Contrôle de l'erreur de type I et du taux de fausses découvertes dans l'analyse de données RNA-seq grâce à un test en composante de variance

Boris Hejblum [1] & Denis Agniel [2]

[1] *Université de Bordeaux, ISPED, INSERM U1219, INRIA SISTM*
Vaccine Research Institute
146 rue Léo Saignat, Bordeaux 33076, France
boris.hejblum@u-bordeaux.fr

[2] *RAND Corporation, Santa Monica, CA, USA*
Department of Biomedical Informatics, Harvard Medical School
Denis_Agniel@rand.org

Résumé. La technologie RNA-seq s'impose comme le nouveau standard pour la mesure de l'expression génique, et son utilisation est toujours plus importante, y compris dans des études cliniques. Il devient alors nécessaire d'adapter les outils statistiques employés pour leur analyse, puisque les données de séquençage se présentent comme des comptages. Il a été proposé de modéliser les comptages RNA-seq comme des variables continues en utilisant des régressions non-paramétriques pour modéliser leur hétéroscédasticité intrinsèque. Dans cet esprit, nous avons développé une méthode efficace pour identifier les transcrits différentiellement exprimés à partir de données RNA-seq. Grâce à un test en composante de variance, cette méthode permet d'identifier les transcrits dont le niveau d'expression est significativement associé à un facteur (ou un groupe de facteurs), conditionnellement à des covariables et sans supposer une quelconque forme paramétrique sur la distribution des comptages (transformés). Malgré la présence d'un estimateur non-paramétrique, notre statistique de test a une forme simple et suit une distribution asymptotique, tous deux pouvant être calculés rapidement. Nous proposons également un test de permutation pour palier au cas des petits échantillons. Ce test présente de bonne propriétés statistiques, illustrées grâce à des données simulées ainsi qu'à des données réelles. En particulier, il fait preuve d'une amélioration de la stabilité et de la puissance statistique comparé aux méthodes actuellement utilisées que sont voom/limma, edgeR, et DESeq2. De plus, nous montrons que ces méthodes échouent toutes les trois à contrôler l'erreur de type I ainsi que le taux de fausses découvertes dans des cas réalistes, tandis que notre méthode se comporte comme attendu. Cette méthode est implémentée dans le package `tcgsaseq` disponible sur le CRAN.

Mots-clés. Données RNA-seq, Hétéroscédasticité, Model mal spécifié, Taux de fausses découvertes, Test en composante de variance

Abstract. As gene expression measurement technology is shifting from microarrays to sequencing, the statistical tools available for their analysis must be adapted since RNA-seq data are measured as counts. It has been proposed to model RNA-seq counts as continuous variables using nonparametric regression to account for their inherent heteroscedasticity. In this vein, we developed a principled, model-free, and efficient method for detecting differentially expressed genes from RNA-seq data. The method can identify the genes whose expression is significantly associated with a factor or a group of factor, through a variance component score test, while accounting for both covariates and heteroscedasticity without assuming any specific parametric distribution for the (transformed) counts. Despite the presence of a nonparametric component, our test statistic has a simple form and limiting distribution, which can be computed quickly. A permutation version of the test is derived for small sample sizes. Applied to both simulated data and real benchmark datasets, we show that our test has very good statistical properties, with an increase in stability and power when compared to state-of-the-art methods voom/limma, edgeR, and DESeq2. In addition, we show that all three methods can fail to control the type I error and the False Discovery Rate under realistic settings while our method behaves as expected. We have made the method available for the community in the R package `tcgsaseq`.

Keywords. False Discovery Rate, Heteroscedasticity, Model Misspecification, RNA-seq data, Type I error, Variance component testing

1 Contexte

La technologie RNA-seq s'impose comme le nouveau standard pour la mesure de l'expression génique, et son utilisation est toujours plus importante, y compris dans des des études cliniques. Les méthodes les plus utilisées à l'heure actuelle sont voom/limma [1], edgeR [2, 3] et DESeq2 [4], toutes trois disponibles dans différents package R sur la plateforme Bioconductor. Ces méthodes s'appuient sur des hypothèses distributionnelles fortes des données RNA-seq (log-normal hétéroscédastique ou négative binomiale, respectivement). Néanmoins, nous mettons en évidence dans nos études de simulations que ces méthodes échouent à contrôler l'erreur de type I ou le taux de fausses découvertes dans des situations réalistes.

2 Méthodes

On considère $\mathbf{y}_i = (y_{i1}^\intercal, ..., y_{iP}^\intercal)^\intercal$ un vecteur de log-comptages par million pour l'échantillon i, associé à P transcrits, ainsi que le vecteur de $\mathbf{x}_i = (x_{i1}, ..., x_{iq})^\intercal$ de covariables décrivant les conditions d'expérience, et enfin au vecteur $\boldsymbol{\phi}_i = (\phi_{i1}, ..., \phi_{is})^\intercal$ des variables d'intérêt dont on veut tester l'association avec l'expression génique. On utilise le modèle de travail

suivant :

$$y_{ij} = \alpha_{0j} + \mathbf{x}_i^\top \boldsymbol{\alpha}_j + \boldsymbol{\phi}_i^\top \boldsymbol{\beta}_j + \boldsymbol{\phi}_i^\top \boldsymbol{\xi}_{ij} + \varepsilon_{ij}$$

où $\boldsymbol{\varepsilon}_j \sim N(0, \Sigma_{\boldsymbol{\varepsilon} j}), \boldsymbol{\xi}_i \sim N(0, \Sigma_{\boldsymbol{\xi} j}), \alpha_{0j}$ représente le niveau d'expression moyen spécifique du transcrit j, $\boldsymbol{\alpha}_j$ représente les effets des covariables sur le niveau d'expression du gene j, $\boldsymbol{\beta}_j$ représente les effets des variables d'intérêt sur le niveau d'expression du gene j tandis que $\boldsymbol{\xi}_{ij}$ représente les effets aléatoires des variables d'intérêt, et enfin $\Sigma_{\boldsymbol{\varepsilon} j}$ est la matrice de covariance des erreurs de mesure (et puisqu'indéxée sur j, cette dernière peut donc dépendre de la moyenne de $\mathbf{y}_j$). Nous supposons également que $\boldsymbol{\xi}_j \perp \boldsymbol{\varepsilon}_j$. Il est important de noter qu'en pratique, il est peu vraisemblable que ce modèle soit bien spécifié. Heureusement, il ne s'agit que d'un modèle de travail et la procédure de test que nous proposons est robuste à la mauvaise spécification de ce modèle.

Nous proposons de tester l'hypothèse nulle suivante :

H_0 : la moyenne d'expression $\mathbf{y}_j$ du transcrit j ne dépend pas des variables d'intérêt $\boldsymbol{\phi}_j$

En construisant le test en composante de variance correspondant, on obtient la statistique de test suivante pour chaque gène :

$$Q_j = \mathbf{q}_j^\top \mathbf{q}_j \qquad \text{où } \mathbf{q}_j^\top = n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{y}_{\mu ij}^\top \Sigma_{\boldsymbol{\varepsilon}_j}^{-1} \boldsymbol{\phi}_i \Sigma_{\boldsymbol{\xi} j}^{1/2}$$

où n est le nombre total d'échantillons, et $\mathbf{y}_{\mu ij} = \mathbf{y}_i - \alpha_{0j} - \mathbf{x}_i^\top \boldsymbol{\alpha}_j$. On montre que cette statistique suit asymptotiquement un mélange de distributions χ_1^2, dont les poids de mélange peuvent se calculer.

Un aspect important de notre approche est également l'estimation de $\Sigma_{\boldsymbol{\varepsilon} j}$ afin de tenir compte de l'hétéroscedasticité intrinsèque de $\mathbf{y}$. Pour cette estimation, on utilise l'ensemble des transcrits mesurés afin d'avoir un maximum d'information sur la relation entre la moyenne et la variance. Dans l'esprit de [1], on utilise des régressions linéaires locales. Ces poids peuvent être calculés soit au niveau du transcrit, soit individuellement pour chaque observation. Notons que la précision de cette estimation affectera la puissance de notre test, mais pas son contrôle de l'erreur de type I.

3 Résultats

Afin d'effectuer une étude de simulation nous avons d'abord généré des données d'expression génique synthétiques sous un modèle binomiale négative, conformément aux hypothèses distributionnelles des méthodes edgeR and DESeq2. Le paramètre de moyenne de la binomiale négative est donné par :

$$\mu_{ij} = \max\{1001 + a_{0i} + x_i + (b_i + \beta + \beta_j)\phi_i x_i, 0\}$$

où $a_{0i} \sim N(0,1), x_i \sim N(\mu_{xi},1), \mu_{xi} \sim \text{Exp}(1/10)$, $b_i \sim N(0,1)$, $\beta_j \sim N(0,1)$, tandis que le paramètre de dispersion de la binomiale negative est fixé à 1. Nous avons étudié à la fois le cas où ϕ_i est une variable binaire (représentant le cas fréquent de la comparaison entre deux groupes), et le cas où l'on compare les échantillons selon une variable d'intérêt quantitative. Dans ce dernier cas, le seule différence avec la modélisation décrite ci-dessus est : $\phi_i \sim N(0,1)$. Enfin, nous nous sommes également intéressé à un modèle largement mal-spécifié en générant les données de la façon suivante :

$$y_{ij} = \log\left\{\frac{(\mu_{ij}+0.5)10^6}{\sum_i \mu_{ij}+1}\right\}$$

$$\text{où} \quad \mu_{ij} = \eta_{ij}\sum_i \eta_{ij}/P + (b_{ij}+\beta)\phi_i, \qquad \eta_{ij} = a_{ij} + \alpha_j + \sum_{k=1}^{3} x_{ik} + \epsilon_{ij},$$

avec $\epsilon_{ij} \sim N(0,\alpha_j^2)$, $\alpha_j \sim \text{Exp}(100)$, $a_{ij} \sim N(0,\alpha_j^2/100)$, $x_{ik} \sim N(100,2500)$, $b_{ij} \sim N(0,1)$. À chaque fois, le paramètre β est laissé variable. Ceci permet d'évaluer l'erreur de type I ainsi que la puissance pour différentes tailles d'effet de manière comparable entre les méthodes (edgeR, DESeq2 et limma/voom testant l'hypothèse $H_1 : \beta \neq 0$).

Dans chacun des scénarios décrits ci-dessus, nous montrons que les méthodes edgeR, DESeq2 ou limma/voom rencontrent des problèmes pour contrôler l'erreur de type I tout comme le taux de fausses découvertes (y compris lorsque le modèle est bien spécifié). En revanche, notre procédure de test se comporte comme attendu, y compris lorsque le modèle est mal spécifié. Les figures 1 et 2 illustrent ces résultats.

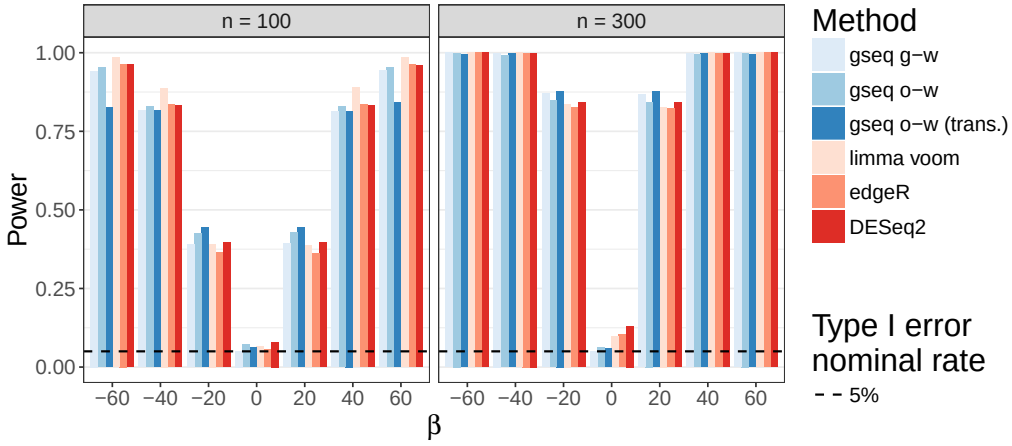

FIGURE 1 : Puissance et erreur de type I pour le modèle binomiale négative et variable d'intérêt continue. gseq désigne notre approche avec différents poids d'hétéroscédasticité.

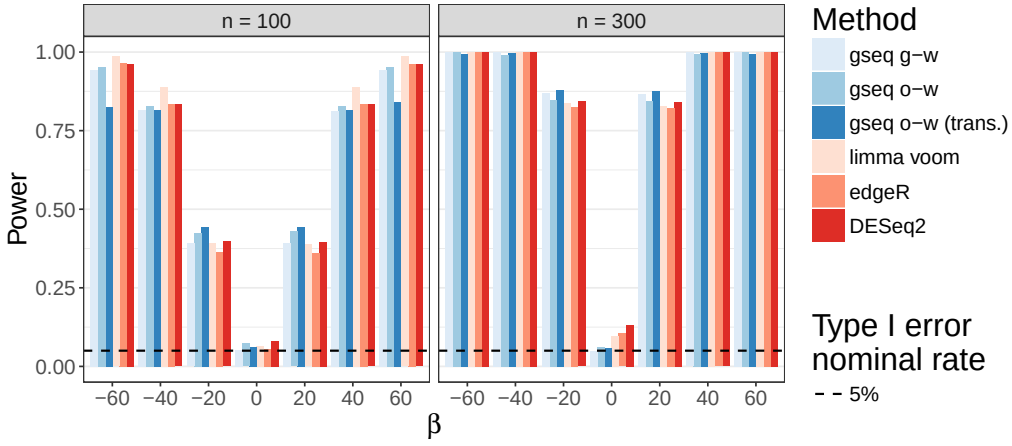

FIGURE 2 : Taux de fausses découvertes pour le modèle non linéaire binomiale-négative et variable d'intérêt continue. gseq désigne notre approche avec différents poids d'hétéroscédasticité.

Références

[1] Charity W Law, Yunshun Chen, et al. voom : Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2) :R29, 2014.

[2] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1) :139–140, 2010.

[3] Davis J. McCarthy, Yunshun Chen, and Gordon K. Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10) :4288–4297, 2012.

[4] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12) :550, 2014.