Examen pratique

Sujet 1 : Analyse des données transcriptomiques

Les données de RNA-Seq ont été générée dans le cadre d'une étude évaluant l'impact du stress osmotique sur l'expression des gènes chez la banane. Dans cette étude 18 échantillons issus de 3 types de bananes triploïdes ont été séquencés en mRNA. Dans cette étude les chercheurs ont pu identifier des gènes différentiellement exprimés et mettre ainsi en lumière différents processus biologiques en lien avec le stress osmotique. Des détails sur la génération et l'interprétation des résultats peuvent être trouvés ici : https://europepmc.org/article/MED/26935041. Répondez aux questions suivantes.

Un document de 6 pages **maximum** devra être rendu accompagné du code qui a permis de produire les résultats. Des points seront attribués à la qualité des figures et des réponses.

1) Sur la base de la section méthodes de l'article dans lequel les données ont été générées, dîtes si la méthodologie vous semble adéquat. Justifiez à la lumière de ce que l'on a vu en cours (modèles, normalisation, etc...)
2) Combien y'a-t-il de gènes ? Combien y'a-t-il d'échantillons dans chaque groupe ? Concluez sur l'impact du design sur les résultats attendus.
3) Après avoir pré-processé convenablement les données, appliquez une analyse en composantes principales (ACP) sur les données transformées. Projetez les individus sur les deux premières composantes principales. ? Combien de variance expliquée est gardée dans les deux premières composantes ? Concluez.
4) Après avoir produit la figure de moyenne contre variance pour l'expression de tous les gènes déterminez s'il y a présence de sur-dispersion. Si oui représentez la graphiquement. Quelle est la force d'association entre moyenne et variance ? Ce résultat est-il attendu ? Justifiez.
5) Rappelez brièvement la structure des données de RNA-Seq et le modèle statistique le plus approprié. Ecrivez le modèle formellement dans le cas d'une analyse d'expression différentielle en supposant aucune autre variable que la condition d'intérêt.
6) Décrivez les grandes étapes sous-jacentes à DESeq2. Vous pouvez utiliser un diagramme pour supporter votre réponse.
7) En vous appuyant sur la fonction *plotDispEsts* dîtes si le modèle d'ajustement de la dispersion est approprié pour ces données. Si non, ajustez-le en conséquence.
8) En utilisant les paramètres par défaut, combien de gènes sont significativement différentiellement exprimés au seuil alpha = 0.05 ? Combien sont sur-exprimés ?

sous-exprimés ? Est-ce que ces résultats sont cohérents avec les hypothèses du modèle ? Justifiez.

9) En choisissant une procédure appropriée, corrigez vos résultats pour la multiplicité. Combien de gènes sont significativement différentiellement exprimés ?
10) Interprétez le modèle pour le gène le plus sur-exprimé/sous-exprimé.
11) Après avoir produit le MAplot que pouvez-vous déduire de la relation entre valeur d'expression moyenne et log-fold change ? Est-ce attendu ?
12) Produisez le volcano plot, affichez en rouge les gènes significativement différentiellement exprimés au seuil alpha = 0.05 avec une valeur absolue de log-fold change supérieure à 2. Combien de gènes ressortent d'intérêt ici ?
13) En utilisant l'option lfcThreshold de la fonction *results*, resortez les gènes significativement différentiellement exprimés au seuil alpha = 0.05 avec un log-fold change supérieur à 2 (valeur absolue). Comparez vos résultats avec la question précédente. Quelle stratégie vous semple la meilleure ? Justifiez.
14) Produisez MAplot et volcanoPlot en utilisant l'approche pénalisée. Justifiez de l'avantage ou du désavantage d'une telle approche dans votre contexte.
15) En utilisant les paramètres qui vous semblent les plus appropriés, réappliquez DESeq2 en ajustant pour le génotype. Comparez vos résultats avec ceux obtenus précédemment.

Dans la seconde partie du travail, il vous sera demandé d'entrainer et de comparer différents modèles d'apprentissage automatique pour prédire le niveau développemental de graines de soja sur la base des données d'expression des gènes. Vous avez à disposition 273 échantillons. Les données à disposition sont normalisées avec le TPM. Vous pourrez cependant utiliser tout autre type de normalisation vue en cours que vous jugerez pertinente.

16) En utilisant deux modèles de clustering appropriés, regroupez les gènes entre eux. Faites la même chose pour les échantillons. Quelles sont vos interprétations ?
17) Entrainez et comparez deux modèles de classification pour prédire le stage développemental sur la base de l'expression des gènes. Justifiez le choix des modèles. Discutez des forces et faiblesses de chacun. Comparez les performances prédictives à l'aide de la bonne métrique. L'emphase devra être mis sur des modèles interprétables. Décrivez votre procédure.
18) Discutez des extensions de votre démarche à d'autres omics. Comment considérer l'intégration d'une autre source biologique ? Quelle stratégie ? etc… Concluez.