Influence du nombre de réplicats dans une analyse différentielle de données RNA-Seq

Sophie Lamarre [1], Stéphane Pyronnet [2], Emeline Sarot [2], Sébastien Déjean [3], Magali San Cristobal [3,4] & Matthieu Vignes [5]

[1] *Plateforme GeT-Biopuces, LISBP, INSA, Toulouse, France*
[2] *Inserm U 1037, CHU de Rangueil, Toulouse, France*
[3] *IMT et INSA Toulouse, France*
[4] *GenPhyse, INRA Toulouse, France*
[5] *MIA-T, INRA Toulouse, France*

Résumé. Cet article traite du nombre de réplicats biologiques dans une analyse différentielle de l'expression des gènes mesurée par la technique RNA-Seq. Nous montrons sur un jeu de données réelles que 4 réplicats semble un minimum pour garantir la précision et la robustesse des résultats et proposons une étude sur données simulées pour répondre à cette question en environnement contrôlé.

Mots-clés. expression génique, analyse différentielle, réplicat biologique

Abstract. This work deals with the question of the number of biological replicates in differential gene analysis as measured by RNA-Seq technique. The analysis of a real data set tends to show that 4 replicates is the bare minimum to achieve acceptable accuracy and robustness in the results. We also propose a simulated framework to better quantify the influence of different parameters to this answer.

Keywords. gene expression, differential analysis, biological replicate

1 Introduction

Le génome est l'ensemble du matériel génétique d'un individu. Le décrypter est un enjeu majeur pour la recherche scientifique. L'information qu'il contient est codée dans l'ADN grâce à un alphabet à 4 lettres (ou nucléotides) A, C, G et T. Elle est partagée par toutes les cellules d'un individu. Cependant, seules certaines parties s'expriment, c'est à dire sont lues et utilisées pour la production de molécules fonctionnelles (les protéines) selon le stade de développement et/ou l'environnement qui sont propres à chaque cellule. Des mécanismes de régulation contrôlent cette expression. Pour simplifier, nous focaliserons notre étude sur l'expression des gènes, les unités codant pour les protéines. Ainsi, dans différentes conditions, l'expression génique est différente pour un même individu. Des techniques statistiques standards existent pour la détection des gènes différentiellement exprimés (DE) dont l'expression est mesurée par la technologie des puces ADN (Dudoit et al. 2002).

Des techniques de séquençage de nouvelle génération telles que le RNA-Seq quantifient directement par fragments l'expression du produit des gènes dans un échantillon. La longueur de ces fragments (ou "reads") varie en fonction des technologies et des bouts du génome observés. Le génome complet d'un organisme s'obtient par assemblage de ces reads et l'expression des gènes par la quantification des reads qui correspondent à la séquence du gène. Toutefois le séquençage à haut débit génère des quantités importantes de données dont la gestion n'est pas totalement maîtrisée, ni automatisée. L'analyse des données brutes requiert une grande puissance de calcul et leur exploitation nécessite le développement ou l'utilisation de méthodes bioinformatiques qui ne sont pas aujourd'hui standardisées. Beaucoup de paramètres sont en constante évolution pour limiter les biais non dûs à l'effet biologique que l'on souhaite mesurer : réglages au niveau manipulation, traitements bioinformatiques et prétraitements statistiques. Comme dans toute analyse statistique, il est également nécessaire de disposer d'un minimum d'individus par condition (réplicats). Le coût de séquençage étant assez élevé et les ressources matérielles parfois limitées, il est encore difficile d'obtenir un grand nombre de réplicats biologiques par condition. Si l'objectif est de déterminer les gènes DE entre deux conditions, on voudra détecter si le nombre de comptages de leur reads entre les deux conditions est significativement différent pour ces gènes, en contrôlant les erreurs possibles.

Les desiderata des biologistes, des financeurs et des statisticiens sont antagonistes ; l'objectif du travail présent est de donner quelques éléments de décision quant au nombre minimum de réplicats. Liu et al. (2014) propose une étude sur données réelles de compromis entre profondeur de lecture (nombre de séquences qui couvrent un locus du génome) et nombre de réplicats. Notre étude sur un jeu de données réelles chez l'homme qui a motivé notre questionnement, et nous proposons un cadre de simulations pour apporter des réponses en environnement où biais techniques, effets biologiques sont maîtrisés. Nous ne disposons pas encore des résultats exploitables mais ils seront bientôt disponibles.

2 Données de RNA-Seq, quelques caractéristiques

Distribution des données Une approche naturelle pour les données de comptage est d'utiliser une distribution de Poisson. Elle modélise bien les évènements rares, comme le séquençage d'une molécule. Mais lorsqu'une corrélation positive existe entre le séquençage de plusieurs fragments pour un même gène, la dispersion observée est plus importante ; Robinson & Smyth (2008) notent une non-adéquation à la sur-dispersion de ces données lors d'observations répetées de la même condition (les réplicats biologiques). Le consensus actuel est d'utiliser une distribution binomiale négative, sorte de modèle de Poisson sur-dispersé. Lu et al. (2005) montrent sur données simulées que l'hypothèse de distribution binomiale négative peut être robuste même lorsque les données ne suivent pas exactement cette distribution.

La paramétrisation comprend la moyenne μ des comptages et la "dispersion" ϕ. La

probabilité d'observer k comptages pour la variable $X \sim \mathcal{NB}(\mu, \phi)$ est:

$$P(X = k) = \frac{\Gamma(k + 1/\phi)}{\Gamma(1/\phi)\Gamma(k + 1)} \left(\frac{1}{1 + \mu\phi}\right)^{1/\phi} \left(\frac{\mu}{1/\phi + \mu}\right)^{k}.$$

On a alors la moyenne de X : $E[X] = \mu$ et sa variance $var(X) = \mu + \phi\mu^2$, ce qui implique la contrainte que $\phi > -1/\mu$ et $\phi > 0$ pour la sur-dispersion des données de RNA-Seq.

Dans notre modélisation, le comptage moyen pourra dépendre du gène et de la condition expérimentale mais pas du réplicat. La paramètre de dispersion sera lui estimé commun sur les conditions, de manière approchée par maximum de vraisemblance conditionnelle pondérée (Robinson & Smyth 2008).

Test d'analyse différentielle Cela permet de définir un test exact d'égalité des moyennes entre les conditions pour chaque gène; données et paramètres de la distribution sont corrigés par un ajustement quantile. Les données ajustées ne sont alors qu'approximativement identiquement distribuées. Pour simplifier, nous limitons la présentation au cas de deux conditions. Une généralisation de ce test est disponible pour des tests à plus de deux conditions et a été utilisée pour l'analyse des données réelles à la Section 3.

Sous l'hypothèse nulle, les valeurs de comptage normalisés observées ne dépendent pas de la condition. L'idée est d'utiliser la somme des comptages dans les deux conditions qui a aussi une distribution conditionnelle binomiale négative. Si cette somme a un comptage anormal, le gène concerné sera déclaré DE. La p-valeur associée est la probabilité que la valeur observée ne soit pas plus extrême que celle attendue sous l'hypothèse nulle de distribution identique de comptage dans les deux conditions. Ce test est un analogue du test exact de Fisher pour les tables de contingence où la loi hypergéométrique est remplacée par la loi binomiale négative. Toutes les analyses présentées ici ont été faites avec le logiciel R et avec le package Bioconductor `edgeR`.

3 Motivation : l'analyse du jeu de données réelles

Nous avons étudié l'action d'une protéine impliquée dans la déstabilisation du processus de traduction des ARN cellulaires. Quatre conditions (design hiérarchique) ont été comparées : ARN non traduits (+) versus ARN traduits (-) *via* un contrôle par induction, en absence (L) ou en présence (H) de la protéine d'intérêt (traitement). La question biologique est de savoir s'il y a une interaction entre ces quatre conditions. Pour chaque gène le contraste d'interaction défini par $[(L-) - (L+)] - [(H-) - (H+)]$ est donc testé. L'hypothèse nulle correspond à la nullité de ce contraste.

Cette étude n'a pas permis d'obtenir des gènes différentiellement exprimés avec une p-valeur corrigée (Benjamini & Hochberg 1995) significative. Pour ne pas perdre le bénéfice d'avoir 5 réplicats par condition, une des particularité forte des caractéristiques de cette

Nombre de réplicats		5	4	3	2
Nombre de gènes DE	min	152	120	93	88
	Q1	152	133	132	122
	mediane	152	147	178	214
	Q3	152	157	261	367
	max	152	207	2445	2328

Table 1: Comptages des gènes DE selon le nombre de réplicats tirés de manière aléatoire

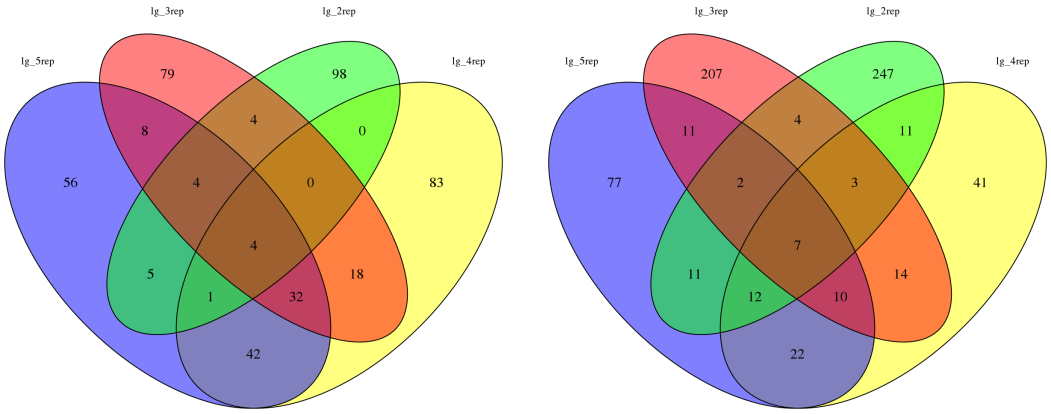

Figure 1: Diagramme de Venn des gènes DE (p-valeur brute à 1%) pour deux essais (essai 10 à gauche et 14 à droite) selon le nombre de réplicats utilisés dans l'analyse.

expérience, nous avons réalisé l'analyse en prenant la p-valeur brute, fixée à 1% (résultats à 5% non montrés ici). Ensuite, nous avons réalisé 30 "essais" sous forme de rééchantillonage des données en enlevant un ou des réplicats ; pour chacun des essais, nous avons tiré aléatoirement sans remise 4 réplicats et réalisé ensuite l'analyse différentielle. La même opération a été réalisée avec 3 réplicats au lieu de 4 puis 2. Pour chacun de ces essais, nous avons exhibé la liste des gènes DE et les avons comptés (voir Tableau 1). Toutes ces listes ont été comparées à la liste des gènes DE avec 5 réplicats. Par exemple, la Figure 1 montre deux exemples de comparaisons sous forme de diagramme de Venn. Le diagramme de Venn de gauche (essai 10) est une situation assez favorable où une intersection satisfaisante entre les analyses à 5 et 4 replicats est observée. Le diagramme de Venn de droite (essai 14) montre une intersection nettement plus limitée.

Les résultats obtenus montrent que globalement avec 4 réplicats, nous obtenons des résultats proches de ceux obtenus avec 5 réplicats. A partir de 3 réplicats, le nombre de gènes DE et la liste de ces gènes devient très aléatoire. De plus les diagrammes de Venn, montrent qu'entre 4 et 5 réplicats, nous obtenons environ 45% de gènes communs avec

une p-valeur brute à 1%. A 3 réplicats, ce pourcentage chute à environ 30% dans des cas favorables et peut être bien en deça. Aussi, très peu de gènes sont communs entre les 4 listes. Enfin, lorsque l'on mesure la stabilité du rééchantillonage, c'est à dire que l'on regarde dans combien d'essais on retrouve chaque gène DE lorsque l'on a de 4 à 2 réplicats, on voit que moins on a de réplicats, moins on retrouve de manière systématique chacun des gènes DE. Pour 4 réplicats, seulement 4 gènes sont retrouvés DE parmi les 30 essais, l'immense majorité des gènes DE n'est pas détectée pour plus de 2 essais. Les chiffres sont encore plus dramatiques pour 3 réplicats où aucun gène n'est DE dans les 30 essais, 1 seul l'est dans plus de 12 essais et environ 75% des gènes ne sont détectés DE que dans un unique essai !

La conclusion de cette partie est que sur un critère de robustesse des résultats, 4 réplicats semble le minimum vital pour arriver à aboutir à des résultats fiables et reproductibles dans notre cas de figure.

Cependant, la fiabilité des gènes DE (en limite de détection) ne permet pas répondre de manière satisfaisante à la question de l'influence du nombre de réplicats dans une analyse différentielle de données RNA-Seq avec seulement ce jeu de données. Nous proposons donc une étude sur données simulées avec différents paramètres qui représenteront de manière simplifiée des caractéristiques de la complexité des jeux de données RNA-Seq traditionnels. Elle permettra des premiers éléments de réponse dans un cadre maîtrisé, où la liste des gènes DE est connue. On pourra quantifier les performances du test mis en jeu selon le nombre de réplicats utilisés. L'alternative de travailler sur un jeu de données réelles dont la liste des gènes DE est (plus ou moins) connue est séduisante mais (i) nous ne disposons pas d'un tel jeu de données avec 5 réplicats et (ii) même avec un bon niveau de confiance sur cette liste, la décision de déclarer un gène DE/non DE comporte toujours un risque.

4 Construction de jeux de données simulées

Par rapport à la Section 3, nous simplifions le design expérimental à $c = 2$ conditions. Nous conservons un maximum de $n = 5$ réplicats et nous considérons $p = 10\,000$ gènes, dont une proportion $q = 5\%$ sera DE. L'idée est de construire des jeux de données simples d'abord (un écart de moyenne entre les 2 conditions pour les gènes DE) puis se complexifiant au fur et à mesure comme nous l'expliquons ci-après:

- Le comptage des gènes non DE suivent une loi $\mathcal{NB}(\mu, \phi)$. Les gènes DE suivent la même loi sous une condition (vue comme un contrôle) et sous la deuxième condition, ils suivent une loi $\mathcal{NB}(\mu', \frac{\mu-\mu'+\phi\mu^2}{(\mu')^2})$ afin d'avoir une moyenne différente et une variance identique. La différence $\mu - \mu'$ donne l'intensité du signal et le paramêtre ϕ contrôle l'intensité du bruit. On pourra aussi simuler des gènes DE activés ($\mu - \mu' > 0$) ou inhibés ($\mu - \mu' < 0$).

- La deuxième configuration des simulations, au lieu de considérer une différence $\mu - \mu'$ fixe, aura une intensité de signal aléatoire ; tous les gènes DE ne le sont pas avec la même intensité. Alors chaque gène i suit une loi $\mathcal{NB}(\mu_{i,1}, \phi)$ dans la condition 1 et une loi identique dans la condition 2 (gènes non DE) où une loi $\mathcal{NB}(\mu_{i,2}, \frac{\mu_{i,1}-\mu_{i,2}+\phi\mu_{i,1}^2}{(\mu_{i,2})^2})$ (gènes DE) ; les lois de $\mu_{i,1/2}$ devront être précisées, *e.g.* gaussiennes de moyennes différentes. On pourra aussi proposer une loi Gamma pour ϕ.

- la troisième proposition aura pour but d'introduire des corrélations entre les gènes sous la forme d'un "modèle binomial négatif multivarié". Le but est de rendre compte du fait que les gènes ne sont pas indépendants et sont organisés sous forme de réseau. On s'inspirera de Kopociński (1999) et de Nikoloulopoulos & D. Karlis (2010) ; la structure de variance-covariance pourra être spécifiée de manière analogue à celle d'un modèle gaussien multivarié mais les calculs des distributions sont assez lourds à notre connaissance. Cette piste ne sera explorée que lorsque les modèles des deux premiers points auront été exploités. Eventuellement, on pourra considérer des distributions multivariées mieux étudiéees (comme la loi de Poisson multivariée) mais toujours lourde sur le plan calculatoire.

Remerciements

Les auteurs remercient David Alter, Hatim Hadria et Kais Sahli pour leur participation à ce travail pendant leur projet étudiant.

Bibliographie

[1] S Dudoit et al. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sinica*, 12:111-139.
[2] MD Robinson & GK Smyth (2008) Small sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics*, 9:321–332.
[3] Y Benjamini & Y Hochberg (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc, B*, 57: 289–300.
[4] Y Liu et al. (2014) RNA-seq differential expression studies: more sequence or more replication ? *Bioinformatics*, 30(3):301–4.
[5] B Kopociński (1999) Multivariate negative binomial distributions generated by multivariate exponential distributions. *Appl. Math.*, 25(4):463–472.
[6] AK Nikoloulopoulos & D Karlis (2010). Modelling multivariate count data using copulas. *Commun. stat.-Simul. C.*, 39:182–197.