TP3 Analyse differentielle

Justine Guégan - j.guegan-ihu@icm-institute.org
Guillaume Meurice - guillaume.meurice@gustaveroussy.fr

08 décembre 2016

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

https://pf-bb.github.io/CentraleSupelec-R-genomics/TP3.html.

Objectif

L'objectif du TP est d'étudier la modulation de l'expression des gènes entre des échantillons de cancer du sein et des échantillons non tumoraux. Pour cela, des expériences de RNA-seq ont été réalisées. Il s'agit ici de mener une analyse différentielle des données de séquencage. Cette analyse différentielle permettra de mettre en évidence quels gènes sont différentiellement exprimés entre les différents sous-type de cancer du sein, et le tissu normal.

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

Données

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

Analyse des données

Lors du TP précedent, vous aviez utilisé une matrice de comptages bruts, que vous aviez nettoyée afin de supprimer les gènes qui ne s'expriment dans aucun échantillon. Nous vous proposons de reprendre les analyses à ce stade.

1. Lecture des fichiers de données

> **Question 1.1** : Chargez en mémoire la table de comptages créée lors du TP2 (fichier `counts_normalized.txt`). Quelle est la classe de l'objet créé ? Transformez le en matrice. Chargez en mémoire le fichier de description des échantillons (fichier `annot_sample.txt`).

2. Création des `MA-plot`

Le MA-plot est une figure permettant de représenter de façon synthètique une comparaison de 2 groupes d'intérêt. Nous vous proposons ici quelques définitions permettant de comprendre comment est construit un MA-plot, en définissant d'abord le **Fold-Change** (FC), puis les valeurs **M** et **A**.

Soit $\bar{x}_1$ et $\bar{x}_2$, définissant respectivement pour un gène j donné, la moyenne des valeurs d'expression pour le groupe 1 et le groupe 2.

- Le FC se défini comme le ratio de la moyenne des valeurs d'expression entre deux groupes d'intérêt. Il se calcule sans la transformation logarithmique. Si on note, pour un gène j donné, $\bar{x}_1$ la moyenne du groupe 1 et $\bar{x}_2$ la moyenne du groupe 2, alors le **FC** du gène j vaut :

$$FC_j = \frac{\bar{x}_1}{\bar{x}_2}.$$

On préfère, pour des raisons pratiques, raisonner sur le *log Fold-Change* (logFC), qui est simplement obtenu en prenant le logarithme naturel du FC. Ainsi :

- si $log(FC_j) > \kappa$, on dit que le gène j est sur-exprimé dans le groupe 1 par rapport au groupe 2,
- si $log(FC_j) < -\kappa$, on dit que le gène j est sous-exprimé dans le groupe 1 par rapport au groupe 2,

avec κ une certaine valeur seuil[1], qui dépend beaucoup de l'expérience. Habituellement, on choisit $\kappa = 1$ car cela signifie que la valeur moyenne d'expression est deux fois plus (resp. moins) grande dans un groupe que dans l'autre.

- La valeur **M** correspond simplement au log Fold-Change (logFC) et est donc définie comme suit :

$$M_j = log2(FC) = log2(\frac{\bar{x}_1}{\bar{x}_2})$$

- La valeur **A** correspond à la moyenne des log2 des moyennes des valeurs d'expression :

$$A_j = \frac{1}{2} * [log2(\bar{x}_1) + log2(\bar{x}_2)]$$

Le MA-plot est le graphique qui affiche la valeur **M** en ordonnée et la valeur **A** en abscisse.

Question 2.1 : La fonction computeMean.

Créez une fonction `computeMean` permettant de calculer, pour tous les gènes, la moyenne des valeurs d'expression pour un groupe d'échantillons donné. Cette fonction prend en entrée les paramètres suivants :

- `condition` : le vecteur de description de la `condition` des échantillons
- `count` : une matrice de comptage
- `label.grp` : le label du groupe d'intérêt.

Cette fonction retourne un vecteur contenant les moyennes d'expression pour l'ensemble des gènes de la matrice de comptage, pour le groupe d'échantillons donné. Cette fonction vous servira pour la majorité des questions de ce TP.

[1] Attention, cette notation de κ pour un seuil de logFC n'est pas universelle, elle est même spécifique de cet énoncé TP. Si vous choisissez d'appeler le seuil ainsi dans votre rapport, n'oubliez pas de le préciser !

Question 2.2 : MA-plots

Affichez les MA-plots pour les groupes suivants :

- HER2 versus NBS
- TNBC versus NBS
- Non-TNBC versus NBS

Pour chaque graphique :

- ajoutez une ligne rouge à y = 0.
- ajoutez deux lignes bleu, respectivement à y = -1, et y = 1.
- Affichez le titre du graphique, ainsi que le nom des axes.
- Interpretez ces figures. Que représentent les gènes situés au dessus de la ligne y = 1 ? Que représentent les gènes situés au dessous de la ligne y = -1 ?

3. Analyse différentielle

Un gène est déclaré différentiellement exprimé si une différence observée ou un changement d'expression entre deux conditions expérimentales est significativement statistique, c'est-à-dire plus grande que la valeur attendue.

Nous avons précedemment calculé les logFC, il faut donc calculer les p-values associées à ces logFC.

Lorsque l'on fait un test d'hypothèses, une manière synthétique de représenter le résultat du test est la p-value. Par définition, la p-value obtenue représente la probabilité sous hypothèse nulle d'obtenir une statistique encore plus atypique que celle obtenue à la suite de notre expérience.

A partir du calcul de la p-value, la conclusion d'un test d'hypothèses se déroule comme suit :

- si la p-value est en dessous du seuil de rejet que j'ai choisi (habituellement : 0.05), je rejette l'hypothèse nulle,
- sinon, je ne peux pas rejeter l'hypothèse nulle.

Question 3.1 : Test de Student

Transformez la matrice `count` en `countLog2`. Calculez, pour tous les gènes (ie les lignes de `countLog2`), les p-values d'un test de Student comparant les deux moyennes des groupes HER2 et NBS en utilisant la fonction `t.test` et ses paramètres par défaut. Combien de ces p-values sont en dessous du seuil classique de 0.05 ?

Question 3.2 : Correction pour les tests multiples

Les tests d'hypothèses n'ont pas été créés dans l'optique d'être utilisés plus de 20 000 fois de façon successive : si on suit la procédure habituelle, on risque de rejeter l'hypothèse nulle à tort beaucoup trop souvent. La conséquence immédiate et néfaste de ces tests multiples est d'augmenter artificiellement le nombre de gènes différentiellement exprimés. Une correction pour les tests multiples est donc nécessaire. La procédure la plus simple est de diminuer le seuil de rejet (c'est la procédure dite de Bonferroni). Nous allons utiliser dans ce TP la procédure qui est utilisée classiquement en transcriptomique : la procédure de Benjamini-Hochberg.

Utilisez sur le vecteur des p-values calculées précédemment la procédure `p.adjust` en attribuant à l'argument `method` la valeur "BH". Après correction, combien de p-values ajustées se trouvent en dessous du seuil de 0.05 ?

Question 3.3 : Répétez les questions 3.1 et 3.2 pour les contrastes “TNBC vs NBS”, et “Non-TNBC vs NBS”

4. Représentations graphiques et conclusions

Une liste de gènes différentiellement exprimés est caractérisée par deux seuils :

- un seuil sur le log Fold-Change,
- un seuil sur la p-value corrigée

Question 4.1 : Combien de gènes passent un seuil en logFC de 1 et un seuil sur la p-value ajustée de 0.05 pour les 3 contrastes suivants :

- TNBC versus NBS
- Non-TNBC versus NBS
- HER2 vesus NBS

Question 4.2 : Volcano plot

Une première représentation graphique permettant de synthétiser ce résultat est une représentation dite en volcan : il s'agit de représenter, pour tous les gènes, un graphe bivarié, avec en abscisse le logFC et en ordonnée -log10(p-value ajustée) (ou -log10(p.value)).

Faites une représentation en volcan de votre analyse différentielle HER2 vs NBS (avec la fonction plot). Représentez sur ce graphe les seuils sur le logFC et la p-value ajustée ou non ajustée (avec la fonction `abline`). Représentez de deux couleurs différentes les gènes sur- et sous-exprimées (avec l'argument `col` de la fonction `plot`). Fixer l'échelle des ordonnées à `ylim=c(0,10)`. Représentez sur la même fenêtre graphique un volcano plot avec les p-values non ajustées et un avec les p-values ajustées. Commentez.

Question 4.3 : Carte de chaleur ou Heatmap

Enfin, représentez à l'aide de la fonction `pheatmap` une représentation des mesures d'expression des conditions HER2 et NBS (contenues dans la matrice `countLog2`) uniquement pour les gènes différentiellement exprimés pour ce contraste, et dont la légende contient les informations contenues dans le fichier d'annotations. Utilisez l'option `scale='row'`. Interprétez la figure.