Modèle bayésien hiérarchique pour l'analyse différentielle des données RNASeq

Florence Jaffrézic et Jean-Louis Foulley

INRA, Jouy-en-Josas (France)

Applibugs, 9 Décembre 2011

Plan

Plan

Introduction

Données d'expression de gènes (transcriptomique).

Mesure de la quantité d'ARN messager de milliers de gènes simultanément.

Détection de gènes différentiellement exprimés.

Enjeu biologique très important.

Enjeu statistique : Très grand nombre de gènes (plusieurs milliers), très peu de réplicats biologiques (5 à 10).

Introduction

Plusieurs méthodes ont été proposées pour l'analyse des données de microarrays (puces à ADN)
(Smyth, 2004 ; Delmar et al., 2005 ; Jaffrézic et al., 2007).

Tests de Student avec paramètres de variance "shrinkés" entre les gènes.

Hypothèse : Données continues Gaussiennes après transformation log2.

Introduction

Nouvelles technologies de séquençage haut-débit.

Plus grande densité de couverture du génome.

Possibilité de découvrir de nouveaux gènes.

Problème statistique : données de comptage, discrètes.

Développement de nouvelles méthodes statistiques.

Introduction

Analyse de données de séquençage : Comptages.

$\Rightarrow$ Modèles de Poisson sur-dispersé ou binomiale négative.

Etude de comparaison de modèles (Hardcastle et Kelly, 2010).
Modèle binomiale négative plus adapté aux données RNASeq.

Introduction

Plusieurs méthodes déjà disponibles dans R :
DESeq, edgeR, baySeq.

Toutes les méthodes proposées reposent sur des estimations approchées des paramètres.

Méthode proposée : Modèle hiérarchique et inférence bayésienne.

Plan

Modèle hiérarchique

y_{ijk} : comptage observé pour la séquence i, le réplicat biologique j et la condition k.

Level (0) :

$$\begin{aligned} y_{ijk} &\sim \mathcal{NB}(p_{ijk}, r_k) \\ p_{ijk} &= r_k/(r_k + \eta_{ijk}) \\ \eta_{ijk} &= M_{jk}\mu_{ik}/M \\ \ln \mu_{ik} &= m_i + \frac{(-1)^k}{2}\delta_i \end{aligned}$$

Pour deux conditions $k = (0, 1)$.
M_{jk} : taille de librairie pour l'échantillon j dans la condition k.
M : plus petite taille de librairie parmi tous les échantillons.

Modèle hiérarchique

Level (0) :

$$
\begin{aligned}
y_{ijk} &\sim \mathcal{NB}(p_{ijk}, r_k) \\
p_{ijk} &= r_k/(r_k + \eta_{ijk}) \\
\eta_{ijk} &= M_{jk}\mu_{ik}/M \\
\ln \mu_{ik} &= m_i + \frac{(-1)^k}{2}\delta_i
\end{aligned}
$$

Level (1) :

$$
\begin{aligned}
r_k &\sim Exp(1) \\
m_i &\sim \mathcal{N}(m, \tau^2) \\
\delta_i &\sim Mixture
\end{aligned}
$$

Modèle hiérarchique

Modèle de mélange à trois composantes pour la variable δ_i :

- séquences non différentiellement exprimées.

- sur ou sous-exprimées dans la condition 1 par rapport à la condition 2.

$$\delta_i = p_0\mathcal{N}(0, \tau_0^2) + p_1\mathcal{N}_{<0}(m_1, \tau_{diff}^2) + p_2\mathcal{N}_{>0}(m_2, \tau_{diff}^2)$$

Modèle hiérarchique

$$\delta_i = p_0\mathcal{N}(0, \tau_0^2) + p_1\mathcal{N}_{<0}(m_1, \tau_{diff}^2) + p_2\mathcal{N}_{>0}(m_2, \tau_{diff}^2)$$

Level (2) : $\pi(m) = 1,\ m1 \sim U(-10, 0),\ m_2 \sim U(0, 10),$
$\tau \sim U(0, A),\ \tau_{diff} \sim U(0, B),\ \tau_0^2 = \tau_{diff}^2/k$

$A = B = 5$, and $k = 100$.

Distribution de Dirichlet pour le vecteur de probabilité $p = (p_0, p_1, p_2)$.

Modèle hiérarchique

Etude de sensibilité au choix des a priori.

Listes de gènes différentiellement exprimés assez similaires avec un prior uniforme $U(0,5)$ ou log-normal sur les écarts-types $log\tau \sim \mathcal{N}(0, 0.1)$ et $log\tau_{diff} \sim \mathcal{N}(0, 0.1)$.

Modèle hiérarchique

Choix des paramètres de la distribution de Dirichlet pour le vecteur de probabilité $P = (p0, p1, p2)$.

$\mathbf{P} \sim \mathit{Dirichlet}(\boldsymbol{\alpha})$, avec $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)$.

$E(p_k) = \pi_k = \alpha_k / \alpha_+$, avec $\alpha_+ = \sum_i \alpha_i$.

Résultats assez robustes au choix de la proportion $\pi_0 = E(p_0)$ (valeurs considérées $\pi_0 = 0.80, 0.90, 0.95$).

Modèle hiérarchique

On définit une variable indicatrice I_i telle que :

$I_i = 1$ si la séquence i est dans la deuxième ou troisième catégorie.

Sélection des séquences différentiellement exprimées par un facteur de Bayes local.

Modèle hiérarchique

$BF_i = post_i/prior_i$, où $post_i$ est l'odd a posteriori :

$$post_i = \frac{P(I_i = 1|\mathbf{y})}{1 - P(I_i = 1|\mathbf{y})} = \frac{E(I_i|\mathbf{y})}{1 - E(I_i|\mathbf{y})}$$

$E(I_i|\mathbf{y})$: moyenne a posteriori de la variable indicatrice.

Odd a priori : $prior_i = p_i/(1 - p_i)$
$p_i = P(I_i = 1)$: proportion a priori de séquences différentiellement exprimées.
$p_i = p_1 + p_2$: deux dernières composantes de la loi de Dirichlet.

Modèle hiérarchique

On transforme ces valeurs de facteur de Bayes local (BF) en deciban par $10 * log_{10}(BF)$ (Gautier et al., 2009).

Utilisation des seuils de décision proposés par Jeffreys (1961).

Estimation des paramètres du modèle par Winbugs.

Plan

Etude de simulation

Le modèle proposé a été comparé à trois méthodes existantes disponibles dans R :

- DESeq (Anders et Huber, 2010).
- baySeq (Hardcastle et Kelly, 2010).
- edgeR (Robinson et al., 2010) :

Etude de simulation

DESeq (Anders et Huber, 2010).

Régression locale pour relier la moyenne et la variance de la distribution binomiale négative.

Méthode d'estimation approchée.

Etude de simulation

baySeq (Hardcastle et Kelly, 2010).

Modèle binomiale négative.

Détermination empirique de la distribution a priori à partir des données .

Estimation bayésienne empirique.

Etude de simulation

edgeR (Robinson et al., 2010).

Modèle binomiale négative.

Estimation approchée des paramètres par vraisemblance conditionnelle pondérée.

1) Dispersion commune.
2) Estimation du paramètre de shrinkage.
3) Paramètre de shrinkage égal à 10.

Etude de simulation

1000 séquences sont simulées suivant une distribution binomiale négative.

100 sont supposées différentiellement exprimées entre les deux conditions.

Les tailles de librairie ℓ_j : échantillonnées dans une distribution uniforme entre 3000 et 9000.

Etude de simulation

Paramètres de dispersion : échantillonnés dans une distribution Gamma (shape=0.85 et scale=0.5).

Les séquences non différentiellement exprimées, sont simulées de moyenne $\lambda_i \ell_j$

λ_i échantillonnés aléatoirement parmi des moyennes empiriques calculées à partir d'un jeu de données réelles (Zhang et al, 1997), comme suggéré par Robinson et al. (2010).

Etude de simulation

Pour les séquences différentiellement exprimées (DE), les données dans la première condition sont simulées de moyenne $\lambda_i \ell_j / \sqrt{b}$ et pour la seconde de moyenne $\lambda_i \ell_j \sqrt{b}$.

Comme dans Hardcastle et Kelly (2010), deux valeurs sont considérées pour le paramètre b : $b = 8$ pour de grandes différences entre les deux conditions et $b = 4$ pour des différences plus faibles.

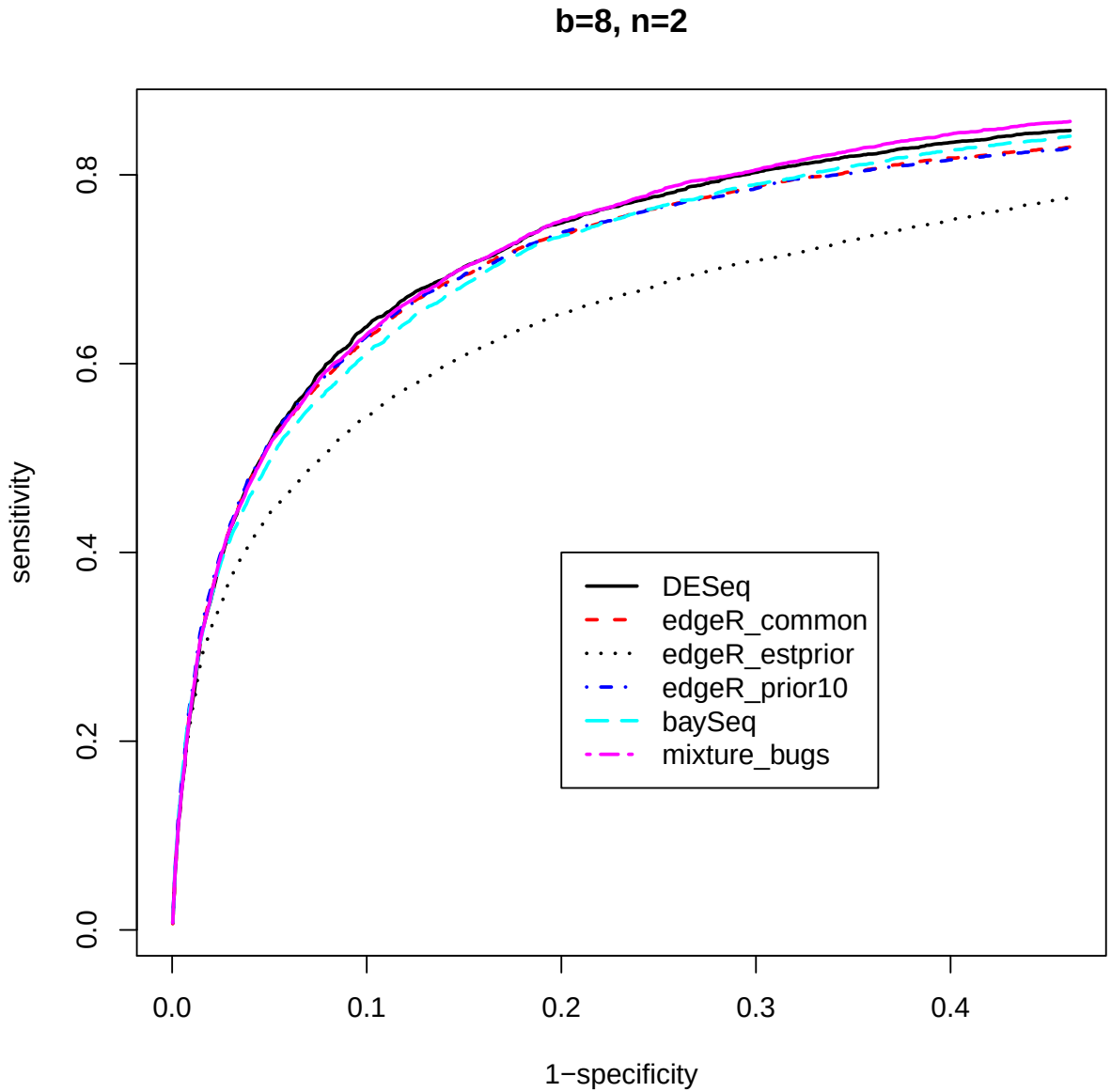
Etude de simulation

Plusieurs valeurs considérées pour les nombres d'échantillons biologiques par condition : 2, 5 et 10.

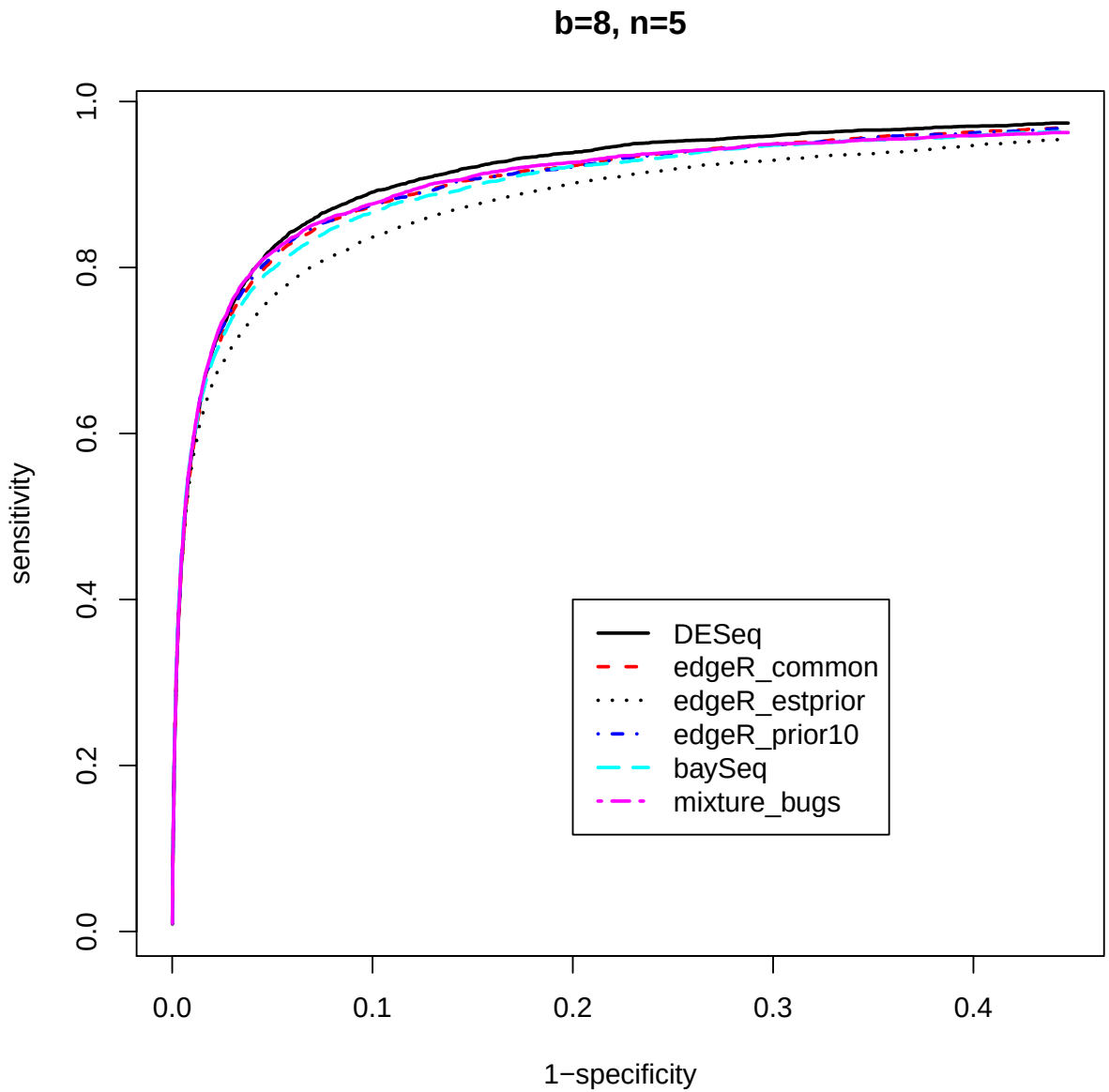
Les méthodes sont comparées sur 50 jeux de données simulés.

Plusieurs critères de comparaison sont utilisés : courbes ROC, nombres de vrais et faux positifs.

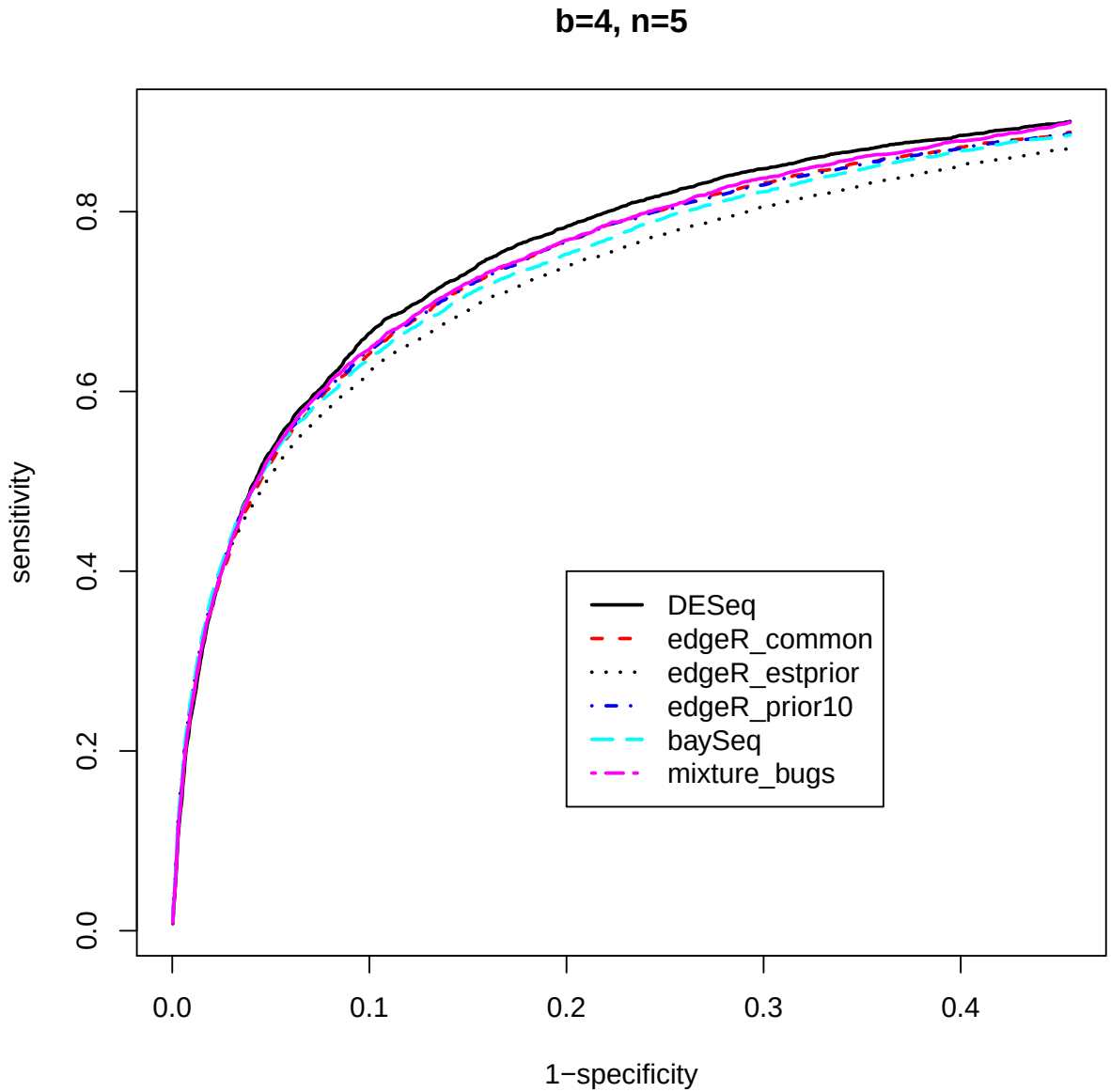
Etude de simulation



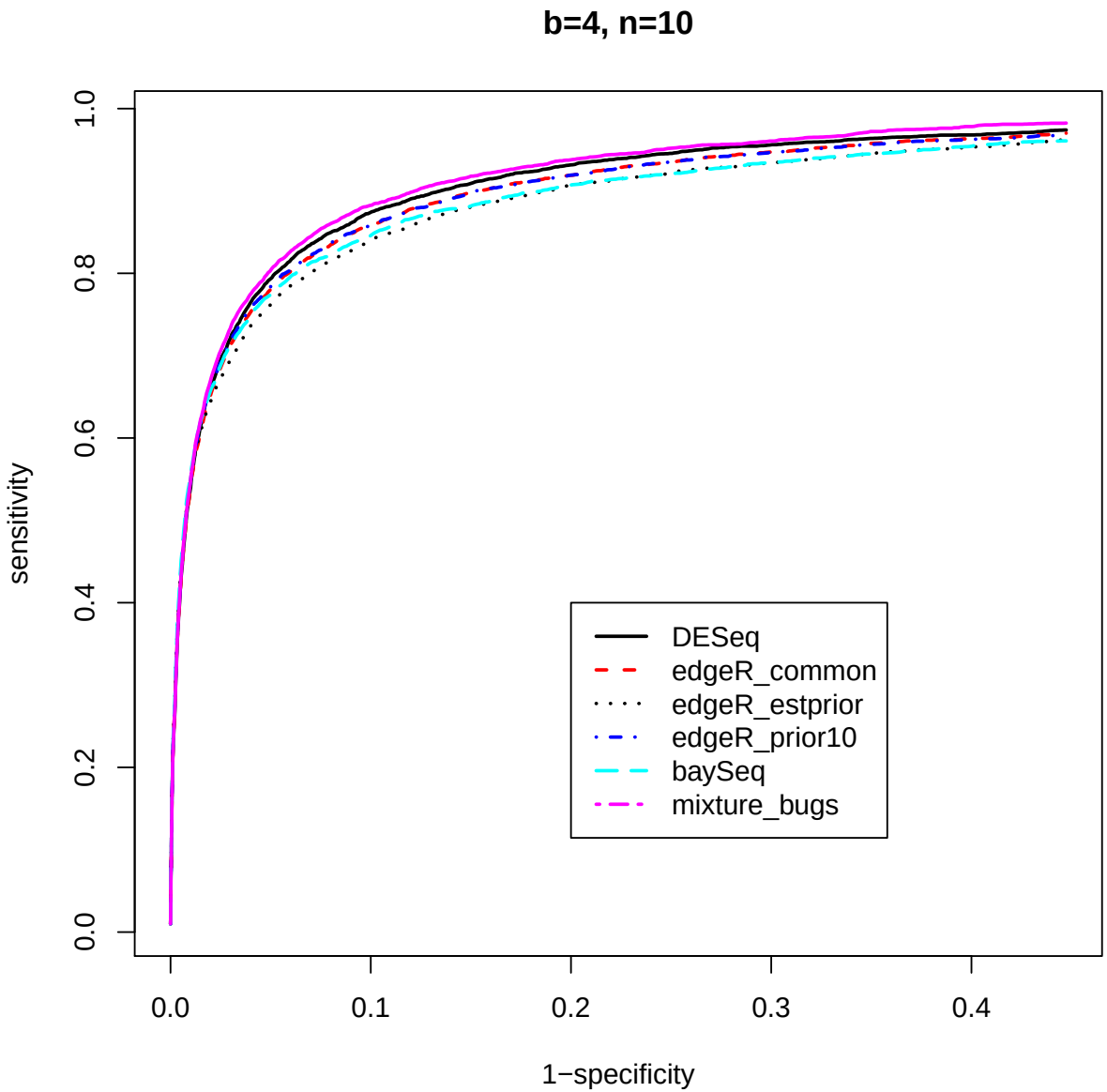
Etude de simulation



Etude de simulation



Etude de simulation



Etude de simulation

Nombre de vrais positifs, pour $b = 8$.

Nombre de réplicats	2	5	10
DESeq	5.1(4.0)	49.4(8.0)	86.2(4.1)
edgeR_common	1.7(2.2)	44.2(8.2)	83.8(4.1)
Mixture_bugs$^{(1)}$	17.5(6.0)	57.0(6.4)	85.4(3.7)

(1) : Seuil de BF = 15.

Etude de simulation

Nombre de faux positifs, pour $b = 8$

Nombre de réplicats	2	5	10
DESeq	1.1(1.3)	5.5(2.4)	8.2(3.7)
edgeR_common	0.4(0.8)	4.5(2.2)	8.1(3.5)
Mixture_bugs	5.6(3.0)	7.4(3.2)	5.1(2.9)

Etude de simulation

Nombre de vrais positifs, pour $b = 4$.

Nombre de réplicats	2	5	10
DESeq	0.8(1.2)	10.2(5.4)	46.3(6.8)
edgeR_common	0.4(0.7)	8.5(4.8)	44.8(6.5)
Mixture_bugs$^{(1)}$	1.8(2.3)	22.5(6.2)	53.1(6.0)

(1) : Seuil de BF = 15.

Etude de simulation

Nombre de faux positifs, pour $b = 4$

Nombre de réplicats	2	5	10
DESeq	0.8(1.3)	2.3(1.7)	6.1(2.9)
edgeR_common	0.5(0.9)	1.8(1.5)	5.4(3.0)
Mixture_bugs	1.3(1.9)	6.4(2.6)	7.4(3.4)

Plan

Analyse de données réelles

Données de miARN chez le cheval obtenues avec un séquenceur Solid
(X. Mata et L. Schibler, INRA-GABI).

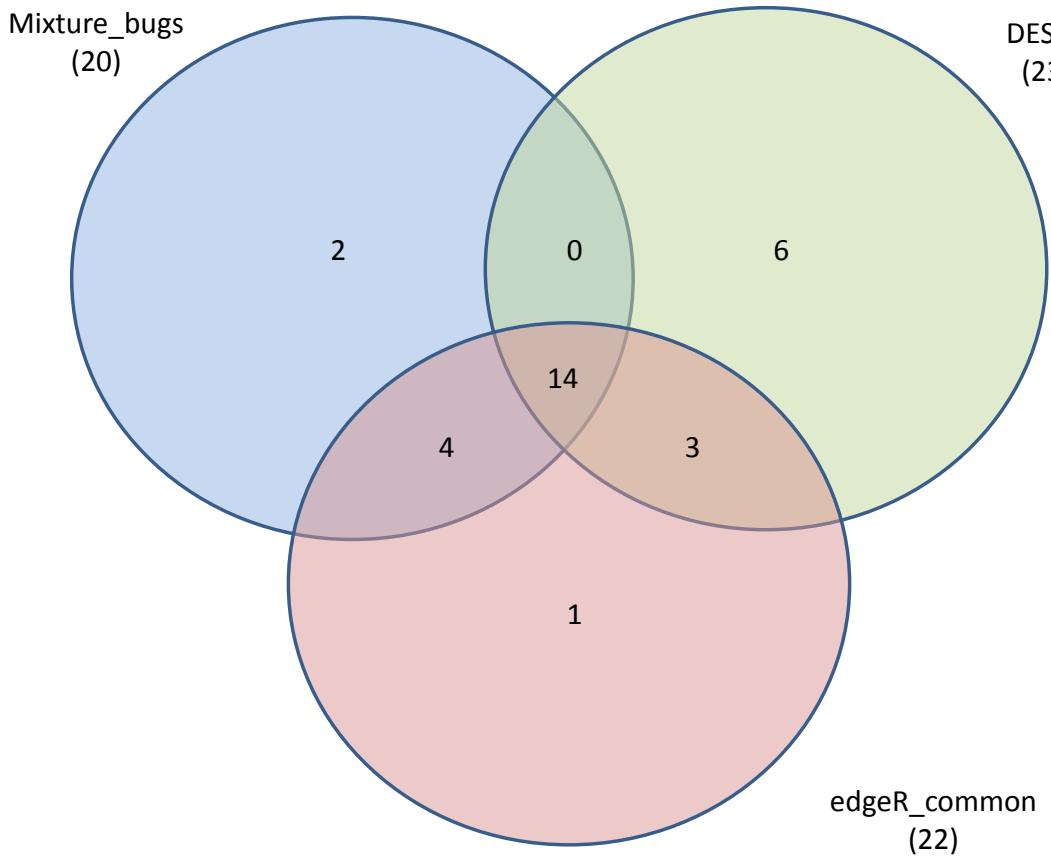
3 réplicats biologiques par condition, chacun correspondant à un pool de 3 animaux.

Au total 253 miARNs.

Plusieurs tissus sont comparés.

Ici comparaison des muscles peaucier-masséter.

Analyse de données réelles



Analyse de données réelles

Toutes les méthodes reposent sur des modèles binomiale négative.

Différentes procédures d'estimation et différents critères de sélection.

$\Rightarrow$ Seulement 14 miARN trouvés en commun
(parmi 20, 22 et 23).

Analyse de données réelles

Recherche bibliographique biologique.

Parmi ces 14 miARNs :

- 5 n'ont pas de rôle évident dans le muscle.
- 3 ont un rôle dans le muscle en général.
- 6 sont potentiellement biologiquement très intéressants.

Analyse de données réelles

Les miARNs trouvés par une seule méthode
DESeq : 6, Mixture_bugs : 2, edgeR : 1.

Ne sont pas connus pour avoir un rôle fonctionnel dans le muscle.

Potentiellement des faux positifs ?

Validation biologique plus approfondie de ces miARNs en cours.

Plan

Discussion

Etude de simulation :

Problème d'estimation du paramètre de shrinkage dans edgeR.

Recommandé d'utiliser le modèle avec dispersion commune.

Très faible puissance de détection pour $n = 2$ avec toutes les méthodes.

Modèle proposé semble plus performant que les autres méthodes pour un faible nombre de réplicats biologiques.

Discussion

Limites de l'approche bayésienne :

Temps de calcul assez longs.

Analyse de données réelles :

Différences assez importantes entre les listes de miARNs trouvées par les différentes méthodes.

Validation biologique de ces miARNs en cours.