mathcal{P}(\mu_{12}), \mathcal{P}(\mu_{13}), \mathcal{P}(\mu_{23}))$$

Etape 2: Sommation des variables

$$Y_1 = X_1 + X_{12} + X_{13}$$
$$Y_2 = X_2 + X_{12} + X_{23}$$
$$Y_3 = X_3 + X_{13} + X_{23}.$$

$$\begin{pmatrix} \mu_1 + \mu_{12} + \mu_{13} & \mu_{12} & \mu_{13} \\ \mu_{12} & \mu_2 + \mu_{12} + \mu_{23} & \mu_{23} \\ \mu_{13} & \mu_{23} & \mu_3 + \mu_{13} + \mu_{23} \end{pmatrix}$$

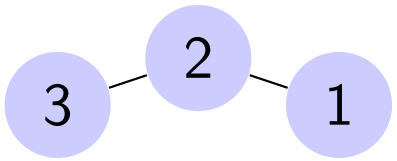
Variance-covariance du vecteur $\mathbf{Y} = (Y_1, Y_2, Y_3)$

Cas général

On simule $\mathbf{X}$ matrice de dimension $n \times (p + \frac{p \times (p-1)}{2})$
contenant en colonne les variables de Poisson indépendantes
On construit $\mathbf{B}$ matrice encodant la structure du graphe simulé
$\mathbf{B} = [\mathbf{I}_{(p)}; \mathbf{P} \odot (\mathbf{I}_{(p)} tri(\mathbf{A})^T)]^T$
$\mathbf{A}$ matrice d'adjacence du réseau
$\mathbf{P}$ matrice de permutation du vecteur $(1, 1, 0, \ldots, 0)$
$\mathbf{I}_{(p)}$ matrice identité de dimension p
$\odot$ produit d'Hadamard

$\Rightarrow \mathbf{Y} = \mathbf{XB}$ suit une loi de Poisson multivariée

$$[\mathbf{I}_{(p)}; \mathbf{P}]^T = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$



$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Simulation de la variabilité inter-échantillons

Pour augmenter la variabilité inter-échantillons des données simulées, on simule les variables $(X_1, \ldots, X_p)$ de $\mathbf{X}$ (définie plus haut) selon des lois de Poisson hiérarchique log-normale:

$X_{ij} \sim \mathcal{P}(\mu_{ij})$ avec $\log(\mu_j) = \theta_j + \epsilon_{ij}$ et $\boldsymbol{\epsilon}_j \sim N_n(0, \sigma_j^2)$

1. On tire d'abord un échantillon $\boldsymbol{\epsilon}_j$ de taille n issu d'une loi N(0,σ_j^2) avec $\sigma_j^2 = 1$
2. On simule n variables de Poisson de moyennes respectives $\exp(\theta_j + \epsilon_{ij})$ avec $\theta_j = 1$

Performance de l'inférence

Inférer un réseau $\Leftrightarrow$ Classifieur binaire (présence/absence d'arcs)

$$\Rightarrow \textit{Spécificité} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$$

$$\Rightarrow \textit{Sensibilité} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

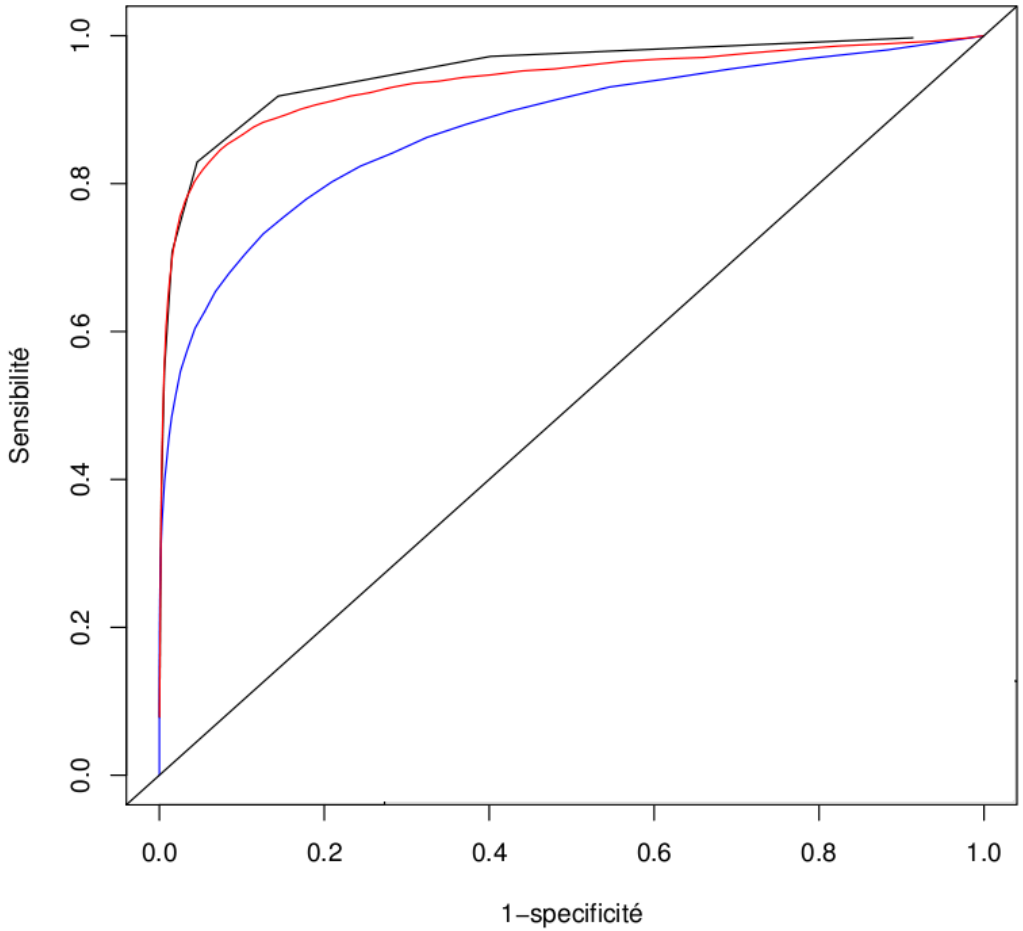
- Nous simulons 90 jeux de données différents à partir de loi de Poisson indépendantes de moyenne 1, comportant $p = 50$ variables, $n = 100$ échantillons.
- Le réseau simulé a une structure *scale-free*.

Résultats sur des données simulées

Modèle graphique gaussien sur données transformées $\mathbf{y} \to \log(\mathbf{y} + 1)$

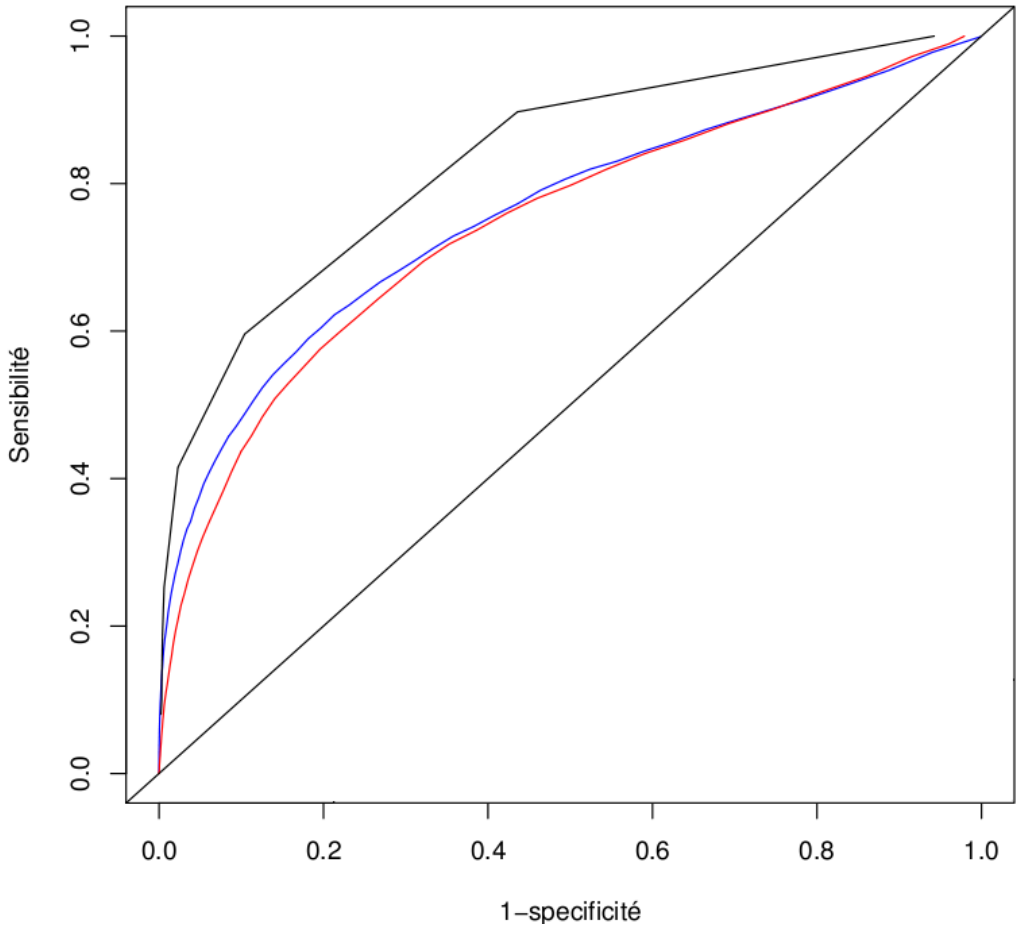
Modèle de Poisson log-linéaire

Modèle hiérarchique Poisson log-normale



Résultats sur des données simulées avec grande dispersion

Modèle graphique gaussien sur données transformées $\mathbf{y} \to \log(\mathbf{y} + 1)$
Modèle de Poisson log-linéaire sur données transformées $\mathbf{y} \to \mathbf{y}^{\alpha}$
Modèle hiérarchique Poisson log-normale



Sélection de modèle

- On effectue la sélection de graphe pour un jeu de données simulés de 50 gènes en sommant des variables de Poisson indépendantes de moyenne 1, avec ajout de surdispersion

		Modèle gaussien $log(données + 1)$	Modèle Poisson $(données)^{\alpha}$	Modèle hiérarchique *données non transformées*
$n = 100$	Sens	0,63	0,63	0,71
	Spéc	0,87	0,90	0,91
$n = 50$	Sens	0,28	0,26	0,38
	Spéc	0,90	0,88	0,98
$n = 10$	Sens	0,12	0,08	0,06
	Spéc	0,88	0,90	0,99

- NB: Sensibilité très faible pour une taille d'échantillon réduite

Application aux données réelles

Application aux données réelles

Les données disponibles au Département de Génétique Animale de l'INRA ne comportaient pas assez de réplicats biologiques ($n \leq 5$).

Jeu de données RNA-seq *Bottomly*, base de données *ReCount*

- 21 souris: lignée C57BL/6J (10 souris) et DBA/2J (11 souris) utilisées pour la recherche en neuroscience
- sélection de 10 gènes parmi les 50 gènes les plus différentiellement exprimés pour la lignée DBA/2J

Modélisation des données

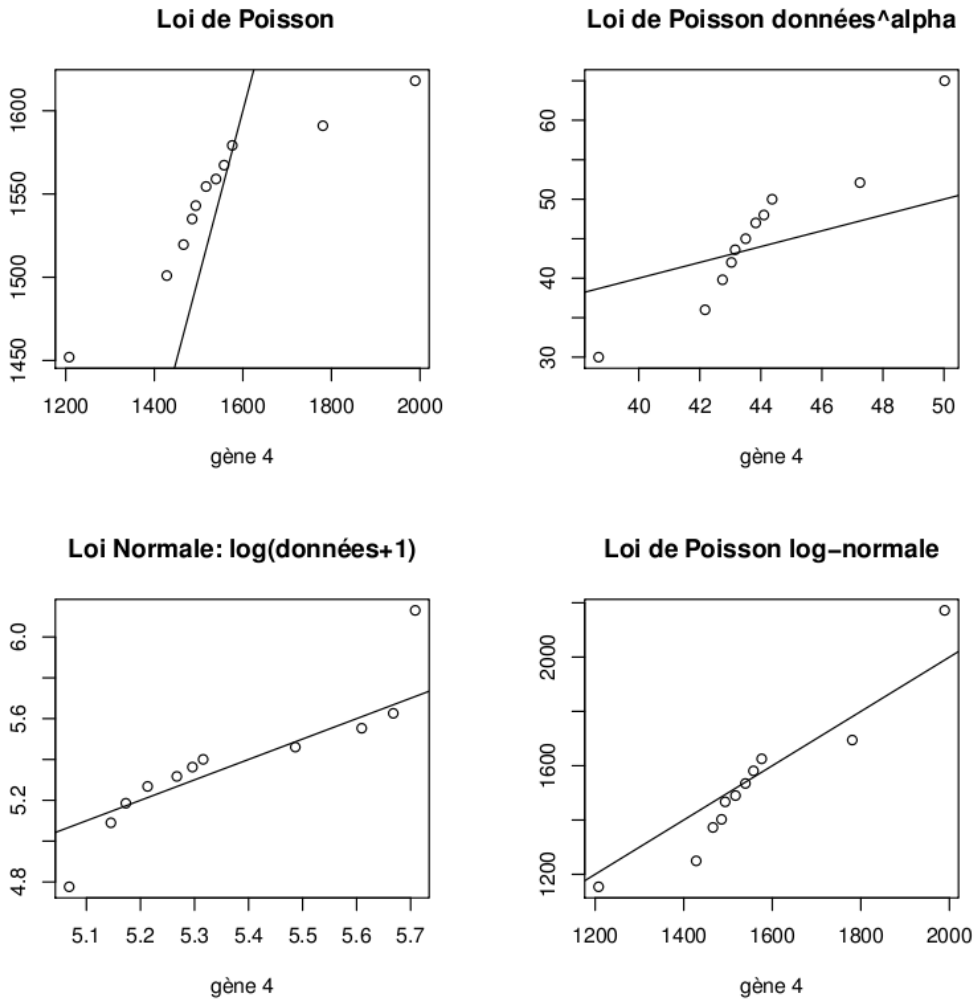
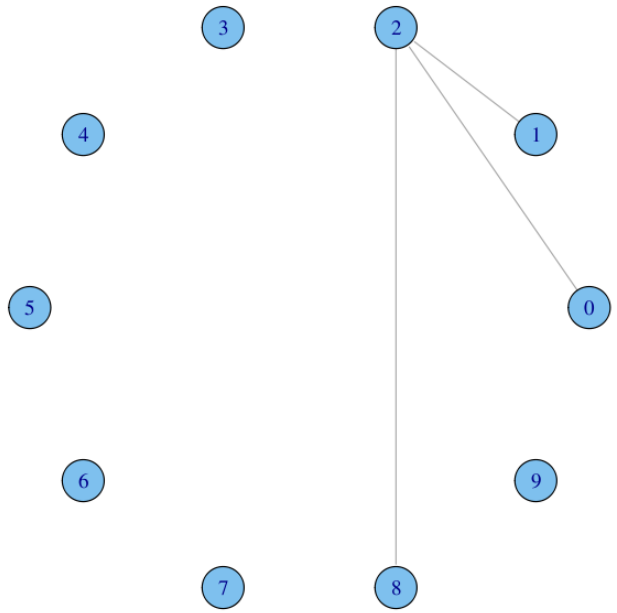


Figure: qqplots pour le gène 4

Réseau inféré par le modèlé log-linéaire

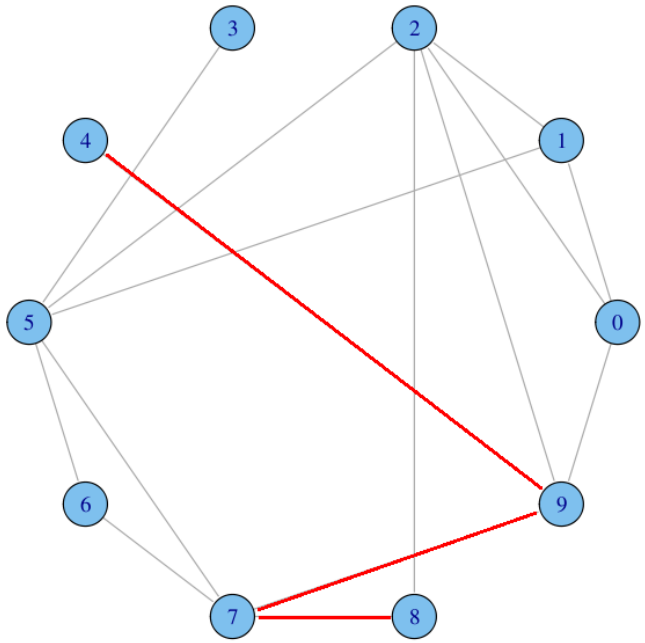
Modèle graphique log-linéaire de Poisson sur données transformées $\mathbf{y} \rightarrow \mathbf{y}^{\alpha}$



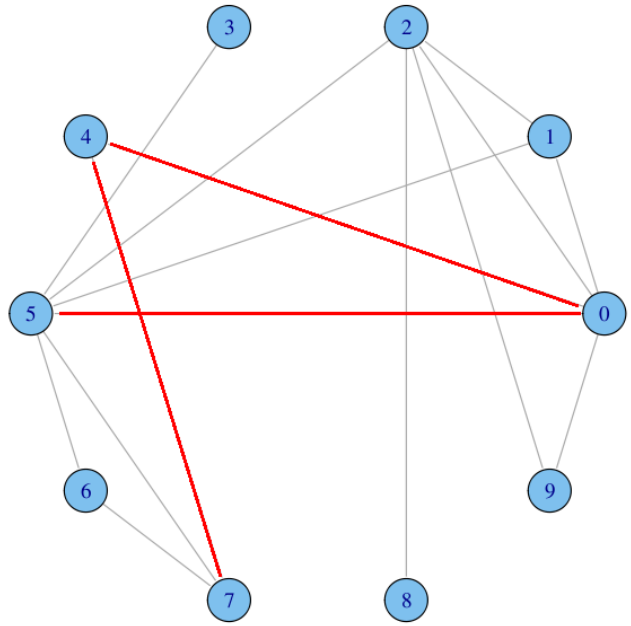
$=>$ Modèle prenant mal en compte la surdispersion

Réseau inféré

1. Modèle graphique gaussien sur données transformées $\mathbf{y} \to \log(\mathbf{y} + 1)$
2. Modèle graphique hiérarchique Poisson log-normale
   - Bootstrap sur les arcs: sur 100 ré-échantillonnages avec remise, on ne garde que les arcs inférés au moins 80 fois.



Modèle graphique gaussien



Modèle hiérarchique poisson log-normale

Discussion

1. Le modèle graphique hiérarchique Poisson log-normale est une bonne alternative pour prendre en compte à la fois le caractère discret et la grande variabilité inter-échantillons des données RNA-seq.
2. Sur les simulations, les méthodes n'étaient pas assez performantes pour des tailles d'échantillons trop petites $=>$ limitation pratique importante pour pouvoir inférer un réseau biologique d'au moins 50 gènes
3. Pour pouvoir inclure plus de gènes dans le réseau inféré:
   - inférer un réseau conjointement pour les deux conditions
     ex: *joint graphical model*
   - réduire le nombre de paramètres à inférer

Merci