metaRNASeq : un package pour la méta-analyse de données RNA-seq

G. Marot[a,b], **F. Jaffrézic**[c] and **A. Rau**[c]

[a]EA2694 Centre d'Etudes et de Recherche en Informatique Médicale
Université Lille 2
1 place de Verdun, 59045 Lille cedex
guillemette.marot@univ-lille2.fr

[b]Equipe Projet Inria MODAL
Inria
40 avenue Halley - Bat A , 59655 Villeneuve d'Ascq cedex
guillemette.marot@inria.fr

[c]UMR1313 Génétique Animale et Biologie Intégrative
INRA
Domaine de Vilvert - 78350 Jouy-en-Josas cedex
florence.jaffrezic@jouy.inra.fr

Mots clefs : Biostatistique, méta-analyse, analyse différentielle, RNA-seq, transcriptomique, séquençage haut débit

Les techniques de séquençage à haut débit telles que le RNAseq sont de plus en plus utilisées pour les analyses de données transcriptomiques. Cependant, en raison du coût encore élevé des expériences, peu de réplicats biologiques sont inclus dans les études, ce qui affecte la capacité de détection des vrais transcrits différentiellement exprimés. Il est probable qu'avec la diminution du coût du séquençage, des expériences soient reconduites pour répondre à des questions déjà posées et gagner en sensibilité en rajoutant des réplicats. Il est donc nécessaire de développer des techniques qui puissent analyser conjointement les résultats d'analyse différentielle de différentes études. Ces méthodes doivent tenir compte de la variabilité biologique et technique à l'intérieur de chaque expérience ainsi que de l'effet inter-études [1]. Nous avons comparé les méthodes de combinaison de p-values déjà utilisées dans des analyses de puces à ADN à un modèle linéaire généralisé (GLM) incluant un effet étude. Ces comparaisons à la fois sur des jeux de données réels et des simulations ont confirmé que le GLM avec effet étude se comportait très bien quand peu d'études étaient disponibles et que l'effet étude était faible. Elles ont aussi montré que les techniques de méta-analyse étaient plus performantes que le GLM étudié quand la variabilité entre études était grande et le nombre d'études important.

Le package metaRNAseq, disponible sur R-Forge, implémente les techniques de méta-analyse présentées dans [1]. Après avoir redonné les principaux résultats du papier correspondant, nous présenterons rapidement la vignette de ce package en insistant sur les différences entre les techniques précédemment utilisées pour les puces à ADN [2] et celles développées pour le séquençage [1]. Ces différences concernent notamment la gestion des conflits entre les gènes sous-exprimés dans une étude et sur-exprimés dans une autre. Nous porterons aussi une attention particulière sur les vérifications préliminaires à effectuer dans une méta-analyse de données RNA-seq pour se placer dans un cadre où les techniques implémentées sont effectivement les meilleures. En particulier, nous insisterons sur la nécessité d'observer des distributions de p-values uniformes sous l'hypothèse nulle dans chaque étude, ce qui est possible en utilisant la

méthode développée par [3].

Références

[1] Rau, A., Marot, G., Jaffrézic, F. (2013). Differential meta-analysis of RNA-seq data from multiple studies. In preparation.

[2] Marot, G. , Foulley, J.-L., Mayer, C.-D., Jaffrézic, F. (2009). Moderated effect size and P-value combinations for microarray meta-analyses, *Bioinformatics*, **25**(20), 2692–2699

[3] Rau, A., Gallopin, M., Celeux, G., Jaffrézic, F. (2013). Independent data-based filtering for replicated high-throughput transcriptome sequencing experiments. Submitted.