Análisis de datos de RNA-seq
PAC2
https://github.com/Turing78/ADO_PEC2

Anna Casademunt

26/5/2020

Contents

Abstract **2**

Objetivos **2**

Materiales y Métodos **2**
1. Datos, tipo y diseño experimental . . . 2
1.1. Naturaleza de los datos . . . 2
1.2. Tipo de experimento . . . 2
1.3. Diseño experimental . . . 2
2. Procedimiento y métodos . . . 3
2.1. Procedimiento general de análisis . . . 3
2.2. Software utilizado . . . 3
3. Pipeline . . . 4
3.1. Importar y seleccionar los datos. . . . 4
3.2. Definir los modelos genéticos . . . 5
3.3. El objeto DESeqDataSet y la fórmula de diseño . . . 5
3.4. Análisis exploratorio y visualización . . . 5
3.5. Análisis de expresión diferencial . . . 7
3.6. Representando los resultados . . . 10
3.7. Anotación y exportación de los resultados . . . 14
3.8. Significación biológica . . . 16
3.9. Eliminando efectos batch ocultos . . . 18

4. Resultados **26**
Gráficos . . . 26
Ficheros . . . 26

5. Discusión **27**

6. Bibliografía **27**

Abstract

Disponemos de muestras de tejido tiroideo procedentes del repositorio GTEx, ver Lonsdale et al. (2013), y pertenecientes a tres grupos: "Non infiltrated tissues" (NIT), "Small local infiltrates" (SFI), "Extensive lymphoid infiltrates" (ELI). Seleccionando 10 muestras de cada grupo, analizaremos los genes diferencialmente expresados separando por grupo y sexo sin considerar el efecto batch, y por grupo eliminando el efecto batch oculto.

Objetivos

Realizaremos un análisis de los datos de expresión (RNA-Seq) para encontrar genes diferencialmente expresados (DEG) entre dos factores:grupo y sexo. Luego repetiremos el mismo análisis por grupo añadiendo al diseño las correcciones SVA para eliminar el efecto batch oculto.

Materiales y Métodos

1. Datos, tipo y diseño experimental

1.1. Naturaleza de los datos

Los datos están preprocesados en una tabla de counts *counts.csv* con 56202 observaciones y 292 variables que corresponden a:

- NIT: 236 muestras
- SFI: 42 muestras
- ELI: 14 muestras

y una tabla *targets.csv* con 292 observaciones de 9 variables, que corresponden a los datos de cada muestra:

- Experiment
- SRA_Sample
- Sample_Name
- Grupo_analisis: 1-NIT, 2-SFI y 3-ELI
- body_site: Thyroid
- molecular_data_type. "Allele-Specific Expression" o "RNA SEQ(NGS)"
- sex: "male" o "female"
- Group: "NIT", "SFI" o "ELI"
- Short_name

Escogeremos aleatoriamente 10 muestras de cada grupo de análisis. Los datos de tipo molecular "Allele-Specific Expression" se tratarán como si fueran de "RNA SEQ(NGS)".

1.2. Tipo de experimento

El tipo de experimento es un análisis de genes diferencialmente expresados de RNA-seq.

1.3. Diseño experimental

Para el diseño del experimento se considerarán dos factores:

- *Grupo-analisis* que renombramos a simplemente grupo y que tiene tres niveles: el nivel de control NIT (tejido no infiltrado), el nivel medio SFI (infiltración pequeña local) y el nivel alto ELI (infiltración extensiva de linfoides). Consideraremos las tres comparaciones posibles: **SFI vs NIT**, **ELI vs NIT** y **ELI vs SFI**.
- *Sexo*, comparación entre los dos niveles, femenino vs masculino, que pueden tener distinta expresión en el tejido tiroideo. En el gráfico [3.4.4. PCA plot] se ve una clara diferenciación.

2. Procedimiento y métodos

2.1. Procedimiento general de análisis

El análisis se ha realizado en R-Studio siguiendo los pasos del estudio https://github.com/aspteaching/omics_data_analysis-case_study_2-rna-seq:

1. Importar y seleccionar los datos
2. Definir los modelos genéticos
3. El objeto DESeqDataSet y la fórmula de diseño
4. Análisis exploratorio y visualización
5. Análisis de expresión diferencial
6. Representando los resultados
7. Anotación y exportación de los resultados
8. Significación biológica
9. Eliminando efectos batch ocultos

- 9.1. Usando SVA con DESeq2
- 9.2. Análisis de expresión diferencial
- 9.3. Representando los resultados
- 9.4. Anotación y exportación de los resultados
- 9.5. Significación biológica

El código R no se muestra en el documento, pero puede encontrarse en https://github.com/Turing78/ADO_PEC2.git.

2.2. Software utilizado

Se han instalado los paquetes para el análisis de RNA-seq que aparecen a continuación.

```
if(!require(EnsDb.Hsapiens.v75)) BiocManager::install("EnsDb.Hsapiens.v75")
if(!require(DESeq2)) BiocManager::install("DESeq2")
if(!require(apeglm)) BiocManager::install("apeglm")
if(!require(BiocParallel)) BiocManager::install("BiocParallel")
if(!require(genefilter)) BiocManager::install("genefilter")
if(!require(org.Hs.eg.db)) BiocManager::install("org.Hs.eg.db")
if(!require(AnnotationDbi)) BiocManager::install("AnnotationDbi")
if(!require(sva)) BiocManager::install("sva")
if(!require(Gviz)) BiocManager::install("Gviz")
if(!require(limma)) BiocManager::install("limma")
if(!require(biomaRt)) BiocManager::install("biomaRt")
if(!require(clusterProfiler)) BiocManager::install("clusterProfiler")
if(!require(enrichplot)) BiocManager::install("enrichplot")

if(!require(stringr)) install.packages("stringr")
if(!require(dplyr)) install.packages("dplyr", dep=TRUE)
if(!require(ggplot2)) install.packages("ggplot2", dep=TRUE)
if(!require(pheatmap)) install.packages("pheatmap", dep=TRUE)
if(!require(RColorBrewer)) install.packages("RColorBrewer", dep=TRUE)
if(!require(ggbeeswarm)) install.packages("ggbeeswarm", dep=TRUE)
if(!require(kableExtra)) install.packages("kableExtra", dep=TRUE)
if(!require(tibble)) install.packages("tibble", dep=TRUE)
if(!require(knitr)) install.packages("knitr", dep=TRUE)
```

3. Pipeline

3.1. Importar y seleccionar los datos.

Importamos los datos del fichero targets y seleccionamos, con la semilla 5154 (parte de mi DNI), 10 muestras de cada grupo de análisis (NIT, SFI y ELI) usando la función *sample*. Lo guardamos en el data frame **subTargets**.

A continuación leemos el fichero counts. Vemos el detalle de las 6 primeras columnas con *str*.

```
## 'data.frame':    56202 obs. of  6 variables:
##  $ GTEX.111CU.0226.SM.5GZXC: int  7 401 4 2 0 0 0 6 16 744 ...
##  $ GTEX.111FC.1026.SM.5GZX1: int  0 1064 0 0 0 1 1 2 8 1442 ...
##  $ GTEX.111VG.0526.SM.5N9BW: int  1 474 1 0 1 1 0 3 7 427 ...
##  $ GTEX.111YS.0726.SM.5GZY8: int  4 395 2 1 0 1 0 3 21 922 ...
##  $ GTEX.1122O.0226.SM.5N9DA: int  2 732 1 1 0 0 1 16 8 1021 ...
##  $ GTEX.1128S.0126.SM.5H12S: int  2 631 0 0 0 0 1 4 16 1053 ...
```

Las columnas de counts corresponden con la columna Sample_names en los targets, aunque los separadores son puntos en lugar de guiones. Sustituimos con *str_replace_all* y generamos el data frame **Datos** con las 30 columnas seleccionadas en el paso anterior. Sustituimos los nombres de las columnas por el valor *SRA_Sample* de **SubTargets**. Mostramos la cabecera de **Datos**.

```
##                   SRS627910 SRS634350 SRS629372 SRS631250 SRS648636 SRS629440
## ENSG00000223972.4         4         4         3         5         1         1
## ENSG00000227232.4       954       700       580       658      1042       524
## ENSG00000243485.2         0         0         0         3         1         1
## ENSG00000237613.2         0         3         1         0         4         0
## ENSG00000268020.2         0         1         1         0         0         0
## ENSG00000240361.1         0         0         0         1         2         0
##                   SRS644461 SRS633874 SRS644565 SRS634136 SRS627158 SRS629299
## ENSG00000223972.4         1         0         2         2         2         1
## ENSG00000227232.4      1051       331       468       754       423       775
## ENSG00000243485.2         3         1         1         3         0         2
## ENSG00000237613.2         2         0         0         2         0         0
## ENSG00000268020.2         0         0         0         0         2         0
## ENSG00000240361.1         0         2         1         1         1         0
##                   SRS631283 SRS639491 SRS631169 SRS644736 SRS374975 SRS333099
## ENSG00000223972.4         4         0         0         0         3         5
## ENSG00000227232.4      1325       834      1002       419       134       489
## ENSG00000243485.2         1         1         1         0         1         1
## ENSG00000237613.2         0         1         0         1         2         3
## ENSG00000268020.2         2         0         0         0         1         2
## ENSG00000240361.1         1         0         1         0         0         1
##                   SRS638114 SRS648152 SRS644012 SRS374817 SRS629456 SRS639261
## ENSG00000223972.4         0         3         3         3         3         3
## ENSG00000227232.4       825      1301       460       406       426       369
## ENSG00000243485.2         1         1         0         4         1         1
## ENSG00000237613.2         0         0         1         1         1         3
## ENSG00000268020.2         0         0         2         0         1         1
## ENSG00000240361.1         1         1         0         1         2         2
##                   SRS623875 SRS624031 SRS637292 SRS408366 SRS374961 SRS389523
## ENSG00000223972.4         1         0         2         6         9         8
## ENSG00000227232.4       483       633       689       820       302      1388
## ENSG00000243485.2         0         2         2         0         4         1
## ENSG00000237613.2         0         1         4         1         2         1
## ENSG00000268020.2         0         0         0         0         2         1
```

```
## ENSG00000240361.1         0         1         2         4         0         0
```

Vemos que el nombre del tránscrito (o nombre de fila) termina con un punto y un número de versión. Lo eliminamos tal y como se ha sugerido en el foro de la PAC.

3.2. Definir los modelos genéticos

Para conseguir los datos de los exones de cada tránscrito usamos la librería **EnsDb.Hsapiens.v75**. La función *exonBy* filtra los valores requeridos a partir de los nombres de las filas de **Datos**.

Observamos que algunos exones no se han encontrado. Comprobamos con la función *which.max* a lo largo de cada fila que los counts en estos tránscritos no superan el 1, por lo que podemos eliminarlos.

3.3. El objeto DESeqDataSet y la fórmula de diseño

Cambiamos los nombres de las variables "Grupo_analisis" a "grupo" y "sex" a "sexo" y las factorizamos de forma que el grupo de referencia es *NIT* y el sexo de referencia *male*. Construimos el objeto DESeqDataSet a partir de **Datos**, **subTargets** y la fórmula de diseño: $\sim$ *grupo + sexo*. Le añadimos la información de los exones obtenida en el apartado anterior como *rowRanges*. Vemos con una tabla que el número de casos está equilibrado por sexos.

```
##
##       male female
##   NIT    5      5
##   SFI    6      4
##   ELI    4      6
```

3.4. Análisis exploratorio y visualización

Hay dos partes en este pipeline. En el primero se realizan transformaciones de los counts para explorar visualmente las relaciones entre las muestras. En la segunda parte se vuelve a los counts crudos originales para el testing estadístico. Esto es crítico ya que los métodos estadísticos dependen de los datos de counts originales para calcular la precisión de las medidas.

3.4.1. Pre-filtrando el dataset

El pre-filtrado no es estrictamente necesario pero eliminando los counts bajos se reduce el tiempo de computación de las transformaciones siguientes. Filtramos un mínimo de 10 counts.

Pasamos de 56156 a 35786 filas a considerar.

3.4.2. La transformación estabilizando la varianza

Aplicamos la función *vst* (variance stabilizing transformation), que es mucho más rápida de calcular y menos sensible a outliers de counts altos que la función *rlog* y está recomendada para datasets grandes. Obtenemos el objeto **vsd**.

3.4.3. Distancias entre las muestras

Visualizamos las distancias en **vsd** en el heatmap siguiente, usando la función *pheatmap*. Las distancias entre las muestras se proveen a la función manualmente. A menor distancia, más fuerte es el tono de azul. Así vemos un primer cuadro de elementos cercanos formado principalmente por elementos del grupo ELI, y otro cuadro mayor con SFI y NIT.

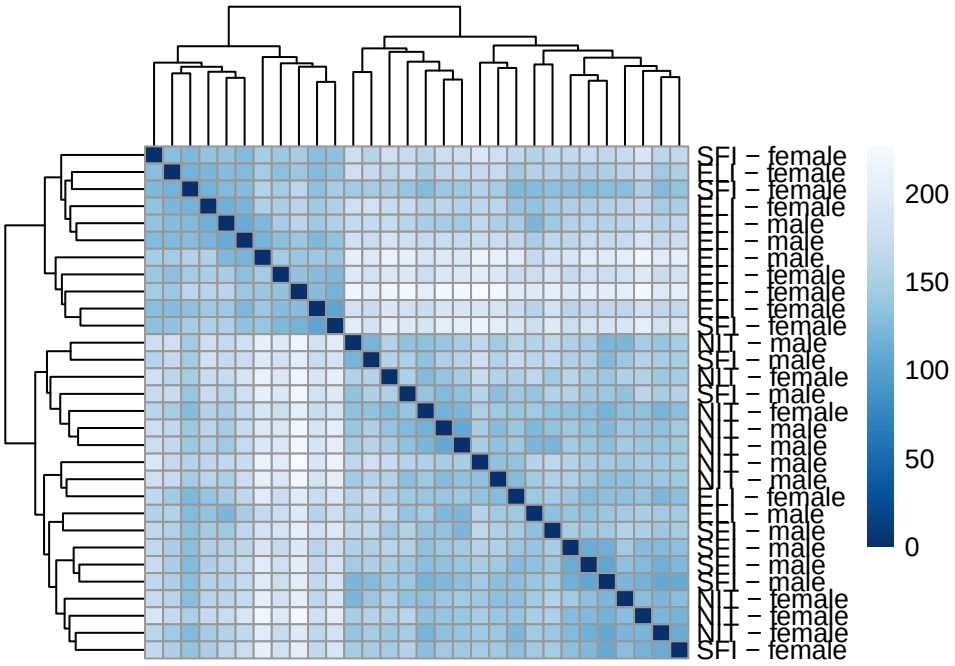

Figure 1: Pheatmap showing the sample distances

3.4.4. Gráfico PCA

Hacemos un gráfico PCA con los datos **vsd**. Cada combinación de grupo tiene un color y cada sexo una forma. Se ve una clara diferenciación en ambos factores.

```
## [1] "data.frame"
```

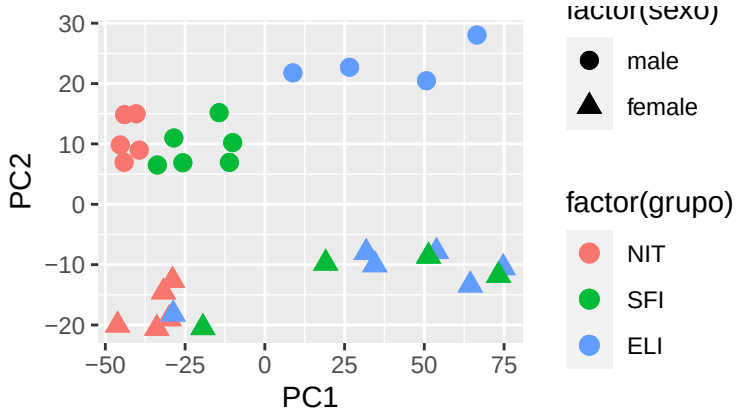

Figure 2: Principal Components Analisis Plot

3.5. Análisis de expresión diferencial

Ejecutamos el pipeline de expresión diferencial sobre los counts crudos con la función *DESeq*. Tiene sentido elevar en este caso el log2 fold change threshold para filtrar los valores más significativos.

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates: 2 workers

## mean-dispersion relationship

## final dispersion estimates, fitting model and testing: 2 workers
```

Mostramos el sumario de la comparación SFI vs NIT:

```
##
## out of 35786 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0.50 (up)    : 273, 0.76%
## LFC < -0.50 (down) : 3, 0.0084%
## outliers [1]       : 178, 0.5%
## low counts [2]     : 5539, 15%
## (mean count < 1)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

Mostramos el sumario de la comparación ELI vs NIT:

```
##
## out of 35786 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0.50 (up)    : 1265, 3.5%
## LFC < -0.50 (down) : 61, 0.17%
## outliers [1]       : 178, 0.5%
## low counts [2]     : 4156, 12%
## (mean count < 1)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

Y finalmente el de ELI vs SFI:

```
##
## out of 35786 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0.50 (up)    : 14, 0.039%
## LFC < -0.50 (down) : 5, 0.014%
## outliers [1]       : 178, 0.5%
## low counts [2]     : 11728, 33%
## (mean count < 5)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

Si consideramos una fracción de falsos positivos del 5% acceptable podemos considerar todos los genes con un p-valor ajustado bajo 0.05 como significativos. Los genes que quedan de cada comparación son los siguientes:

SFI_NIT	ELI_NIT	ELI_SFI
243	1154	13

3.5.1. SFI vs NIT

Mostramos los genes significativos de los que se ha reducido más la expresión:

```
## # A tibble: 6 x 7
##   EnsemblID       baseMean log2FoldChange lfcSE  stat   pvalue   padj
##   <chr>              <dbl>          <dbl> <dbl> <dbl>    <dbl>  <dbl>
## 1 ENSG00000199315  11922.          -2.63 0.581 -3.67 0.000240 0.0309
## 2 ENSG00000222297    174.          -2.45 0.509 -3.83 0.000126 0.0176
## 3 ENSG00000250536     61.9         -2.20 0.466 -3.65 0.000259 0.0330
## 4 ENSG00000269619   2252.           2.44 0.538  3.61 0.000310 0.0391
## 5 ENSG00000143674    321.           2.47 0.554  3.55 0.000382 0.0473
## 6 ENSG00000237668     92.0          2.53 0.563  3.60 0.000316 0.0396
```

Y aquellos en los que la expresión se ha incrementado más:

```
## # A tibble: 6 x 7
##   EnsemblID       baseMean log2FoldChange lfcSE  stat   pvalue     padj
##   <chr>              <dbl>          <dbl> <dbl> <dbl>    <dbl>    <dbl>
## 1 ENSG00000243970    430.           25.8  1.46  17.4 9.81e-68 2.95e-63
## 2 ENSG00000133138     56.3          22.1  1.37  15.8 3.90e-56 5.86e-52
## 3 ENSG00000133055     85.4          22.0  1.70  12.7 8.59e-37 3.69e-33
## 4 ENSG00000133247     62.2          22.0  1.47  14.6 2.20e-48 2.21e-44
## 5 ENSG00000133107     36.1          21.3  1.48  14.0 9.24e-45 5.56e-41
## 6 ENSG00000243771     19.9          21.2  1.69  12.3 1.21e-34 4.05e-31
```

3.5.2. ELI vs NIT

Mostramos los genes significativos de los que se ha reducido más la expresión:

```
## # A tibble: 6 x 7
##   EnsemblID       baseMean log2FoldChange lfcSE  stat        pvalue        padj
##   <chr>              <dbl>          <dbl> <dbl> <dbl>         <dbl>       <dbl>
## 1 ENSG00000229957   5095.           -8.74 1.43  -5.75 0.00000000879 0.000000933
## 2 ENSG00000258476     75.9          -8.43 1.79  -4.44 0.00000897    0.000476
## 3 ENSG00000254591    359.           -5.42 1.07  -4.60 0.00000416    0.000241
## 4 ENSG00000242284    332.           -4.45 0.997 -3.96 0.0000746     0.00314
## 5 ENSG00000145700    103.           -3.74 0.861 -3.77 0.000166      0.00635
## 6 ENSG00000166796     19.8          -3.70 0.918 -3.49 0.000487      0.0160
```

Y aquellos en los que la expresión se ha incrementado más:

```
## # A tibble: 6 x 7
##   EnsemblID       baseMean log2FoldChange lfcSE  stat   pvalue     padj
##   <chr>              <dbl>          <dbl> <dbl> <dbl>    <dbl>    <dbl>
## 1 ENSG00000243970    430.           25.8  1.46 17.4  9.73e-68 1.53e-63
## 2 ENSG00000133055     85.4          25.7  1.69 14.9  4.88e-50 2.56e-46
## 3 ENSG00000133247     62.2          24.8  1.46 16.6  5.11e-62 5.36e-58
## 4 ENSG00000243749     25.2          24.4  1.70 14.0  8.19e-45 2.86e-41
## 5 ENSG00000133138     56.3          24.3  1.37 17.4  5.68e-68 1.53e-63
## 6 ENSG00000266813     16.9          23.9  2.51  9.34 1.01e-20 5.47e-18
```

3.5.3. ELI vs SFI

Mostramos los genes significativos de los que se ha reducido más la expresión:

```
## # A tibble: 6 x 7
##   EnsemblID       baseMean log2FoldChange lfcSE  stat      pvalue    padj
##   <chr>              <dbl>          <dbl> <dbl> <dbl>       <dbl>   <dbl>
```

```
## 1 ENSG00000229957   5095.           -7.61 1.45  -4.92 0.000000880 0.00629
## 2 ENSG00000227108     58.2          -3.40 0.677 -4.28 0.0000188   0.0407
## 3 ENSG00000175800   1505.           -3.23 0.626 -4.37 0.0000126   0.0360
## 4 ENSG00000264831     35.2           2.52 0.467  4.33 0.0000147   0.0360
## 5 ENSG00000119912    189.            2.95 0.538  4.55 0.00000525  0.0209
## 6 ENSG00000260511     31.6           3.14 0.620  4.25 0.0000212   0.0423
```

Y aquellos en los que la expresión se ha incrementado más:

```
## # A tibble: 6 x 7
##   EnsemblID       baseMean log2FoldChange lfcSE  stat       pvalue     padj
##   <chr>              <dbl>          <dbl> <dbl> <dbl>        <dbl>    <dbl>
## 1 ENSG00000115204     21.4           6.24 1.05   5.49 0.0000000410 0.000978
## 2 ENSG00000259234    328.            4.21 0.800  4.64 0.00000343   0.0164
## 3 ENSG00000107099     15.0           4.04 0.802  4.41 0.0000102    0.0346
## 4 ENSG00000115363     77.2           4.04 0.705  5.02 0.000000516  0.00616
## 5 ENSG00000266867    140.            3.42 0.674  4.33 0.0000151    0.0360
## 6 ENSG00000260510   2211.            3.39 0.592  4.88 0.00000105   0.00629
```

3.5.4. Female vs Male

Mostramos el sumario de la comparación Female vs Male:

```
##
## out of 35786 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0.50 (up)    : 17, 0.048%
## LFC < -0.50 (down) : 34, 0.095%
## outliers [1]       : 178, 0.5%
## low counts [2]     : 4849, 14%
## (mean count < 1)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

Mostramos los genes significativos de los que se ha reducido más la expresión:

```
## # A tibble: 6 x 7
##   EnsemblID       baseMean log2FoldChange lfcSE  stat    pvalue      padj
##   <chr>              <dbl>          <dbl> <dbl> <dbl>     <dbl>     <dbl>
## 1 ENSG00000272567     159.         -10.3  0.583 -16.9 5.09e- 64 1.04e- 60
## 2 ENSG00000273017     204.         -10.2  0.562 -17.2 3.72e- 66 8.17e- 63
## 3 ENSG00000273018    3728.          -9.82 0.322 -28.9 4.69e-184 2.06e-180
## 4 ENSG00000272814     912.          -9.79 0.338 -27.5 1.08e-166 3.71e-163
## 5 ENSG00000273013    2063.          -9.75 0.279 -33.1 2.06e-240 1.58e-236
## 6 ENSG00000272797    2284.          -9.53 0.270 -33.4 1.46e-244 1.49e-240
```

Y aquellos en los que la expresión se ha incrementado más:

```
## # A tibble: 6 x 7
##   EnsemblID       baseMean log2FoldChange lfcSE  stat   pvalue     padj
##   <chr>              <dbl>          <dbl> <dbl> <dbl>    <dbl>    <dbl>
## 1 ENSG00000270442  20482.           11.2  0.236 45.6  0.       0.
## 2 ENSG00000270441     35.9           5.43 0.337 14.6  1.77e-48 3.41e-45
## 3 ENSG00000236292     18.3           4.59 0.971  4.21 2.56e- 5 2.18e- 2
## 4 ENSG00000266647    545.            4.17 0.940  3.91 9.37e- 5 6.69e- 2
## 5 ENSG00000166743    384.            4.13 0.749  4.85 1.23e- 6 1.30e- 3
## 6 ENSG00000243478   1184.            3.83 0.827  4.03 5.65e- 5 4.45e- 2
```

3.6. Representando los resultados

3.6.1. Counts-Plot

Realizamos un Counts-Plot para la comparación ELI vs NIT tras eliminar el efecto batch oculto. Se observa que el gen más significativo tiene un counts más elevado en SFI para mujeres que para hombres.

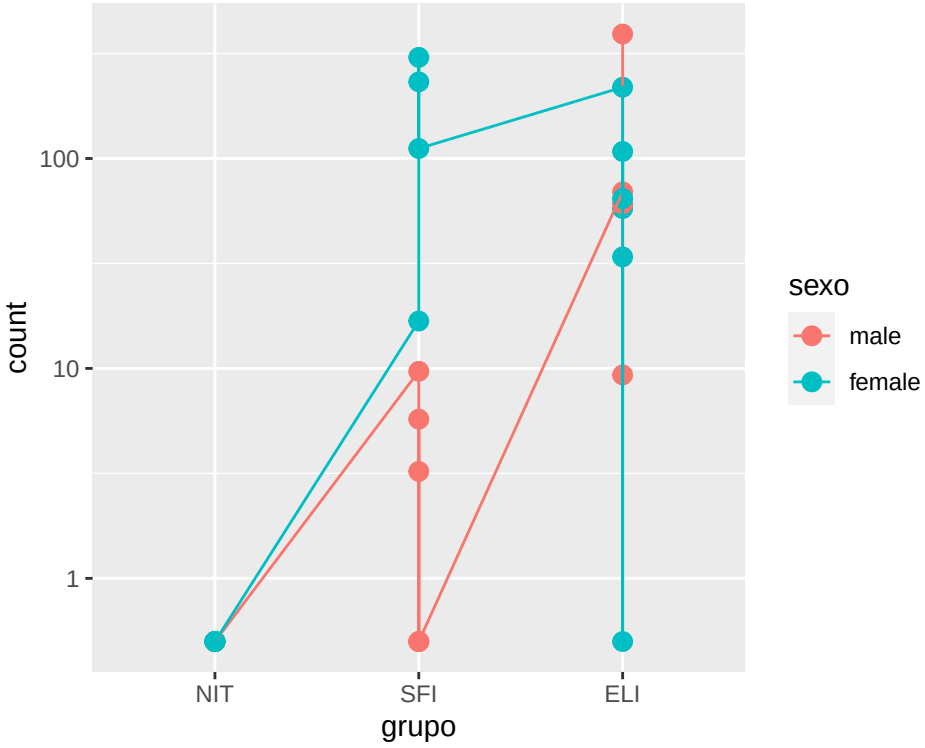

Figure 3: Counts-Plot for ELI vs NIT

3.6.2. Diagrama de Venn

Este diagrama nos sirve para ver las coincidencias entre los diferentes contrastes.

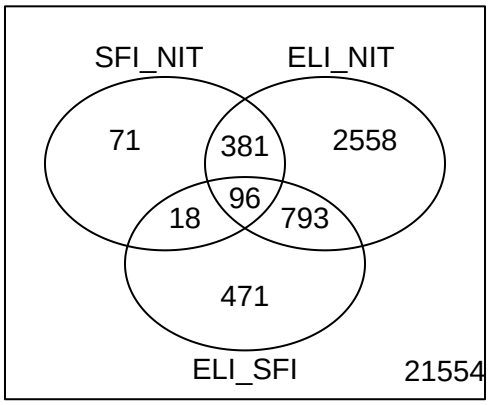

Figure 4: Venn's Diagram

3.6.3. Clustering de los genes

Seleccionamos los 20 genes con la varianza más alta, usando los datos **vsd** y hacemos un *pheatmap*.

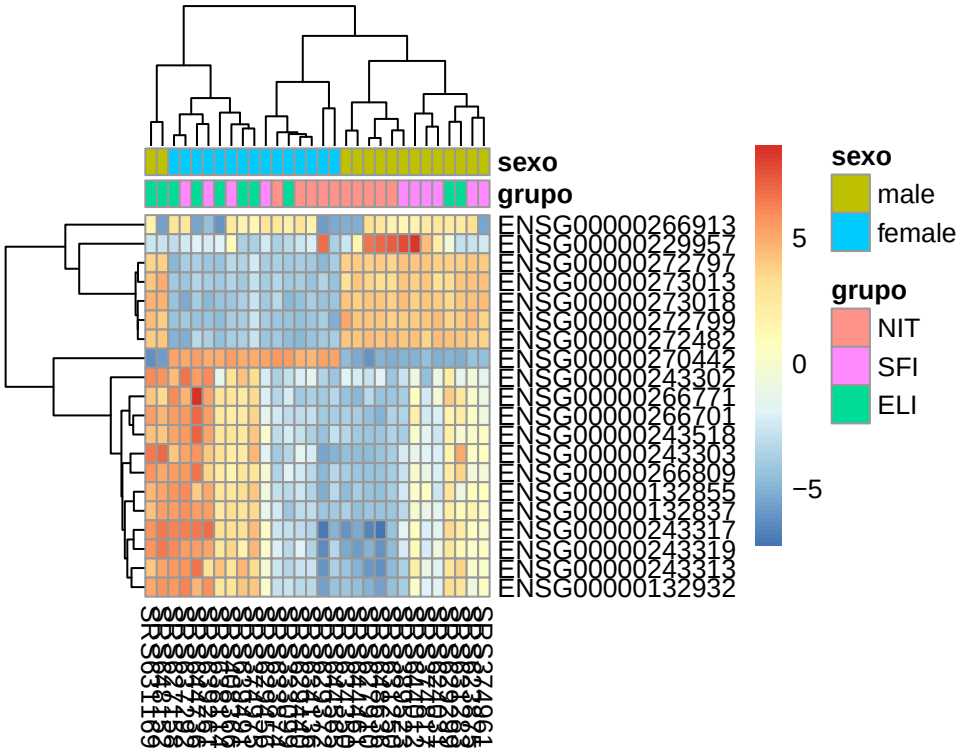

Figure 5: Pheatmap showing the gene clustering

3.6.4. Representando los cambios de expresión en el espacio genómico

Hemos añadido previamente los datos de los exones a **dds** como *rowranges*. Esto nos permite realizar un gráfico en el espacio genómico.

Añadimos el símbolo del gen con la librería **org.Hs.eg.db** para mostrar los genes en el gráfico. Si no se encuentra se una el EnsemblID.

El package *Gviz* nos sirve para representar los GRanges y la metadata asociada: los cambios en el log fold debidos a la infiltración. Se usa una ventana de un millón de pares de bases upstream y downstream des de el gen con el p valor más pequeño. Realizamos 3 gráficos:

3.6.4.1. SFI vs NIT

Representación de los cambios de expresión en el espacio genómico para SFI vs NIT. Observamos cambios significativos en los genes ENGS00000182109 y PPIEL.

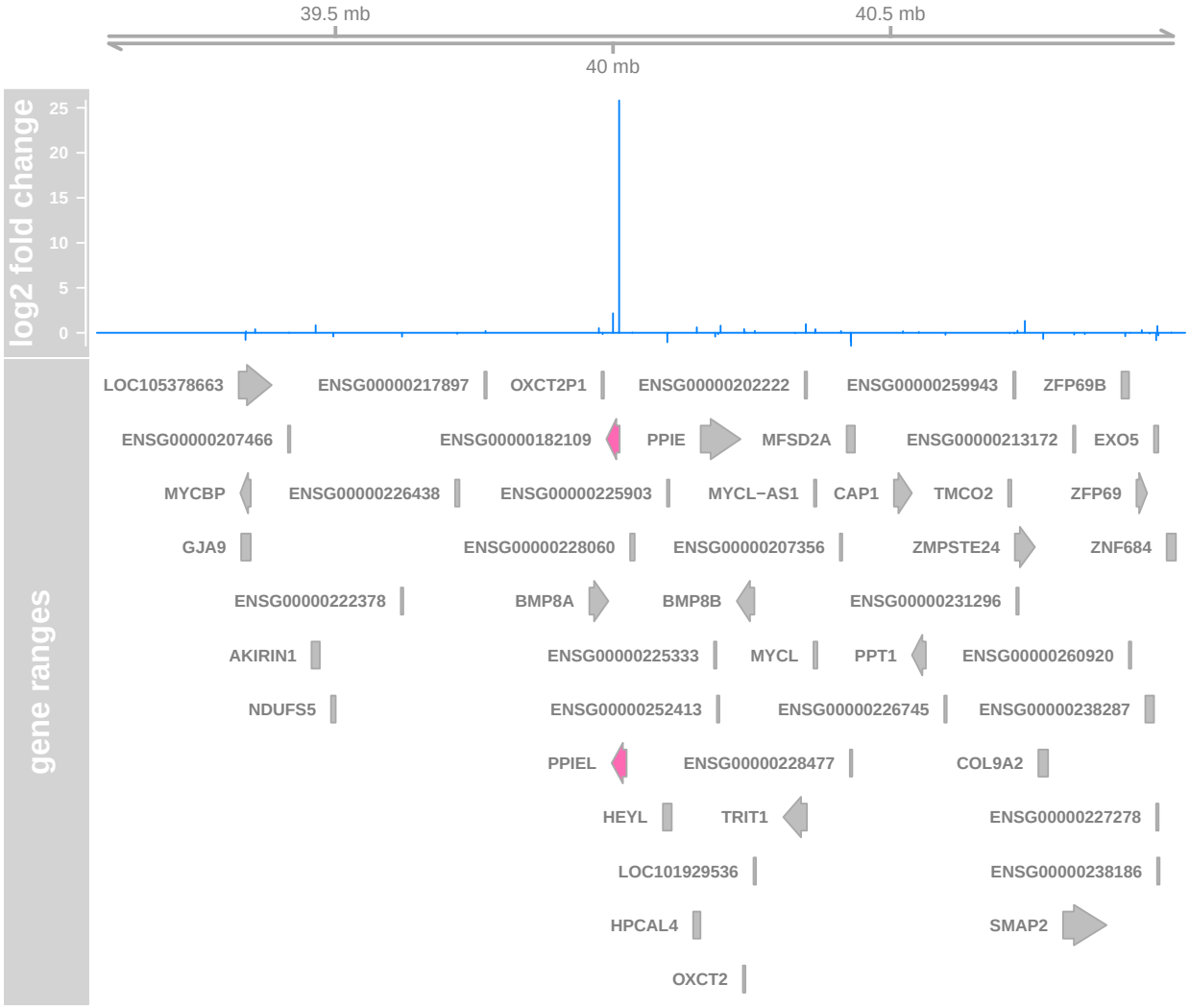

Figure 6: log2fold change in the genomic space for SFI vs NIT

3.6.4.2. ELI vs NIT

Representación de los cambios de expresión en el espacio genómico para SFI vs NIT. Observamos cambios significativos en los genes MORC4, RNF128, TBC1D8 y FDRMPD3-AS1.

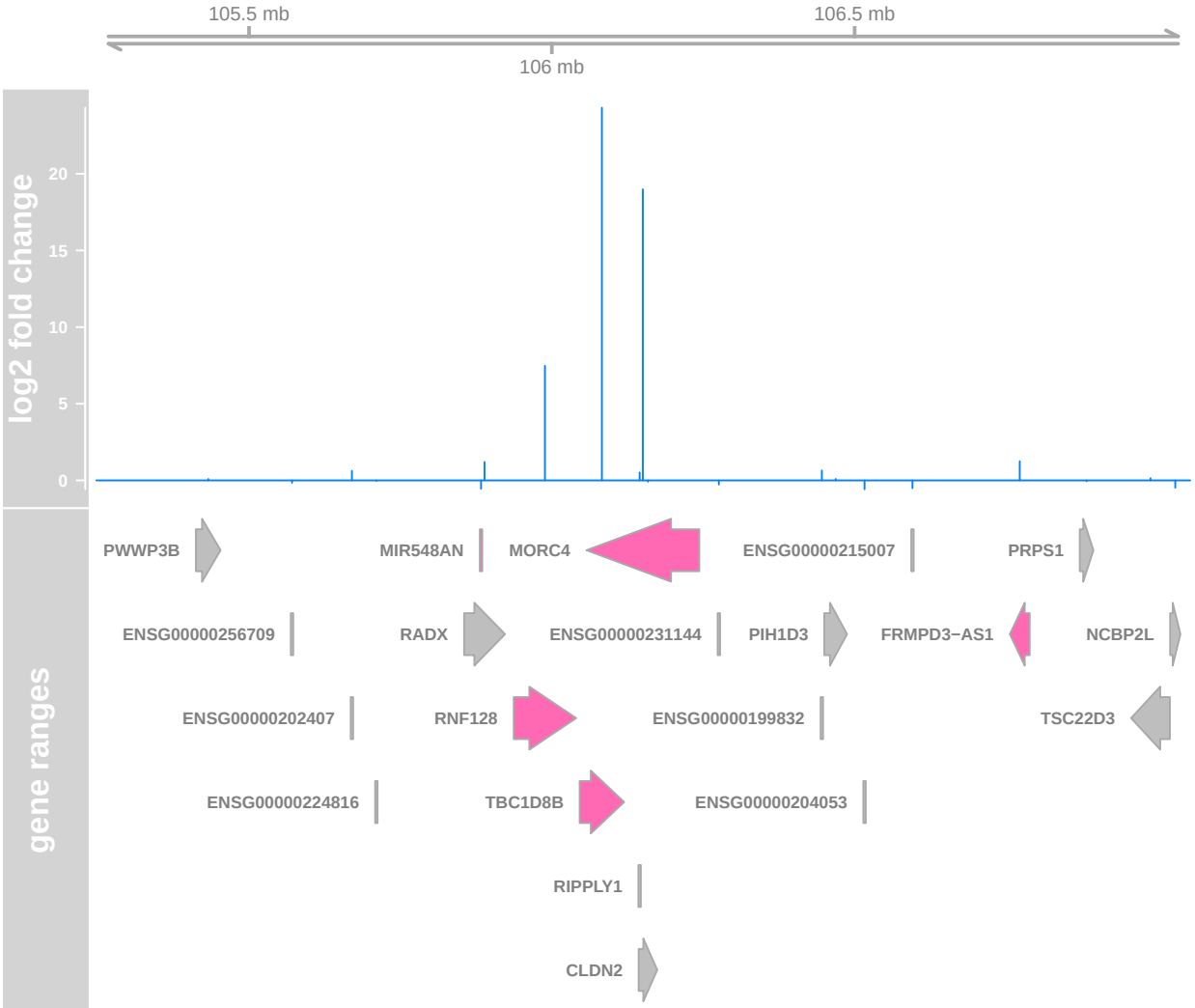

Figure 7: log2fold change in the genomic space for ELI vs NIT

3.6.4.3. Female vs Male

En este caso se representan los cambios en el log2 fold por sexo. Observamos cambios significativos en el gen GEMIN5.

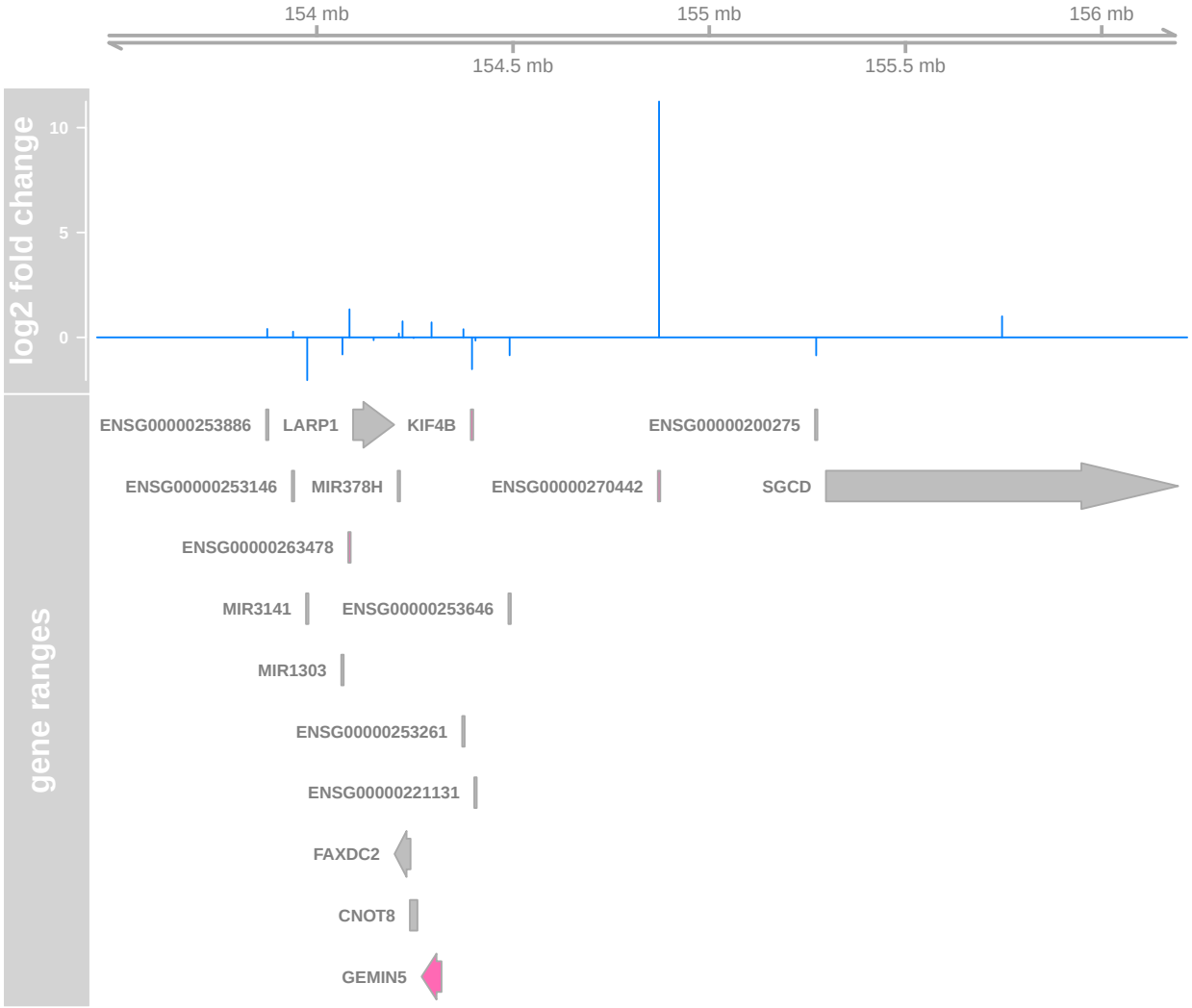

Figure 8: log2fold change in the genomic space for female vs male

3.7. Anotación y exportación de los resultados

Usamos la función *mapIds* de la librería **AnnotationDbi** para añadir información adicional a nuestra tabla de genes, que de momento sólo contiene EnsemblID. Los ordenamos por p-valor y mostramos los primeros 5, recortando la parte del medio del ID de Ensembl para ahorrar espacio.

3.7.1. SFI vs NIT

```
## # A tibble: 6 x 10
##   EnsemblID baseMean log2FoldChange lfcSE  stat   pvalue     padj symbol entrez
##   <chr>        <dbl>          <dbl> <dbl> <dbl>    <dbl>    <dbl> <chr>  <chr>
## 1 ENS243970   430.            25.8  1.46  17.4 9.81e-68 2.95e-63 PPIEL  728448
## 2 ENS133138    56.3           22.1  1.37  15.8 3.90e-56 5.86e-52 TBC1D~ 54885
## 3 ENS133247    62.2           22.0  1.47  14.6 2.20e-48 2.21e-44 KMT5C  84787
```

```
## 4 ENS133226     7.22           19.1  1.30  14.3 1.87e-46 1.41e-42 SRRM1  10250
## 5 ENS133107    36.1            21.3  1.48  14.0 9.24e-45 5.56e-41 TRPC4  7223
## 6 ENS243664    18.2            20.6  1.58  12.7 4.80e-37 2.41e-33 <NA>   <NA>
## # ... with 1 more variable: genename <chr>
```

3.7.2. ELI vs NIT

```
## # A tibble: 6 x 10
##   EnsemblID baseMean log2FoldChange lfcSE  stat   pvalue     padj symbol entrez
##   <chr>        <dbl>          <dbl> <dbl> <dbl>    <dbl>    <dbl> <chr>  <chr>
## 1 ENS133138    56.3            24.3  1.37  17.4 5.68e-68 1.53e-63 TBC1D~ 54885
## 2 ENS243970   430.             25.8  1.46  17.4 9.73e-68 1.53e-63 PPIEL  728448
## 3 ENS133247    62.2            24.8  1.46  16.6 5.11e-62 5.36e-58 KMT5C  84787
## 4 ENS133226     7.22           21.3  1.28  16.2 3.34e-59 2.63e-55 SRRM1  10250
## 5 ENS133107    36.1            23.9  1.48  15.8 1.82e-56 1.14e-52 TRPC4  7223
## 6 ENS133055    85.4            25.7  1.69  14.9 4.88e-50 2.56e-46 MYBPH  4608
## # ... with 1 more variable: genename <chr>
```

3.7.3. ELI vs SFI

```
## # A tibble: 6 x 10
##   EnsemblID baseMean log2FoldChange lfcSE  stat  pvalue    padj symbol entrez
##   <chr>        <dbl>          <dbl> <dbl> <dbl>   <dbl>   <dbl> <chr>  <chr>
## 1 ENS115204     21.4           6.24 1.05   5.49 4.10e-8 9.78e-4 MPV17  4358
## 2 ENS115363     77.2           4.04 0.705  5.02 5.16e-7 6.16e-3 EVA1A  84141
## 3 ENS229957   5095.           -7.61 1.45  -4.92 8.80e-7 6.29e-3 <NA>   <NA>
## 4 ENS260510   2211.            3.39 0.592  4.88 1.05e-6 6.29e-3 <NA>   <NA>
## 5 ENS259234    328.            4.21 0.800  4.64 3.43e-6 1.64e-2 ANKRD~ 729911
## 6 ENS119912    189.            2.95 0.538  4.55 5.25e-6 2.09e-2 IDE    3416
## # ... with 1 more variable: genename <chr>
```

3.7.4. Female vs Male

```
## # A tibble: 6 x 10
##   EnsemblID baseMean log2FoldChange lfcSE  stat    pvalue      padj symbol
##   <chr>        <dbl>          <dbl> <dbl> <dbl>     <dbl>     <dbl> <chr>
## 1 ENS270442   20482.          11.2  0.236  45.6 0.        0.        <NA>
## 2 ENS272799    4513.          -9.41 0.265 -33.6 1.17e-247 1.80e-243 <NA>
## 3 ENS272797    2284.          -9.53 0.270 -33.4 1.46e-244 1.49e-240 <NA>
## 4 ENS273013    2063.          -9.75 0.279 -33.1 2.06e-240 1.58e-236 <NA>
## 5 ENS272482    5646.          -9.04 0.260 -32.8 6.90e-236 4.25e-232 <NA>
## 6 ENS273032     500.          -8.17 0.245 -31.2 2.58e-214 1.32e-210 DGCR5
## # ... with 2 more variables: entrez <chr>, genename <chr>
```

3.7.5. Exportando los resultados

Observamos que para muchos de los tránscritos no hay símbolo, entrezID ni nombre. Si extraemos el biotipo de gen usando la librería **biomart** obtenemos la siguiente tabla de la comparación ELI vs NIT (omitiendo los biotipos que reúnen menos de 500 tránscritos):

	SYMBOL	NO SYMBOL
lncRNA	2358	5779
miRNA	687	12
misc_RNA	5	1011
processed_pseudogene	43	6095
protein_coding	11932	47
snoRNA	212	307
snRNA	15	1077
unprocessed_pseudogene	17	1346

La mayoría de los tránscritos con símbolo son codificadores de proteina, y la mayoría sin símbolo son pseudogenes. Una buena explicación de qué son los speudogenes puede verse en Zheng and Gerstein (2006). Al considerar si se deben eliminar estos datos del fichero de resultados hay que tener en cuenta que se han realizado estudios en los que los pseudogenes han resultado ser relevantes, por ejemplo: Poliseno, Marranci, and Pandolfi (2015).

Por lo tanto guardamos todos los datos en la carpeta de resultados como results_*comparación*.

3.8. Significación biológica

Interpretar los resultados comporta analizar la lista de genes obtenida y ver que funciones, procesos biológicos o pathways de moléculas aparecen con mayor frecuencia. Al filtrar los genes la selección debería ser menos restrictiva que la realizada anteriormente, por lo que quitamos el filtro por lfcThreshold. Usamos el paquete **ClusterProfile**, siguiendo las instrucciones que se encuentran en Yu, Wang, and Dall'Olio (2020).

Para empezar necesitamos un vector de genes para cada comparación escogida con el valor fold change ordenado de mayor a menor y el EntrezID como nombre de la fila. Se muestra el número de genes filtrado en cada caso:

SFIvsNIT	ELIvsNIT	ELIvsSFI
502	2765	1773

Hemos eliminado los tránscritos que no tienen entrezID asociados. Si no lo hubiéramos hecho los números serían:

SFIvsNIT	ELIvsNIT	ELIvsSFI
1219	6696	4098

Hay funciones separadas para Gene Onthology y KEGG (pathway), *enrichGO* y *enrichKEGG*. Ejecutar la segunda opción no da resultados por lo que no se incluye.

Mostramos el gráfico de red para la comparación SFI vs NIT, que es la única que da resultados:

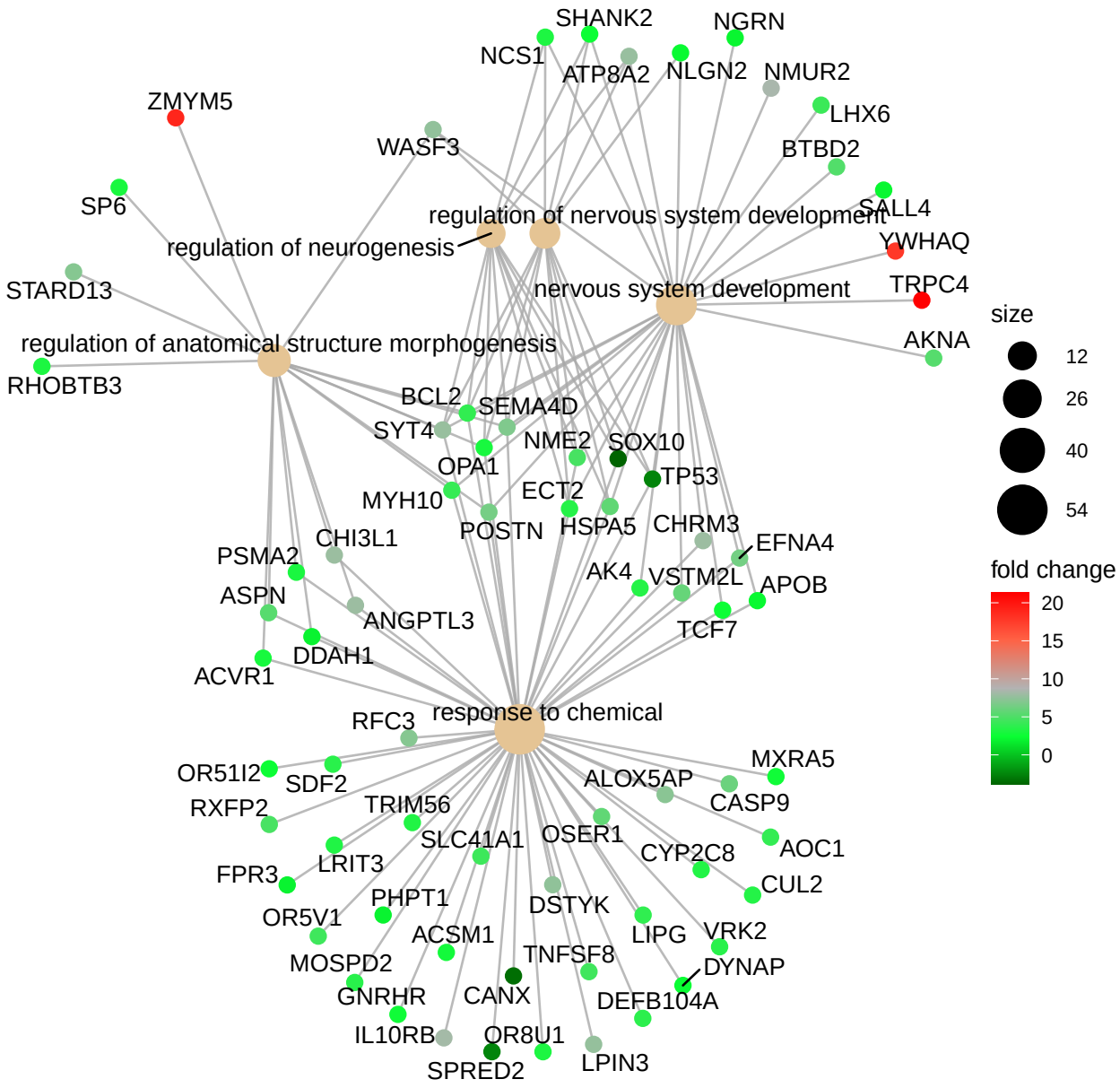

Figure 9: Net diagram showing the gene ontology for SFIvsNIT

Table 1: First rows and columns for ClusterProfiler GO on SFIvsNIT comparison

ONTOLOGY	ID	Description	GeneRatio	BgRatio
BP	GO:0051960	regulation of nervous system development	14/157	16/347
BP	GO:0007399	nervous system development	30/157	44/347
BP	GO:0042221	response to chemical	54/157	91/347
BP	GO:0022603	regulation of anatomical structure morphogenesis	17/157	22/347

Y la tabla correspondiente:

3.9. Eliminando efectos batch ocultos

Suponiendo que los 30 experimentos seleccionados no se han realizado en el mismo batch, es interesante intentar eliminar los efectos batch ocultos.

3.9.1. Usando SVA con DESeq2

Para ello usamos la librería **sva**. Ésta crea las variables sustitutas SV1 Y SV2, que añadimos como columnas a los datos **dds** creando la nueva tabla **ddssva** y luego también al diseño.

```
## Number of significant surrogate variables is:  2
## Iteration (out of 5 ):1  2  3  4  5
```

3.9.2. Análisis de expresión diferencial

Ejecutamos el pipeline de expresión diferencial con el nuevo diseño usando la función *DESeq* sobre los datos **ddssva**. Usamos el mismo log2 fold change threshold para filtrar los valores más significativos.

```
## using pre-existing size factors

## estimating dispersions

## gene-wise dispersion estimates: 2 workers

## mean-dispersion relationship

## final dispersion estimates, fitting model and testing: 2 workers
```

Mostramos el sumario de la comparación SFI vs NIT:

```
##
## out of 35786 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0.50 (up)    : 16, 0.045%
## LFC < -0.50 (down) : 181, 0.51%
## outliers [1]       : 0, 0%
## low counts [2]     : 0, 0%
## (mean count < 0)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

Mostramos el sumario de la comparación ELI vs NIT:

```
##
## out of 35786 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0.50 (up)    : 82, 0.23%
## LFC < -0.50 (down) : 643, 1.8%
```

```
## outliers [1]       : 0, 0%
## low counts [2]     : 694, 1.9%
## (mean count < 0)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

Y finalmente el de ELI vs SFI:

```
##
## out of 35786 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0.50 (up)    : 17, 0.048%
## LFC < -0.50 (down) : 59, 0.16%
## outliers [1]       : 0, 0%
## low counts [2]     : 0, 0%
## (mean count < 0)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

Si consideramos aceptable una fracción de falsos positivos del 5% podemos considerar todos los genes con un p-valor ajustado bajo 0.05 como significativos. Los genes que quedan de cada comparación son los siguientes:

SFI_NIT	ELI_NIT	ELI_SFI
178	625	44

3.9.2.1. SFI vs NIT

Mostramos los genes significativos de los que se ha reducido más la expresión:

```
## # A tibble: 6 x 7
##   EnsemblID       baseMean log2FoldChange lfcSE   stat   pvalue     padj
##   <chr>              <dbl>          <dbl> <dbl>  <dbl>    <dbl>    <dbl>
## 1 ENSG00000243970  430.            -23.8  1.31 -17.8  1.47e-70 5.24e-66
## 2 ENSG00000238317    0.463         -20.2  3.66  -5.38 7.61e- 8 2.87e- 5
## 3 ENSG00000223904    0.714         -19.8  3.65  -5.29 1.22e- 7 4.42e- 5
## 4 ENSG00000266780    7.56          -18.5  1.29 -14.0  1.54e-44 2.76e-40
## 5 ENSG00000266630   31.4           -18.1  1.35 -13.0  9.93e-39 5.92e-35
## 6 ENSG00000243749   25.2           -17.9  1.52 -11.5  2.26e-30 8.98e-27
```

Y aquellos en los que la expresión se ha incrementado más:

```
## # A tibble: 6 x 7
##   EnsemblID       baseMean log2FoldChange lfcSE  stat   pvalue     padj
##   <chr>              <dbl>          <dbl> <dbl> <dbl>    <dbl>    <dbl>
## 1 ENSG00000229926    8.17           30.0   3.68  8.01 1.16e-15 2.76e-12
## 2 ENSG00000223799    0.376          21.2   3.68  5.62 1.89e- 8 8.06e- 6
## 3 ENSG00000242978    2.15           19.7   2.50  7.69 1.50e-14 3.16e-11
## 4 ENSG00000165244    0.607          19.1   2.08  8.97 3.03e-19 1.08e-15
## 5 ENSG00000113303    0.595          18.3   2.39  7.47 7.93e-14 1.23e-10
## 6 ENSG00000101782   53.3             8.01  1.99  3.77 1.62e- 4 3.36e- 2
```

3.9.2.2. ELI vs NIT

Mostramos los genes significativos de los que se ha reducido más la expresión:

```
## # A tibble: 6 x 7
##   EnsemblID       baseMean log2FoldChange lfcSE  stat   pvalue     padj
##   <chr>              <dbl>          <dbl> <dbl> <dbl>    <dbl>    <dbl>
## 1 ENSG00000243970   430.            -24.2  1.36 -17.3 2.75e-67 9.66e-63
```

```
## 2 ENSG00000243749    25.2           -21.5  1.54 -13.7 1.63e-42 7.13e-39
## 3 ENSG00000243539    16.3           -20.8  1.31 -15.4 1.26e-53 1.47e-49
## 4 ENSG00000243532    11.8           -20.7  1.24 -16.3 9.56e-60 1.68e-55
## 5 ENSG00000266630    31.4           -20.4  1.36 -14.6 2.33e-48 1.63e-44
## 6 ENSG00000243385     4.15          -20.2  1.50 -13.1 1.76e-39 6.87e-36
```

Y aquellos en los que la expresión se ha incrementado más:

```
## # A tibble: 6 x 7
##   EnsemblID       baseMean log2FoldChange lfcSE  stat   pvalue     padj
##   <chr>              <dbl>          <dbl> <dbl> <dbl>    <dbl>    <dbl>
## 1 ENSG00000242640    4.71           25.5   3.88  6.44 1.21e-10 3.30e- 8
## 2 ENSG00000139880    0.386          22.8   3.50  6.37 1.85e-10 4.80e- 8
## 3 ENSG00000048991    0.381          18.9   2.98  6.16 7.39e-10 1.74e- 7
## 4 ENSG00000213218    0.546          18.7   2.53  7.16 7.88e-13 2.97e-10
## 5 ENSG00000258476   75.9             9.77  2.01  4.62 3.83e- 6 4.24e- 4
## 6 ENSG00000229957 5095.              9.01  1.59  5.36 8.55e- 8 1.33e- 5
```

3.9.2.3. ELI vs SFI

Mostramos los genes significativos de los que se ha reducido más la expresión:

```
## # A tibble: 6 x 7
##   EnsemblID       baseMean log2FoldChange lfcSE  stat   pvalue     padj
##   <chr>              <dbl>          <dbl> <dbl> <dbl>    <dbl>    <dbl>
## 1 ENSG00000273051    1.56           -34.0  3.71 -9.02 1.83e-19 6.54e-15
## 2 ENSG00000229926    8.17           -31.8  3.70 -8.47 2.54e-17 4.55e-13
## 3 ENSG00000261837    0.417          -31.4  3.72 -8.29 1.11e-16 1.32e-12
## 4 ENSG00000157657    1.70           -21.6  3.73 -5.66 1.48e- 8 4.42e- 5
## 5 ENSG00000272253    2.88           -21.0  3.72 -5.51 3.56e- 8 8.50e- 5
## 6 ENSG00000242978    2.15           -20.5  2.52 -7.97 1.64e-15 1.18e-11
```

Y aquellos en los que la expresión se ha incrementado más:

```
## # A tibble: 6 x 7
##   EnsemblID       baseMean log2FoldChange lfcSE  stat   pvalue          padj
##   <chr>              <dbl>          <dbl> <dbl> <dbl>    <dbl>         <dbl>
## 1 ENSG00000242640    4.71           24.8   3.71  6.53 6.50e-11 0.000000291
## 2 ENSG00000139880    0.386          20.4   3.35  5.94 2.91e- 9 0.0000116
## 3 ENSG00000213218    0.546          18.0   2.44  7.16 8.12e-13 0.00000000415
## 4 ENSG00000048991    0.381          17.0   2.87  5.74 9.34e- 9 0.0000304
## 5 ENSG00000229957 5095.              9.32  1.52  5.80 6.81e- 9 0.0000244
## 6 ENSG00000258476   75.9             8.76  1.92  4.30 1.67e- 5 0.0176
```

3.9.3. Representando los resultados

3.9.3.1. Counts-Plot

Realizamos un Counts-Plot para la comparación ELI vs NIT. Se observa un cambio: el gen más significativo aún tiene un counts más elevado en SFI para mujeres que para hombres, pero ahora tiene un counts más elevado para hombres en el grupo ELI.

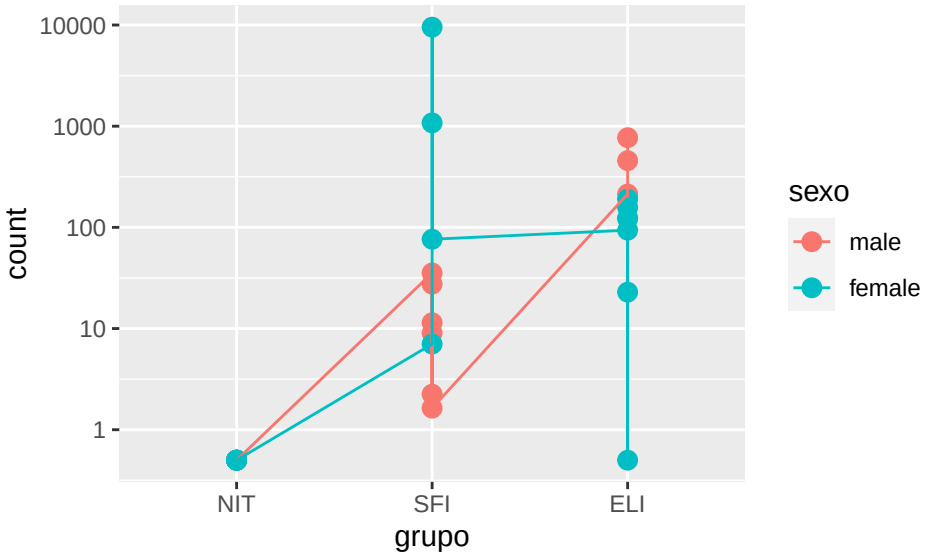

Figure 10: Counts-Plot for ELI vs NIT

3.9.3.2. Diagrama de Venn

Observamos las coincidencias entre los diferentes contrastes tras eliminar el efecto batch oculto. Vemos que los genes significativos de la comparación ELI vs SFI han aumentado, tanto los comunes con ELI vs NIT como los no comunes con ninguna comparación.

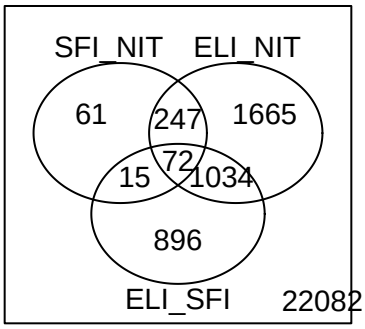

Figure 11: Venn's Diagram with SVA

3.9.3.3. Representando los cambios de expresión en el espacio genómico

Repetimos los gráficos para SFI vs NIT y ELI vs NIT para los datos **ddssva**.

3.9.3.3.1. SFI vs NIT

El gráfico no ha variado. Observamos de nuevo cambios significativos en los genes ENGS00000182109 y PPIEL.

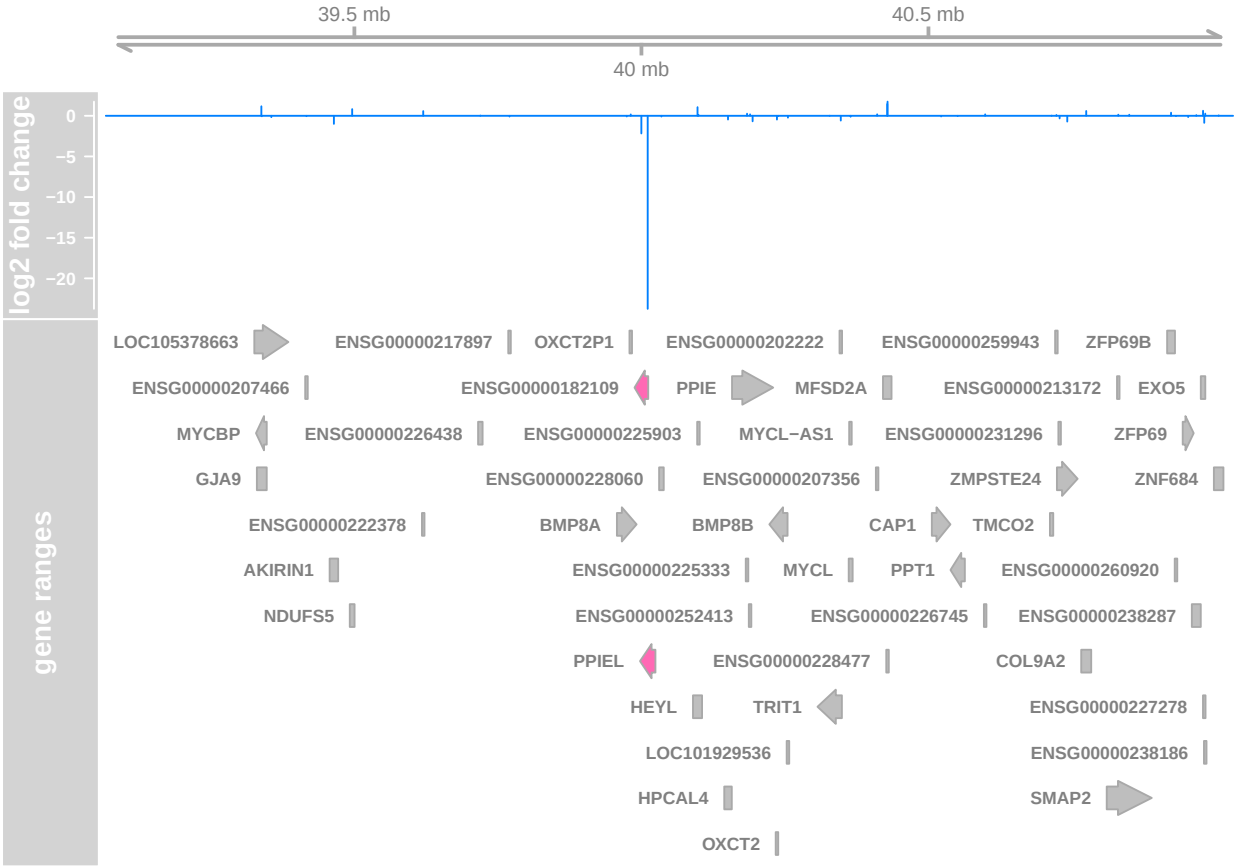

Figure 12: log2fold change in the genomic space for SFI vs NIT (SVA)

3.9.3.3.2. ELI vs NIT

El gráfico es completamente distinto. Observamos que los genes expresados significativamente son ahora los dos mismos que para SFI vs NIT, sumandole BMP8B, MFSD2A y COL9A2.

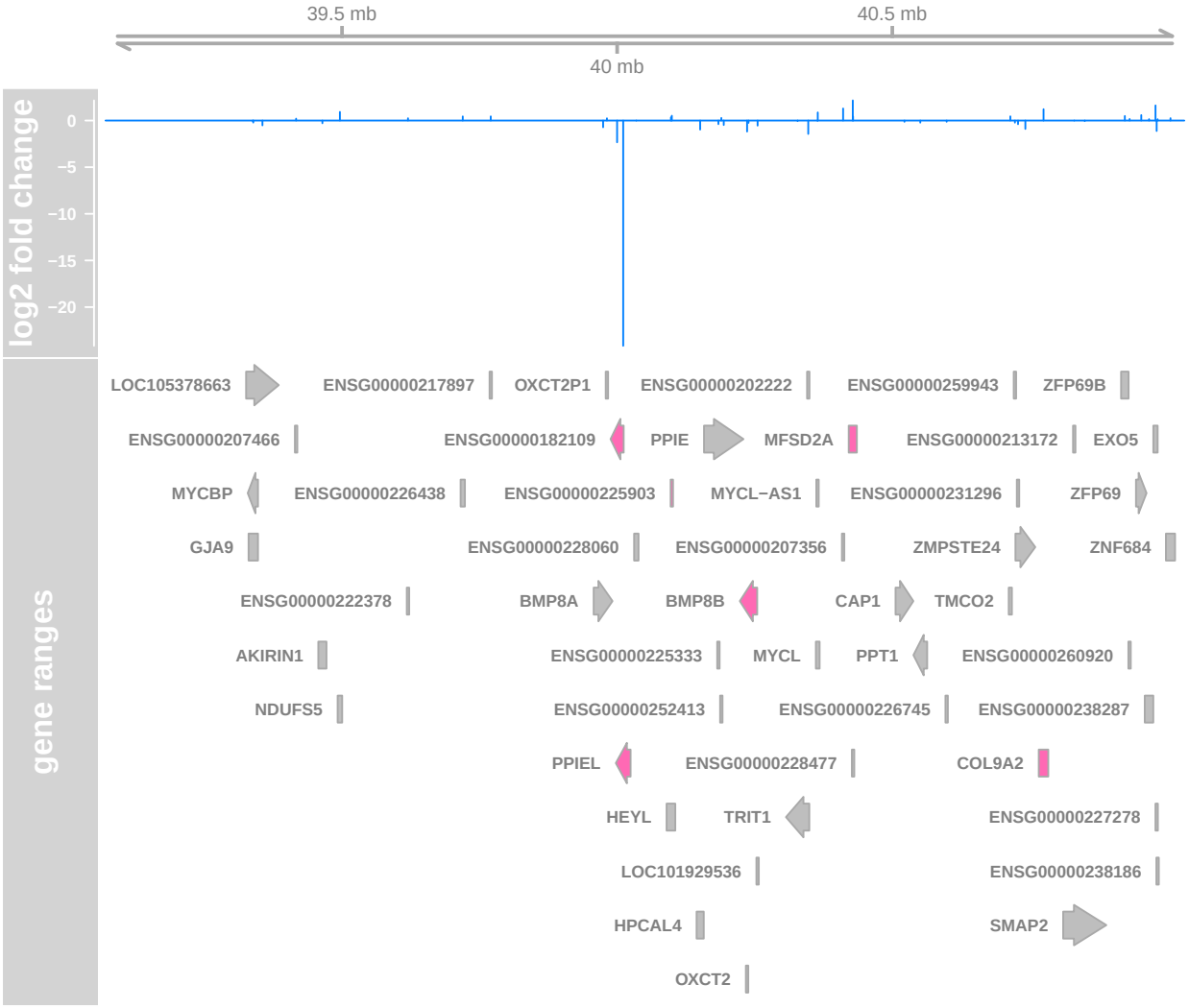

Figure 13: log2fold change in the genomic space for ELI vs NIT (SVA)

3.9.4. Anotación y exportación de los resultados

Usamos la función *mapIds* de la libreria **AnnotationDbi** para añadir información adicional a nuestra tabla de genes, que de momento sólo contiene EnsemblID. Los ordenamos por p-valor y mostramos los primeros 5, recortando la parte del medio del ID de Ensembl para ahorrar espacio.

3.9.4.1. SFI vs NIT

```
## 'select()' returned 1:many mapping between keys and columns
## 'select()' returned 1:many mapping between keys and columns
## 'select()' returned 1:many mapping between keys and columns
```

```
## # A tibble: 6 x 10
##   EnsemblID baseMean log2FoldChange lfcSE  stat   pvalue     padj symbol entrez
```

```
##   <chr>        <dbl>          <dbl> <dbl> <dbl>    <dbl>    <dbl> <chr>  <chr>
## 1 ENS243970   430.           -23.8  1.31 -17.8 1.47e-70 5.24e-66 PPIEL  728448
## 2 ENS266780     7.56         -18.5  1.29 -14.0 1.54e-44 2.76e-40 MIR44~ 10084~
## 3 ENS243532    11.8          -17.6  1.25 -13.6 2.07e-42 2.47e-38 <NA>   <NA>
## 4 ENS266645     2.91         -16.9  1.21 -13.6 3.80e-42 3.40e-38 MIR42~ 10042~
## 5 ENS251785     4.21         -17.7  1.32 -13.0 7.75e-39 5.55e-35 <NA>   <NA>
## 6 ENS266630    31.4          -18.1  1.35 -13.0 9.93e-39 5.92e-35 <NA>   <NA>
## # ... with 1 more variable: genename <chr>
```

3.9.4.2. ELI vs NIT

```
## 'select()' returned 1:many mapping between keys and columns
## 'select()' returned 1:many mapping between keys and columns
## 'select()' returned 1:many mapping between keys and columns
```

```
## # A tibble: 6 x 10
##   EnsemblID baseMean log2FoldChange lfcSE  stat   pvalue     padj symbol entrez
##   <chr>        <dbl>          <dbl> <dbl> <dbl>    <dbl>    <dbl> <chr>  <chr>
## 1 ENS243970   430.           -24.2  1.36 -17.3 2.75e-67 9.66e-63 PPIEL  728448
## 2 ENS243532    11.8          -20.7  1.24 -16.3 9.56e-60 1.68e-55 <NA>   <NA>
## 3 ENS243539    16.3          -20.8  1.31 -15.4 1.26e-53 1.47e-49 <NA>   <NA>
## 4 ENS266780     7.56         -20.1  1.32 -14.8 9.13e-50 8.01e-46 MIR44~ 10084~
## 5 ENS266630    31.4          -20.4  1.36 -14.6 2.33e-48 1.63e-44 <NA>   <NA>
## 6 ENS251785     4.21         -19.9  1.34 -14.5 1.41e-47 8.23e-44 <NA>   <NA>
## # ... with 1 more variable: genename <chr>
```

3.9.4.3. ELI vs SFI

```
## 'select()' returned 1:many mapping between keys and columns
## 'select()' returned 1:many mapping between keys and columns
## 'select()' returned 1:many mapping between keys and columns
```

```
## # A tibble: 6 x 10
##   EnsemblID baseMean log2FoldChange lfcSE  stat   pvalue     padj symbol entrez
##   <chr>        <dbl>          <dbl> <dbl> <dbl>    <dbl>    <dbl> <chr>  <chr>
## 1 ENS273051    1.56           -34.0  3.71 -9.02 1.83e-19 6.54e-15 <NA>   <NA>
## 2 ENS229926    8.17           -31.8  3.70 -8.47 2.54e-17 4.55e-13 <NA>   <NA>
## 3 ENS261837    0.417          -31.4  3.72 -8.29 1.11e-16 1.32e-12 <NA>   <NA>
## 4 ENS165244    0.607          -17.5  2.14 -7.97 1.62e-15 1.18e-11 ZNF367 195828
## 5 ENS242978    2.15           -20.5  2.52 -7.97 1.64e-15 1.18e-11 <NA>   <NA>
## 6 ENS113303    0.595          -18.6  2.39 -7.58 3.47e-14 2.07e-10 BTNL8  79908
## # ... with 1 more variable: genename <chr>
```

3.9.4.4. Exportando los resultados

Los guardamos en la carpeta de resultados como SVA_results_*comparación*.

3.9.5. Significación biológica

Re-interpretamos los resultados con los nuevos datos **ddssva**. Este es el numero de genes que se obtienen para cada combinación:

SFIvsNIT	ELIvsNIT	ELIvsSFI
323	2454	2128

Sólo se obtienen resultados para la comparación ELI vs SFI:

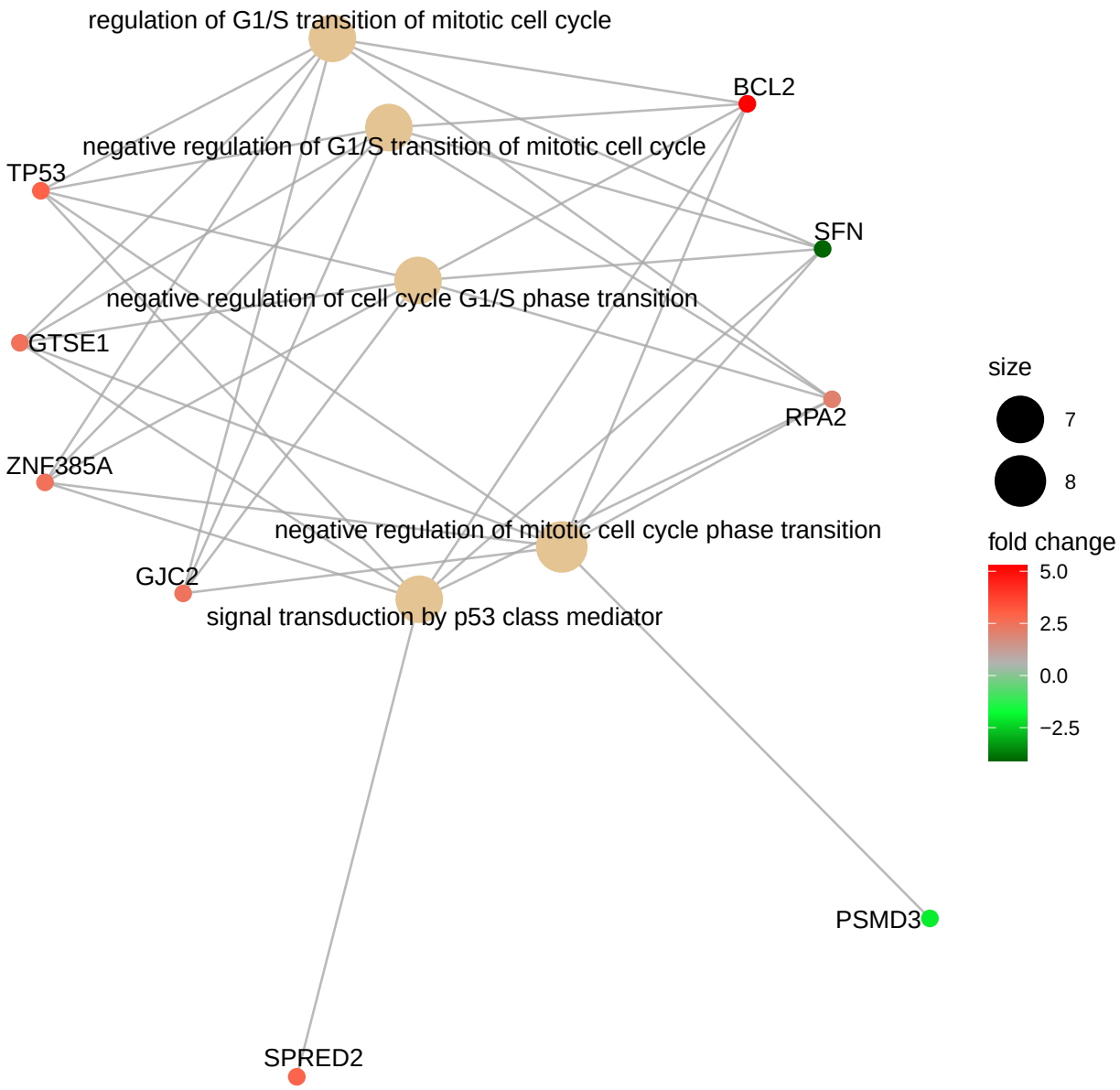

Figure 14: Net diagram showing the gene ontology for ELIvsSFI

Table 2: First rows and columns for ClusterProfiler GO on ELIvsSFI comparison

ONTOLOGY	ID	Description	GeneRatio	BgRatio
BP	GO:1902807	negative regulation of cell cycle G1/S phase transition	7/108	14/1543
BP	GO:2000134	negative regulation of G1/S transition of mitotic cell cycle	7/108	14/1543
BP	GO:1901991	negative regulation of mitotic cell cycle phase transition	8/108	23/1543
BP	GO:0072331	signal transduction by p53 class mediator	7/108	19/1543

Esta es la tabla resultante:

4. Resultados

Se obtienen los siguientes ficheros como resultados del análisis:

Gráficos

- 3.4.3. Distancias entre las muestras
- 3.4.4. Gráfico PCA
- 3.6.1. Counts-Plot
- 3.6.2. Diagrama de Venn
- 3.6.3. Clustering de los genes
- 3.6.4.1. SFI vs NIT
- 3.6.4.2. ELI vs NIT
- 3.6.4.3. Female vs Male
- 3.9.3.1. Counts-Plot
- 3.9.3.2. Diagrama de Venn
- 3.9.3.3.1. SFI vs NIT
- 3.9.3.3.2. ELI vs NIT

Ficheros

- results_ELIvsNIT.csv - Resultados anotados para ELI vs NIT
- results_ELIvsSFI.csv - Resultados anotados para ELI vs SFI
- results_FemalevsMale.csv - Resultados anotados para Female vs Male
- results_SFIvsNIT.csv - Resultados anotados para SFI vs NIT
- ClusterProfileGO.Results.SFIvsNIT.csv - Resultados GO de la comparación SFIvsNIT.
- ClusterProfileGO_Barplot.SFIvsNIT.pdf - Barplot de GO para la comparación SFIvsNIT.
- ClusterProfileGO-cnetplot.SFIvsNIT.pdf - Cnetplot de GO para la comparación SFIvsNIT.
- SVA_results_ELIvsNIT.csv - Resultados anotados para ELI vs NIT sin efecto batch oculto
- SVA_results_ELIvsSFI.csv - Resultados anotados para ELI vs SFI sin efecto batch oculto
- SVA_results_SFIvsNIT.csv - Resultados anotados para SFI vs NIT sin efecto batch oculto
- SVA_ClusterProfileGO.Results.ELIvsSFI.csv- Resultados de GO para la comparación ELIvsSFI sin efecto batch oculto.
- SVA_ClusterProfileGO_Barplot.ELIvsSFI.pdf - Barplot de GO para la comparación ELIvsSFI sin efecto batch oculto.
- SVA_ClusterProfileGO_cnetplot.ELIvsSFI.pdf - Cnetplot de GO para la comparación ELIvsSFI sin efecto batch oculto.

5. Discusión

No tenemos información de los lotes, por lo que me ha parecido apropiado hacer el análisis tanto directamente con los datos como eliminando el efecto batch oculto. Los resultados son bastante distintos, se debería escoger el más adecuado.

6. Bibliografía

Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. "The Genotype-Tissue Expression (GTEx) project." *Nature Genetics* 45 (6): 580–85. https://doi.org/10.1038/ng.2653.

Poliseno, Laura, Andrea Marranci, and Pier Paolo Pandolfi. 2015. "Pseudogenes in human cancer." Higher Education Press. https://doi.org/10.3389/fmed.2015.00068.

Yu, Guangchuang, Li-Gen Wang, and Giovanni Dall'Olio. 2020. "Package ' clusterProfiler '." https://bioconductor.org/packages/release/bioc/manuals/clusterProfiler/man/clusterProfiler.pdf.

Zheng, Deyou, and Mark B. Gerstein. 2006. "A computational approach for identifying pseudogenes in the ENCODE regions." *Genome Biology* 7 Suppl 1 (Suppl 1). BioMed Central: S13. https://doi.org/10.1186/gb-2006-7-s1-s13.