TP1 - Découverte du jeu de donnée

Guillaume Meurice - guillaume.meurice@aphp.fr

07 juin 2023

Consignes

Les données nécessaires à cette séance se trouvent sur le site suivant :

- https://tinyurl.com/epitaTP1
- mot de passe : `25Kg.bwF`

Répondez aux questions dans un document Rmarkdown produisant un fichier **HTML**.

Prenez soin de systématiquement décrire les figures, en définissant les axes représentés, et les données affichées. Décrivez vos obsevations et vos interprétations.

Nommez votre rapport en suivant la convention suivante : NOM_PRENOM_TP1.pdf ou NOM_PRENOM_TP1.html et envoyer le avant le 17 juin 2023.

Pré-requis

Ce TP nécessite le chargement de 2 packages :

- package pheatmap
- package RColorBrewer

```
install.packages("pheatmap")
install.packages("RColorBrewer")
```

Objectif

L'objectif du TP est de découvrir un jeu de données RNAseq, de se familiariser avec celui-ci, et d'en proposer une première analyse non supervisée.

Données

Les ARNm de 17 échantillons de tumeurs du sein de 3 types - HER2 positif (HER2), triple négatif (TNBC), non triple négatif (NonTNBC) - et de 3 échantillons de sein normal (épithelium) ont été séquencés par Illumina HiSeq2000.

Les données brutes sont disponibles sur le site **Sequence Reads Archive** (SRA), hebergé au NCBI, sous la référence SRP032789.

Ces données ont été utilisées pour des travaux scientifiques publiés dans les 2 articles suivants :

- J.Eswaran et al. Transcriptomic landscape of breast cancers through mRNA sequencing. Scientific Report (2012) article
- J.Eswaran et al. RNA sequencing of cancer reveals novel splicing alterations. Scientific Report (2013) article

Afin d'obtenir la table d'expression utilisée dans ce TP, les données brutes (fastq) ont préalablement été nettoyées, et alignées sur le génome de référence humain hg19 avec l'outil STAR. Puis la quantification a été faite avec l'outil `feature-count`. Cet outil génère un tableau avec en ligne les gènes, et en colonnes les individus. On y retrouve l'annotation des gènes ainsi que les valeurs d'expression brutes, aussi appelées "comptages".

Un peu de biologie

Question 1 : Que sont les cancers "TNBC" ?

Question 2 : A l'aide du site GeneCards (https://www.genecards.org/), donner la définition des gènes `ERBB2`, `PGR`, `ESR1`.

Manipulation des données

Chargez en mémoire le fichier de description des échantillons `annot_sample.txt`.

Question 4 : Indiquez le nombre d'échantillons par `condition` (indice : fonction `table`). Représenter cette répartition sous forme de `pie chart` (fonction `pie`).

Question 5 : Que pensez vous du design expérimental de cette étude (taille des groupes, etc) ?

Chargez en mémoire la table de comptage (fichier `counts.txt`)

Question 6 : A partir du fichier de comptages, créer un tableau `annotgene`, contenant uniquement les données d'annotations des gènes et un tableau `comptage`, contenant uniquement les valeurs de comptages.

Sur le tableau de comptage, modifier les noms des lignes pour qu'ils correspondent aux noms des gènes, et les noms de colonnes pour qu'ils correspondent aux noms d'échantillons (`sampleName`).

Un biais important en RNA-seq est la profondeur de séquençage de chaque échantillon, aussi appelée taille de la librairie, et qui correspond à la quantité total de matériel séquencé par échantillons.

Question 7 : Calculer la profondeur de séquencage pour chaque échantillons.

Question 8 : Représentez en barplot la taille de librairie de chaque échantillon. Adaptez la couleur des barplot en fonction de la `condition` (1 couleur par `condition`). Est ce que la profondeur de séquencage est équivalente pour tous les échantillons ? Que pourriez-vous suggérer ?

Chargez en mémoire le fichier de comptages normalisés `counts_normalized.txt`.

Question 9 : Afin d'observer l'effet de la normalisation sur les données, représenter sous forme de `boxplot` les données avant et après normalisation. Représentez ces 2 graphiques sur une même fenêtre.

Question 10 : Afin d'avoir une meilleure représentation des données, transformez les matrices de comptages non normalisés et normalisés en log2 (fonction `log2`). Représentez une nouvelle fois les 2 boxplots, après transformation des valeurs en log2 et commentez le résultats obtenus.

En RNA-seq, il est courant d'avoir des valeurs de comptages à 0, ce qui correspond à des gènes non exprimés.

Question 11 : A partir de la matrice de `comptage normalisés`, affichez un graphique représentant le nombre de gènes ayant des comptages nuls en fonction des échantillons (utilisez le meme code couleur pour les échantillons que précedemment) . Que remarquez-vous ?

Question 12 : Combien de gènes ne sont jamais exprimés chez tous les échantillons ? Créez une nouvelle matrice ne contenant pas ces gènes. Affichez les dimensions de cette nouvelle matrice. Transformez cette matrice en log2. Quelle est la valeur minimale par échantillon ? Pourquoi ? Afin d'éviter cette valeur, gênante pour la suite des analyses, nous allons ajouter un pseudocount de 1 avant de passer en log2. Créez cette nouvelle matrice.

Analyse non supervisée

`IMPORTANT` : Pour la suite du TP, nous travaillerons uniquement sur cette nouvelle matrice de comptages, c'est à dire la matrice de comptages normalisés, en log2, sans les gènes non exprimés chez tous les échantillons.

Nous allons visualiser la proximité relative des observations, grâce à une **Analyse en Composantes Principales**. Il s'agit d'une méthode d'analyse multivariée par réduction de dimension. Les composantes principales sont des combinaisons linéaires des variables. Elles ont pour contraintes de maximiser la variance entre les observations et d'être orthogonales entre elles. Le nombre de composantes est égal au rang de la matrice des données. On utilise la fonction `prcomp` de `R base`.

Afin de pouvoir construire le graphique de l'ACP, suivez les étapes ci-dessous : > * Transposez votre matrice de comptage grâce à la fonction `t()`. > * Calculez les composantes principales grâce à la fonction `prcomp` avec les options `scale=TRUE` et `center=TRUE`. > * Représentez graphiquement les observations, c'est à dire les échantillons, en fonction des deux premières composantes (les composantes principales sont disponibles dans la sous-variables `$x` de votre object issu de la fonction `prcomp`)

- colorez les points en fonction de la colonne "condition" et changez la taille des points (paramètre `cex` de la fonction `plot`) en fonction du nombre de gènes non exprimés du tableau de données.
- Ajoutez les lignes x=0 et y=0 en pointillé.

Question 13 : Commentez le graphique obtenu. Est ce que les résultats de l'ACP correspondent à ce que vous attendiez ? Peut-on associer les composants principaux à une information biologique ?

Il est possible de visualiser la proximité relative des observations, grâce à une autre méthode : **le clustering**, et en particulier le **clustering hiérarchique**. Cette methode consiste à calculer une distance entre les profils transcriptomique, puis à les regrouper de proche en proche. Il existe de nombreuses méthodes pour calculer une distance entre deux profils (distance Euclidienne, distance de Manhatan . . .) et de nombreuses méthodes pour agréger les profils les plus semblables entre eux.

Afin de pouvoir construire le graphique, suivez les étapes ci-dessous :

- Chargez les library `pheatmap` et `RColorBrewer` si ce n'est pas déjà fait.
- Transposez votre matrice de comptage grâce à la fonction `t()`.
- Appliquez la fonction `dist` (avec les paramètres par défaut) à la matrice de comptage `transposée`.

Question 14 : Quelle est la classe de l'objet généré ? Quelle est la méthode par defaut utilisée pour calculer la distance ?

- Créer un vector de couleurs de dégradé de bleu grâce à la commande suivante : `colors <- colorRampPalette( rev(brewer.pal(9, "Blues")) )(255)`.

- Générez une heatmap `sample-to-sample` grâce à la fonction `pheatmap`, et y ajouter une ligne d'annotations correspondant à l'annotation `condition`.

Question 15 : Est ce que les résultats du clustering correspondent à ce que vous attendiez ? Les résultats sont ils concordants avec ceux obtenus par l'ACP ?