Projet 2021/2022 – Prototype machine

Vous devez créer une interface graphique pour que les biologistes puissent visualiser les résultats d’une analyse transcriptomique portant sur le champignon *Trichoderma reesei*. Pour cela, on vous demande de créer une web-application *via* la solution Shiny présente sous Rstudio : <https://shiny.rstudio.com/>

Les données sont à récupérer sur le dépôt Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>) sous le numéro d’accession : **GSE82287**. Ces données correspondent à des données transcriptomiques après alignement sur le génome et comptage. Vous disposez donc de 36 échantillons, correspondant à :

* de 43 à 48 : 6 réplicats en Glucose à 24h
* de 49 à 54 : 6 réplicats en Glucose à 48h
* de 55 à 60 : 6 réplicats en Lactose à 24h
* de 61 à 66 : 6 réplicats en Lactose à 48h
* de 67 à 68 : 2 réplicats en 90% Glucose et 10% Lactose à 24h
* de 69 à 72 : 4 réplicats en 75% Glucose et 25% Lactose à 24h
* de 73 à 74 : 2 réplicats en 90% Glucose et 10% Lactose à 48h
* de 75 à 78 : 4 réplicats en 75% Glucose et 25% Lactose à 48h

Les biologistes ont construit ce design expérimental pour aider à la compréhension fine des mécanismes de production de cellulases en condition lactose par rapport à une condition non inductrice qu’est le glucose. Pour cela, deux temps ont été choisi, un où l’induction est encore en cours (24h), et l’autre où l’induction est terminée et la production stable (48h). Il se pose également une question quant à une certaine linéarité par rapport à la concentration en lactose dans la régulation. Pour répondre à ces questions, plusieurs analyses différentielles peuvent être regardées, toutes par rapport à la condition de glucose à 24 ou 48h.

A partir de ces données transcriptomiques, vous devez réaliser six analyses différentielles :

* Lactose vs Glucose à 24h
* Lactose vs Glucose à 48h
* 90% Glucose et 10% Lactose vs Glucose à 24h
* 90% Glucose et 10% Lactose vs Glucose à 48h
* 75% Glucose et 25% Lactose vs Glucose à 24h
* 75% Glucose et 25% Lactose vs Glucose à 48h

*Note : dans ces données, l’identifiant des gènes est celui du JGI (Joint Genome Institute). Deux autres types d’identifiants existent : ceux de l’annotation officielle (Tr…) et ceux de FungiDB (TRIREDRAFT\_...). La table de correspondance entre les différents identifiants est fournie en complément de ce document dans le fichier MappingTable\_geneID.xlsx. Vous aurez également besoin du fichier d’annotation structurelle du génome de Trichoderma reesei. Fichier également fourni avec ce document : QM6aAnnotationIFPEN2021strict.gff*

L’interface graphique devra répondre aux spécifications des biologistes suivantes :

* Pour les trois types de comparaison et chaque temps 24h et 48h pris séparément, afficher sur le même graphique, les **distributions des RPKM moyens entre les deux conditions** et faite un **test statistique** pour indiquer si les niveaux d’expression moyen des gènes sont significativement différents entre les conditions étudiées. Afficher également le résultat/conclusion du test statistique.
* Proposer une visualisation du **nombre de gènes sous- et sur-exprimés dans chaque des comparaisons**. Demander à l’utilisateur la comparaison qu’il souhaite voir en utilisant un *widget* interactif. Ajouter la possibilité de voir toutes les données sur une seule et même figure *(asctuce : utiliser les facets)*.
* Extraire les gènes différentiellement exprimés dans au moins une des six analyses. A partir de ces gènes, construisez la matrice d’expression des gènes composés de ces gènes et de tous les logFC associés à ces gènes dans les 6 comparaisons. Réaliser un clustering (K-means) de ces données et proposer une visualisation montrant les **résultats du clustering**. Le nombre de classe sera laissé au choix de l’utilisateur au travers d’un *widegt*. Afin d’aider l’utilisateur dans le choix du nombre de classe, un calcul de silhouette sera effectué et rendu dans l’interface. Le rendu du clustering devra comprendre *a minima* : le **nombre de gènes par classe** (et la représentative de la classe), les **profils d’expression des gènes par classe** et les **profils d’expression moyens** sur les différentes classes. *(astuce : afin de faciliter la lecture des résultats, choisissez une couleur par classe)*

*Bonus :*

* *permettez à l’utilisateur de rentrer ses données directement via l’interface. Pour cela, inspirer vous de cet exemple :* [*https://shiny.rstudio.com/gallery/file-upload.html*](https://shiny.rstudio.com/gallery/file-upload.html)*.*
* *Permettre à l’utilisateur de définir le design des analyses différentielles au travers de l’interface. Proposer alors une interface qui permette à l’utilisateur de choisir le nombre d’analyses différentielles, avec pour chaque, la condition de référence et la condition de test.*

Devra être renvoyé par mail au plus tard le **31 janvier 2022**, le dossier complet nécessaire à l’exécution de l’application. Assurez-vous que le dossier contient tous les codes/données/etc nécessaire à sa bonne exécution. Un code non fonctionnel engendrera une note inférieure à 10.

Au-delà de l’aspect purement fonctionnel, la notation se fera sur la pertinence des visualisation choisie et la clarté de leur mise en place.