PEC2 Análisis de Datos Ómicos

Rita Ortega Vallbona

6 de junio, 2020

Contents

1 Abstract

2 Objetivos

3 Materiales y métodos

3.1 Los Datos

Los datos proporcionados en el enunciado provienen del repositorio GTEx (*Genotype-Tissue Expression*), que recoge información de expresión específica de 54 tipos de tejido sano, proveniente de 1000 individuos ("GTEx Portal," n.d.). Este **portal** permite el acceso a los datos de expresión, imágenes de histología, etc.

Obtenemos los datos de targets y counts de los archivos csv proporcionados en el enunciado: targets.csv y counts.csv.

Estos son datos de expresión (RNA-seq) pertenecientes a un análisis del tiroides, donde se comparan tres tipos de infiltración en 292 muestras:

- *Not infiltrated tissues* (**NIT**): 236 muestras
- *Small focal infiltrates* (**SFI**): 42 muestras
- *Extensive lymphoid infiltrates* (**ELI**): 14 muestras

Con este script extraemos 10 muestras de cada grupo del archivo targets.csv y subseteamos las columnas escogidas en el archivo counts.csv:

```
> # Separamos el dataframe que recoge los targets por grupos
> NIT <- subset(all_targets, Group == "NIT")
> SFI <- subset(all_targets, Group == "SFI")
> ELI <- subset(all_targets, Group == "ELI")
>
> # Seleccionamos 10 muestras de cada grupo y las unimos en un
> # único #dataframe que recoge los targets con los que
> # trabajaremos
> set.seed(params$seed.extract)
> NIT10 <- NIT[sample(nrow(NIT), size = 10, replace = FALSE), ]
> SFI10 <- SFI[sample(nrow(SFI), size = 10, replace = FALSE), ]
> ELI10 <- ELI[sample(nrow(ELI), size = 10, replace = FALSE), ]
>
> mytargets <- rbind(NIT10, SFI10, ELI10, deparse.level = 0)
>
> # Extraemos los nombres de las muestras y cambiamos los
> # guiones #por puntos para que coincidan con los nombres de
> # las muestras en #el dataframe de counts
> sample_names <- mytargets[, 3]
> s_names <- gsub("-", ".", sample_names)
>
> # Subseteamos las columans escogidas del dataframe de counts
> mycounts <- dplyr::select(all_counts, s_names)
> row.names(mycounts) <- all_counts$X
```

De este modo hemos obtenido dos datasets: `mytargets` que recoge los detalles de cada una de las 30 muestras con las que vamos a trabajar, y `mycounts`, que representa la tabla de contajes de estas 30 muestras.

3.2 Preprocesado de los datos

Para poder identificar los tipos de muestra con mayor facilidad, procedemos a renombrar las columnas de `mycounts` con los nombres cortos asignados en la tabla `mytargets`.

```
> # Primero cambiamos los guiones por puntos
> newShortNames <- gsub("_", ".", mytargets$ShortName)
> mytargets$ShortName <- gsub("-", "", newShortNames)
>
> # Asignamos los nombres cortos a las columnas de mycounts
> colnames(mycounts) <- mytargets$ShortName
```

Antes de proceder con el análisis de los datos, debemos evaluar la calidad de los datos crudos con los que vamos a trabajar.

3.2.1 Filtraje

3.2.2 Normalización

3.3 Identificación de genes diferencialmente expresados

3.4 Anotación de los resultados

3.5 Busca de patrones de expresión y agrupación de las muestras (comparación entre las distintas comparaciones)

3.6 Análisis de significación biológica ("*Gene Enrichment Analysis*")

4 Resultados

5 Discusión

6 Apéndice

Bibliografía

"GTEx Portal." n.d. https://www.gtexportal.org/home/.