

TP3 Analyse différentielle

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'expressions, 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-jacents, 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 précédent, vous aviez généré les MA-plot, des volcanoplots et des heatmap 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 dans 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 paraît-il cohérent avec ce que l'on pourrait attendre ?

2 Analyse fonctionnelle : sur-representation de catégories fonctionnelles

Question 2.1 : Réaliser une analyse de sur-représentation de catégories fonctionnelles pour les gènes significativement différentiellement surexprimés dans le sous type HER2. * Commencez par ouvrir le fichier `dataMAplot_HER2_vs_NBS.txt` dans excel * Sélectionner la liste des gènes sur et sous-exprimés en utilisant un seuil de $\log_{2}FC > 1$ ou ≤ -1 et une *p.value ajustée* ≤ 0.01 . en utilisant l'outil ToppFun (<https://toppgene.cchmc.org/enrichment.jsp>). Choisissez les identifiants HGNC Symbol (*Entry type*) et copier/coller votre liste de gènes d'intérêt, et lancer l'analyse. Le site vous indique alors le nombre de symbols retrouvés dans sa propre base de connaissance. Dans le tableau listant les base de données à interroger, sélectionnez :

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

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

Question 2.2 : Quels sont les processus biologiques les plus récurrents (donnez quelques mot-clefs) ?

Question 2.3 : Quelles sont les trois voies de signalisation de la base de données **KEGG** significativement enrichie ? En affichant plus de données **Show 16 more annotations**, quelles voies de signalisation pourrait 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` :

1. Sourcez (commande `source`) le script `biocLite.R`, qui se trouve à l'URL <http://www.bioconductor.org/biocLite.R>.
2. Utilisez le script pour installez le package `pathview` : `> biocLite("pathview")`. Le script installe automatiquement les dépendances du package.

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

note : pour comprendre en détail el fonctionnement de ce package, vous pouvez visualisez la “vignette” en utilisant le code suivant : `> browseVignettes(package = "pathview")`

Question 3.1 : Quelles sont les gènes différentiellement exprimées au seuil de p-valeurs ajustées de 1% et de la valeur absolue de logFC de 1 ? Construisez une matrice d’une colonne contenant ces logFC. Les noms de rang de cette matrice seront les nom des gènes.

```
updown = read.table("data_TP4/dataMAplot_HER2_vs_NBS.txt", sep="\t", header=TRUE, row.names=1)
IDX = intersect(which(updown$Adj.p.value < 0.01), which(abs(updown$M) > 1))
genes = as.matrix(updown[IDX, "M"])
rownames(genes) = rownames(updown)[IDX]
```

Question 3.2 : A l’aide de la fonction `pathview()` du package que vous venez d’installer, visualisez les *pathways* suivant (identifiant : nom) :

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

chez l’Homme (**hsa**) de la base de connaissance **KEGG**, en affichant le logFC pour les gènes différentiellement exprimés sélectionné précédemment.

```
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")
```

Question 3.3 : Quel gène semble jouer en rôle central dans le contrôle de cette voie de signalisation ? Est-il différentiellement exprimé dans votre analyse ? Quelle hypothèse pouvez-vous faire pour expliquer cette observation ?