Lundi 08/01/2018

Séance de TP : Analyse statistiques de données de transcriptome obtenues par puces d'expression

Dans cette séance de TP, vous allez travailler sur un jeu de données de transcriptome obtenues chez une levure à l'origine d'infections nosocomiales, *Candida Parapsilosis.* L'objectif est d'identifier les gènes dont l'expression est modifiée en réponse à un manque d'oxygène (hypoxie) dans le milieu de culture des cellules (*C.parapsilosis*). Les données utilisées sont issues de la publication scientifique Guida et al. (BMC Genomics,2011). Vous aurez à procéder à l'analyse différentielle des transcrits entre les conditions hypoxique et normoxique. Pour chaque condition, plusieurs réplicats biologiques ont été préparés. Pour les puces, 2 échantillons normoxiques et 2 échantillons hypoxiques ont été hybridées à 4 puces. Les données vous sont présentées ici en log2 du ratio des intensités: log2(Hypoxia/Normoxia).

Les fichiers nécessaires à ce TP sont sur moodle dans le dossier TP_analyse_transcriptome. Téléchargez dans votre répertoire de travail.

Fichier de données :
Microarrays_logValues_parapsilosis.txt

Outils d'analyse :
R avec les packages de bioconductor (http://www.bioconductor.org)
limma : package "gold standard" pour l'analyse des puces qui peut aussi être utilisé pour le RNASeq

Script:
TP_transcriptome_arrays_M1BI_2017-2018.R

→ Suivez le script "TP_transcriptome_arrays_M1BI_2017-2018.R" **pas à pas** et répondez au fur et à mesure aux différentes questions ci-dessous.

Q1. *Les données de puce semblent-elles avoir été normalisées?*

Q2. *Les réplicats vous semblent-ils bien corrélés entre eux?*

Q3. *Que pensez-vous des moyennes des log2FC par rapport aux écarts-types?*

Q4. *Avec le test de Student, quel est l'effet de la correction de Bonferroni sur le nombre de gènes différentiellement exprimés? Au final, combien de gènes sont-ils up ou down régulés avec la correction de Bonferroni?*

Q5. *Que pensez-vous des valeurs de FC obtenues par limma comparées à celles de Student? Quelle différence observez-vous entre les valeurs t modérées de limma et les valeurs t de Student? Comment se comporte la statistique t en fonction de l'écart type? Concluez.*

Q6. *Avec limma, combien de gènes sont-ils up ou down régulés au seuil ajusté de 0.0001? Les gènes identifiés avec Student sont ils retrouvés?*

Q7. *Proposez une liste de 20 gènes différentiellement exprimés.Discutez des résultats et des perspectives de travail.*