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/TP4.html.

Objectif

Dans les TP2 et 3, nous avons appris à manipuler des données d'expression, à évaluer la variabilité de profils transcriptionnels, et à rechercher des gènes différentiellement exprimés entre deux groupes d'échantillons. Aujourd'hui, nous nous pencherons sur l'analyse fonctionnelle de ces gènes : quels sont les mécanismes biologiques sous-jascent, si tant est que nous puissions les identifier ?

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 généré des MA-plots, des volcanoplots et des heatmaps permettant de représenter les gènes différentiellement exprimés entre deux groupes d'échantillons. Afin de repartir sur la même base, nous vous donnons aujourd'hui les fichiers suivants :

- Fichier des comptages normalisés : counts_normalized.txt
- Fichier d'annotation des échantillons : annot_sample.txt
- Fichiers de liste de gènes : ces fichiers contiennent simplement la liste des gènes sur- et sous-exprimés pour les 3 contrastes :
    - gene_list_HER2_vs_NBS_up.txt
    - gene_list_HER2_vs_NBS_dn.txt
    - gene_list_TNBC_vs_NBS_up.txt
    - gene_list_TNBC_vs_NBS_dn.txt
    - gene_list_NTNBC_vs_NBS_up.txt
    - gene_list_NTNBC_vs_NBS_dn.txt
- Fichiers MA-plot : ces fichiers contiennent en ligne les gènes, et en colonnes, les valeurs **M**, **A**, ainsi que les p.value et p.value ajustée.
    - dataMAplot_HER2_vs_NBS.txt

- dataMAplot_TNBC_vs_NBS.txt
- dataMAplot_NTNBC_vs_NBS.txt

note : pour répondre aux questions de ce TP, il faudra charger ces fichiers dans R. Pour cela, n'oubliez pas les options `header = TRUE`, `row.names = 1` et `sep = "\t"`.

1. Un peu de Biologie

Question 1.1: Trouvez une brève définition des 3 sous types de cancer. Identifiez pour chaque sous-type, le ou les gènes caractéristiques. Ces informations sont définies dans les 2 articles indiqués sur le site web du TP.

Question 1.2:

- Retrouvez les gènes cités précédemment dans notre jeu de données (différents noms peuvent exister pour un même gène = alias. Les noms de gènes sont appelés des *HGNC Gene Symbol*).
- Création des graphiques permettant de visualiser les mesures d'expression de ces gènes : Créez un graphique par gène avec en abscisse les 4 conditions et en ordonnée les comptages normalisés. Supprimer les labels des abscisses en précisant l'argument `xaxt='none'`. Vous pouvez ensuite personnaliser les labels des abscisses grâce à la fonction `axis()`.
  Cela vous parait-il cohérent avec ce que l'on pourrait attendre ?

2 Analyse fonctionnelle : sur-représentation de catégories fonctionnelles

Question 2.1 : Réalisez une analyse de sur-représentation de catégories fonctionnelles pour les gènes significativement différentiellement sur- et sous-exprimés au seuil de logFC de 1 pour le sous type HER2. * Commencez par ouvrir le fichier `dataMAplot_HER2_vs_NBS.txt` dans Excel. * Selectionnez la liste des gènes sur et sous-exprimés en utilisant un seuil de `logFC > 1 ou <= -1` et une `p.value ajustée <= 0.01`. * En utilisant l'outil ToppFun (https://toppgene.cchmc.org/enrichment.jsp), choississez les identifiants HGNC Symbol (*Entry type*) et copiez/collez votre liste de gènes d'interêt, puis lancez l'analyse. L'application vous indique alors le nombre de symbols retrouvés dans sa propre base de connaissance. Dans le tableau listant les bases de données à interroger, sélectionnez :

- GO:Biological Process
- Pathway : BioSystems : KEGG
- Disease

Vous pouvez laisser les options de calcul par défaut (mode de correction FDR, cutoff de la p.value à 0.05).

Question 2.2 : Quels sont les processus biologiques `GO` les plus récurrents parmis l'ensemble des 34 premiers termes significatifs (donnez quelques mot-clefs) ?

Question 2.3 : Quelles sont les trois voies de signalisation de la base de données **KEGG** significativement enrichies ? En affichant plus de données `Show XX more annotations`, quelles voies de signalisation pourraient être intéressantes compte-tenu du contexte biologique ?

Question 2.4 : Que vous suggère le tableau n°3 ?

3 Visualisation de Pathways

Nous vous proposons de visualiser en détail la modulation des gènes de voies de signalisation.

Installez le package Bioconductor `pathview` :

```
source('http://www.bioconductor.org/biocLite.R')
biocLite("pathview")
```

note : pour comprendre en detail le fonctionnement de ce package, vous pouvez visualiser la "vignette" en utilisant le code suivant : `> browseVignettes(package = "pathview")`.

Question 3.1 : Quelles sont les gènes différentiellement exprimés au seuil de p.values ajustées de 1% et au seuil de valeur absolue de logFC de 1 ? Construisez une matrice d'une colonne contenant ces logFC. Les noms de lignes de cette matrice seront les noms des gènes.

Question 3.2 : A l'aide de la fonction `pathview()` du package que vous venez d'installer, et de la matrice de gène précédemment construite, visualisez les *pathways* suivant (identifant : nom) :

- hsa05224 : Breast cancer
- hsa04012 : ERBB signaling pathway
- hsa05200 : Pathway in cancer

```
library(pathview)
pathview(genes, pathway.id = "hsa05224", gene.idtype = "SYMBOL")
pathview(genes, pathway.id = "hsa04012", gene.idtype = "SYMBOL")
pathview(genes, pathway.id = "hsa05200", gene.idtype = "SYMBOL")
```