T4_Sobrerrepresentacion

Francisco Martínez Picó

7/14/2020

En esta viñeta se muestra el **análisis de sobrerrepresentación** de grupos de genes, tanto del experimento de *Microarrays (gse69762)* como del experimento de *RNAseq (PRJNA601724)*. Realizaremos lo que es conocido como un test hipergeométrico o, más clásicamente, **test de Fisher unilateral**.

Cargando paquetes

Antes de empezar, cargamos los paquetes necesarios:

- *Biobase*: este paquete para trabajar con Bioconductor contiene estructuras estandarizadas de datos para representar información genómica.
- *EnrichmentBrowser*: utilizado para la descarga de grupos de genes de las BBDD que queramos (en este caso, KEGG) así como el análisis de sobre-representación.
  - *SummarizedExperiment*: paquete utilizado para trabajar con el objeto RangedSummarizedExperiment, que contendrá toda la información de nuestro experimento de RNAseq.
- *org.Ce.eg.db*: para realizar la anotación de los genes del experimento de RNAseq. Este experimento se realiza sobre *C. elegans*. Los ID incluidos son los de WormBase y necesitamos los ENTREZ.

```
pacman::p_load(Biobase)
pacman::p_load(SummarizedExperiment)
pacman::p_load(EnrichmentBrowser)
pacman::p_load(org.Ce.eg.db)
```

Microarray

Cargamos los datos del experimento de Microarrays **gse69762**, correspondientes a la *Tarea 1*.

```
data(gse69762, package = 'franciscomartinez')
```

Descargamos los grupos de genes para *Homo sapiens* de la base de datos GeneOntology.

```
hsaGOgsc = getGenesets(org = 'hsa', db = 'go') # Bajamos info de grupos de genes para Homo sapiens y la
```

Para realizar este análisis de sobre-representación se utilizará el paquete **EnrichmentBrowser**.

```
se69762 = makeSummarizedExperimentFromExpressionSet(gse69762) # Si ya tenemos SummarizedExperiment no s

se69762 = probe2gene(se69762) # Se cambian los identificadores por los ENTREZID. Además introduce la va

se69762 = deAna(expr = se69762) # t-test moderado

se69762.oraGO = sbea(method = "ora", se = se69762, gs = hsaGOgsc,
                     perm = 0, alpha = 0.2) # Análisis de la sobrerrepresentación con un test de Fisher

gsRanking(se69762.oraGO) # Visualizar.
```

RNAseq

Cargamos los datos del experimento de RNAseq **PRJNA601724**, correspondientes a la *Tarea 3*.

```
data('PRJNA601724', package = 'franciscomartinez')
```

A continuación, descargamos los grupos de genes. De nuevo realizamos esta descarga utilizando la función getGenesets del paquete EnrichmentBrowser de la base de datos KEGG, pero en esta ocasión para el organismo objeto de nuestro estudio (*C. elegans*).

```
celKEGGgsc = getGenesets(org = 'cel', db = 'kegg') # Para C. elegans y la base de datos KEGG.
```

Como tenemos tres grupos y los métodos están preparados para trabajar únicamente con dos, decidimos trabajar arbitrariamente sólo con los grupos de *E.coli* y *Chryseobacterium*.

```
sel = colData(PRJNA601724)[,"Treatment"] == "E.coli" |
  colData(PRJNA601724)[,"Treatment"] == "Chryseobacterium"

sel = which(sel)

nuevo_se = PRJNA601724[,sel]
```

Seguidamente pasamos a añadir los ENTREZ ID a nuestros datos, ya que con los ID de WormBase (los cuales ya están incluídos) no es suficiente. Esto es debido a que la información de grupos de genes descargada de KEGG viene con los ID de ENTREZ.

```
# Para añadir más información de anotación de los genes:
genesInfo = AnnotationDbi::select(org.Ce.eg.db, keys = rownames(nuevo_se), columns = c("ENTREZID", "SYM

# Nos quedamos con la primera coincidencia de WormBase:
posiciones = match(unique(genesInfo[,1]),genesInfo[,1]) # Para WORMBASE
genesInfo = genesInfo[posiciones,]
```

Y de nuevo utilzamos el paquete EnrichmentBrowser para llevar a cabo el análisis.

```
se601724 = probe2gene(probeSE = nuevo_se, chip = org.Ce.eg.db) # Se cambian los identificadores por los

se69762 = deAna(expr = se69762) # t-test moderado

se69762.oraGO = sbea(method = "ora", se = se69762, gs = hsaGOgsc,
                     perm = 0, alpha = 0.2) # Análisis de la sobrerrepresentación con un test de Fish

gsRanking(se69762.oraGO) # Visualizar.
```

De esta forma se finaliza el análisis de la sobrerrepresentación de grupos de genes.