Proceso de análisis de datos de ultrasecuenciación

PAC 2 - ANÁLISIS DE DATOS ÓMICOS (UOC)

María José Ojeda-Montes

14 de junio, 2020

Contents

ANÁLISIS DE ULTRASECUENCIACIÓN

Obtención de los datos

El objetivo principal de este estudio es **analizar el efecto en la expresión genética debido a la infiltración mediante métodos en tiroides.**

– *Enlace al repositorio de github*: https://github.com/MJoseom/ADO_RNAseq

El análisis de RNAseq se realiza a partir de los datos ya divididos en dos archivos: targets.csv y counts.csv empleando las herramientas de Bioconductor. Para realizar la práctica, se han seguido diferentes tutoriales [1] [2] [3] [4].

- **Archivo *targets.csv***

Se trata del archivo de datos **targets.csv** con 292 observaciones, en este caso corresponde a datos de expresión (RNA-seq) pertenecientes a un análisis del tiroides. Para cada observación, se han anotado 9 variables o características. En este caso, se compara tres tipos de infiltración: 14 muestras pertenecientes al grupo *Extensive lymphoid infiltrates* (ELI), 236 muestras del grupo *Not infiltrated tissues* (NIT) y 42 muestras del grupo *Small focal infiltrates* (SFI). En la base de datos tenemos 0 missing values y 0 valores nulos. En este caso, las variables Grupo_analisis, body_site, molecular_data_type, sex, Group son consideradas factores.

Las variables o características consideradas son: Experiment, SRA_Sample, Sample_Name, Grupo_analisis, body_site, molecular_data_type, sex, Group, ShortName.

- **Archivo *counts.csv***

Se trata del archivo de datos **counts.csv** con 292 muestras para las que se han analizado la expresión de 56202 genes en tiroides anotados con el Gencode ID como índice de cada fila. En este caso, el nombre de las columnas de la matriz *Count* coincide con la variable *Sample_Name* de los datos del archivo **targets.csv**. En la base de datos tenemos 0 missing values y 0 valores nulos. En este caso, las variables 292 son consideradas numéricas.

Selección de un subset

Se realiza una selección de 30 muestras de forma aleatoria, 10 para cada grupo de estudio (NIT, SFI, ELI).

1. Determinar la distribución de los grupos de estudio en función del sexo

Table 1: Proporción en función del sexo y el grupo para la totalidad de los datos

	female	male
ELI	3.08	1.71
NIT	28.08	52.74
SFI	5.82	8.56

2. Seleccionar 10 muestras de cada grupo de forma aleatoria del archivo targets.csv

```
set.seed(params$seed)
# Selección del subset
```

```r
setTarget <- dataTarget %>% group_by(Group) %>% sample_n(10)
# Modificación la notación de la columna de muestras para evitar errores con la matriz de Count
setTarget$Sample_Name <- gsub("-", ".", setTarget$Sample_Name)
# Exportación la tabla con el subset de Targets
write.csv(setTarget, file= file.path(dataDir, "setTargets.csv"))
```

3. Analizar las características del dataSet

El dataset de *Targets* creado seleccionando muestras aleatorias de cada uno de los grupos, presenta 30 muestras para las que se han anotado 9 características o variables. En este caso, disponemos de tres tipos de infiltración: 10 muestras pertenecientes al grupo *Extensive lymphoid infiltrates* (ELI), 10 muestras del grupo *Not infiltrated tissues* (NIT) y 10 muestras del grupo *Small focal infiltrates* (SFI). En la base de datos tenemos 0 missing values y 0 valores nulos.

Table 2: Número de muestras en función del sexo y el grupo para el dataset seleccionado

	female	male
ELI	7	3
NIT	5	5
SFI	3	7

En este caso, dado que se seleccionan el mismo número de muestras para todos los grupos, no se mantien la misma relación de muestras en función del sexo y del grupo.

El identificador de las muestras seleccionadas para el dataset son:

Table 3: Nombre de la muestra del dataSet

Muestra	Sexo	Tipo de infiltración
GTEX.ZYY3.1926.SM.5GZXS	female	ELI
GTEX.11XUK.0226.SM.5EQLW	female	ELI
GTEX.R55G.0726.SM.2TC6J	female	ELI
GTEX.14BMU.0226.SM.5S2QA	female	ELI
GTEX.YFC4.2626.SM.5P9FQ	female	ELI
GTEX.TMMY.0826.SM.33HB9	female	ELI
GTEX.11NV4.0626.SM.5N9BR	male	ELI
GTEX.14ABY.0926.SM.5Q5DY	male	ELI
GTEX.YJ89.0726.SM.5P9F7	male	ELI
GTEX.PLZ4.1226.SM.2I5FE	female	ELI
GTEX.ZF28.0826.SM.4WKGJ	male	NIT
GTEX.S7SF.0226.SM.5SI7H	female	NIT
GTEX.12WSJ.0326.SM.5GCMT	female	NIT
GTEX.U8T8.2326.SM.3DB96	male	NIT
GTEX.12WSD.0926.SM.5GCNL	female	NIT
GTEX.11VI4.0226.SM.5GU6C	female	NIT
GTEX.111CU.0226.SM.5GZXC	male	NIT
GTEX.ZAB5.0726.SM.5P9JG	male	NIT
GTEX.ZLWG.0526.SM.4WWFB	female	NIT
GTEX.13N1W.0826.SM.5MR5J	male	NIT
GTEX.12ZZX.1226.SM.5EGHS	female	SFI
GTEX.11EQ8.0826.SM.5N9FG	male	SFI
GTEX.12584.0826.SM.5FQSK	male	SFI

Muestra	Sexo	Tipo de infiltración
GTEX.131XF.1826.SM.5EGKG	male	SFI
GTEX.12ZZY.0826.SM.5EQMT	male	SFI
GTEX.13FXS.0726.SM.5LZXJ	male	SFI
GTEX.131YS.0726.SM.5P9G9	female	SFI
GTEX.WYVS.0326.SM.3NM9V	female	SFI
GTEX.SIU8.0626.SM.2XCDN	male	SFI
GTEX.TKQ1.0126.SM.33HB3	male	SFI

4. Seleccionar los datos de expresión génica correspondientes al dataset del archivo counts.csv

Buscamos la coincidencia de los identificadores de la tabla 3 en las columnas las columnas correspondientes a las muestras del dataSet.

```
# Selección de las columnas con el identificador del dataset
setCount <- dataCount[,  sampleName]
# Eliminación de la versión en el nombre de cada gen con la notación de Ensambl
row.names(setCount) <- sapply(strsplit(row.names(setCount), split = ".", fixed=TRUE), function(a) a[1])
# Exportación la tabla con el subset de Counts
write.csv(setCount, file= file.path(dataDir, "setCounts.csv"))
```

El dataset de *Counts* creado presenta 30 muestras para las que se han analizado la expresión de 56202 genes en tiroides anotados con el Gencode ID como índice de cada fila. En este caso, el nombre de las columnas coincide con los 10 muestras de cada grupo (*i.e.*, NIT, SFI, ELI) seleccionados de forma aleatoria en la Tabla 3. En la base de datos tenemos 0 missing values y 0 valores nulos.

Filtraje no específico

En la matriz *setCount* cada fila representa un gen con el código Ensembl y cada columna una librería de RNA secuenciada. Los valores corresponden a los datos sin procesar de fragmentos que se asignaron de forma única al gen respectivo en cada una de las librerías. Se empleará el paquete `Deseq2` para realizar el análisis de expresión diferencial de los datos obtenidos por ultrasecuenciación. En este caso se determina como condición el efecto del tipo de infiltración (*i.e.*, NIT, SFI, ELI) para cada una de las muestras.

```
dds <- DESeqDataSetFromMatrix(countData = setCount,
                              colData = setTarget,
                              design = ~ Group)
dds
```

```
## class: DESeqDataSet
## dim: 56202 30
## metadata(1): version
## assays(1): counts
## rownames(56202): ENSG00000223972 ENSG00000227232 ... ENSG00000210195
##   ENSG00000210196
## rowData names(0):
## colnames(30): GTEX.ZYY3.1926.SM.5GZXS GTEX.11XUK.0226.SM.5EQLW ...
##   GTEX.SIU8.0626.SM.2XCDN GTEX.TKQ1.0126.SM.33HB3
## colData names(9): Experiment SRA_Sample ... Group ShortName
```

En primer lugar, se realiza un filtro con el fin de eliminar todas aquellas filas que presentan únicamente 0 o bien un único valor de expresión genética. De este modo, se disminuye el tamaño de la matriz y aumentaremos la agilidad computacional. Así, pasamos de tener 56202 genes a reducir la cantidad de genes representativos hasta 43525.

Normalización

Con el objetivo de estabilizar la varianza entre los diferentes contajes de expresión génica, se ha aplicado el método VST (variance stabilizing transformation) y el rlog (regularized-logarithm transformation) del propio paquete `DESseq2`.

- VST

```
# VSD method
vsd <- vst(dds, blind = FALSE)
head(assay(vsd), 3)[,1:2]
```

```
##                 GTEX.ZYY3.1926.SM.5GZXS GTEX.11XUK.0226.SM.5EQLW
## ENSG00000223972               4.957860                 4.041418
## ENSG00000227232              10.206114                 8.957012
## ENSG00000243485               4.420783                 4.041418
```

```
head(colData(vsd), 3)
```

```
## DataFrame with 3 rows and 10 columns
##                           Experiment  SRA_Sample              Sample_Name
##                          <character> <character>              <character>
## GTEX.ZYY3.1926.SM.5GZXS    SRX568364   SRS627095  GTEX.ZYY3.1926.SM.5GZXS
## GTEX.11XUK.0226.SM.5EQLW   SRX619829   SRS644736 GTEX.11XUK.0226.SM.5EQLW
## GTEX.R55G.0726.SM.2TC6J    SRX204036   SRS374975  GTEX.R55G.0726.SM.2TC6J
##                          Grupo_analisis body_site        molecular_data_type
##                                <factor>  <factor>                   <factor>
## GTEX.ZYY3.1926.SM.5GZXS               3   Thyroid Allele-Specific Expression
## GTEX.11XUK.0226.SM.5EQLW              3   Thyroid              RNA Seq (NGS)
## GTEX.R55G.0726.SM.2TC6J               3   Thyroid              RNA Seq (NGS)
##                              sex    Group   ShortName        sizeFactor
##                          <factor> <factor> <character>         <numeric>
## GTEX.ZYY3.1926.SM.5GZXS    female      ELI   ZYY3-_ELI 0.873267401835296
## GTEX.11XUK.0226.SM.5EQLW   female      ELI   11XUK_ELI 0.901884946019608
## GTEX.R55G.0726.SM.2TC6J    female      ELI   R55G-_ELI  0.29590023913757
```

- rlog

```
# rlog method
rld <- rlog(dds, blind = FALSE)
```

```
## rlog() may take a few minutes with 30 or more samples,
## vst() is a much faster transformation
```

```
head(assay(rld), 3)[,1:2]
```

```
##                 GTEX.ZYY3.1926.SM.5GZXS GTEX.11XUK.0226.SM.5EQLW
## ENSG00000223972               1.8457820                1.0262641
## ENSG00000227232              10.0237124                9.0651040
## ENSG00000243485               0.3170474                0.1270095
```

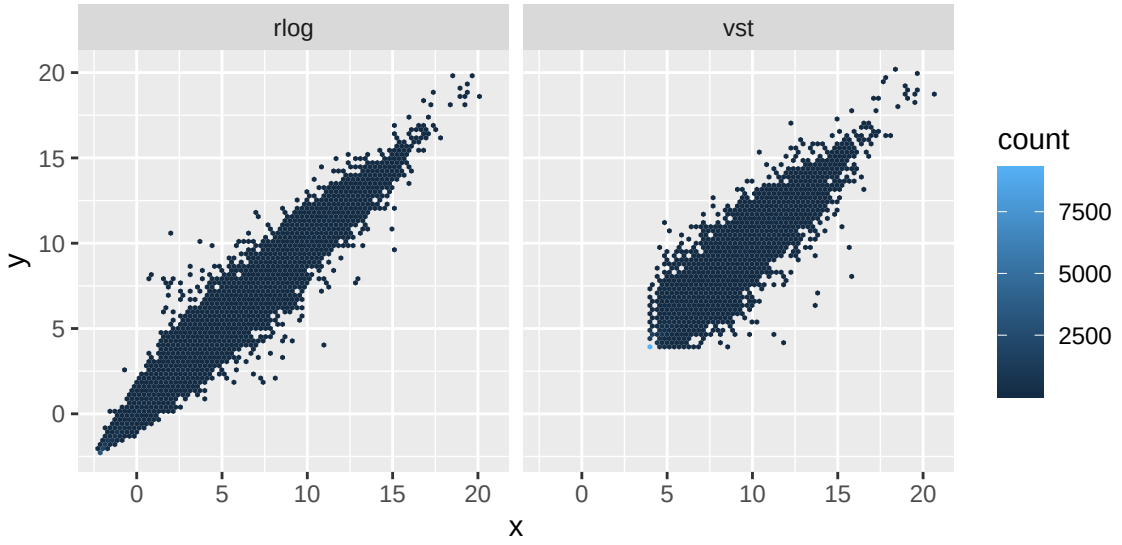

Figure 1: Diagrama de dispersión de recuentos transformados de dos muestras por los métodos VST y rlog

La gráfica de dispersión de la Figura 1 nos muestra la transformación normalizada de *counts* en las dos primeras muestras empleando el método VST y rlog. Observamos que el rango de valores de rlog para el eje de las `x` es de -2.3 a 24.8, en cambio, el rango de valores para VST es menor para los valores inferiores, situándose en un rango de 4 a 21.7.

Un paso interesante en el análisis de RNAseq es determinar la similitud entre las muestras a partir de determinar la distancia entre éstas.

Para ello, realizamos una matriz de correlación de las distancias y con el fin de visualizarlas de una forma gráfica, realizamos un heatmap empleando el paquete `pheatmap`.

En el heatmap de la Figura 2 observamos a partir del dendograma que hay 4 grandes clusters de muestras que presentan una cierta similitud. La distancia entre las muestras con mayor similitud, se encuentran con valores de 100 aproximadamente, tal y como se muestra en la paleta de colores del heatmap. Y éstas principalmente se concentran en los dos primeros clusters (primera rama izquierda del dendograma) pertenecientes a los grupos `NIT` y `SFI`. En general, las muestras de los diferentes grupos, se encuentran bastante agrupadas en el caso de la infiltración tipo `ELI` y la `NIT`. En cambio, las muestras del grupo `SFI` se encuentran más dispersas, mostrando mayor distancia entre ellas y por lo tanto, mayor variabilidad.

En el gráfico de la Figura 3 se han representado las dos primeras componentes del PCA (Principal Component Analysis) que explicarían el 72% de la variabilidad de las muestras, mayoritariamente por la PC1 que abarca

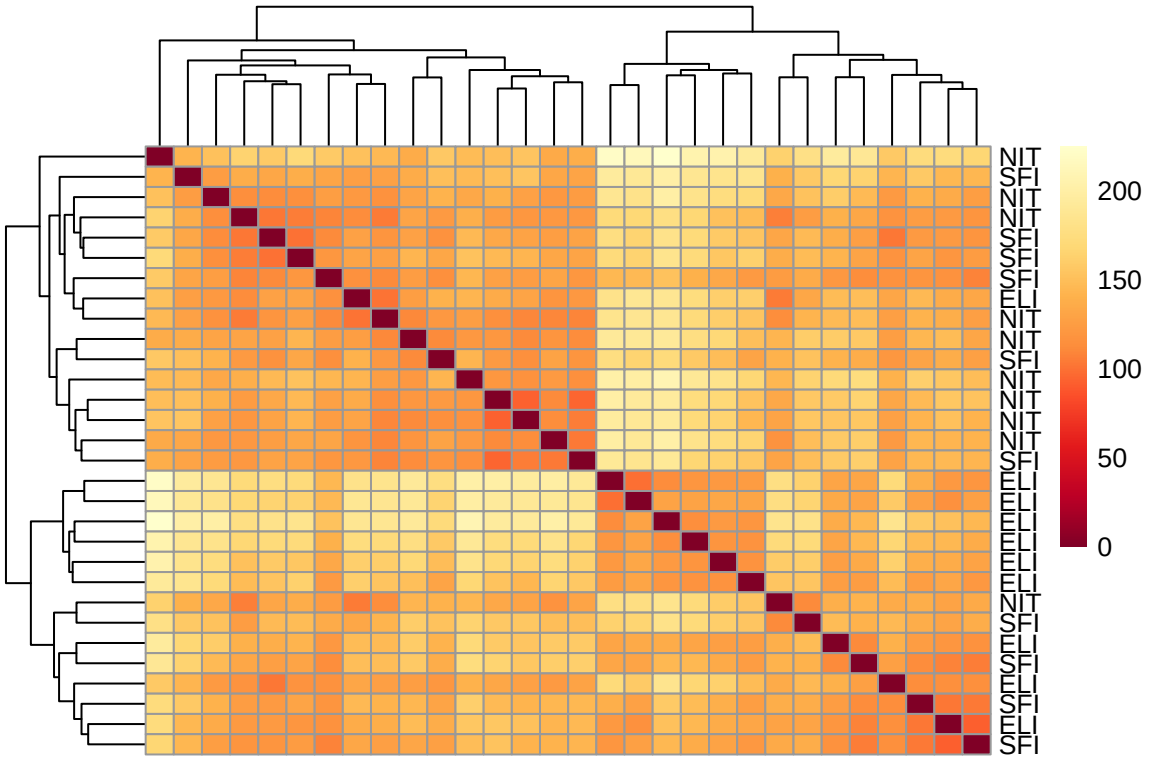

Figure 2: Heatmap de distancias de muestra a muestra usando los valores transformados de VST

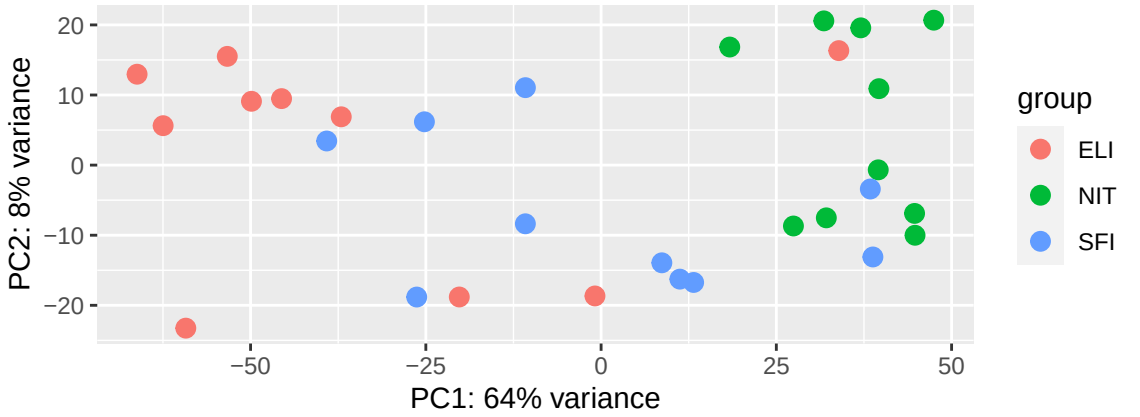

Figure 3: Visualización de las dos primeras componentes principales del PCA usando los valores transformados de VST

el 64%. Observamos como en el caso anterior, que los grupos `NIT` y `ELI`, en menor medida, se encuentran agrupados. La variabilidad de la primera componente tiene una contribución muy alta del grupo `NIT`, que como podemos observar sitúa todas las muestras a la derecha de la gráfica, mientras que el grupo `ELI` se encuentra mayoritariamente a la izquierda. En cambio, las muestras de `SFI` se encuentran más repartidas por el centro del gráfico.

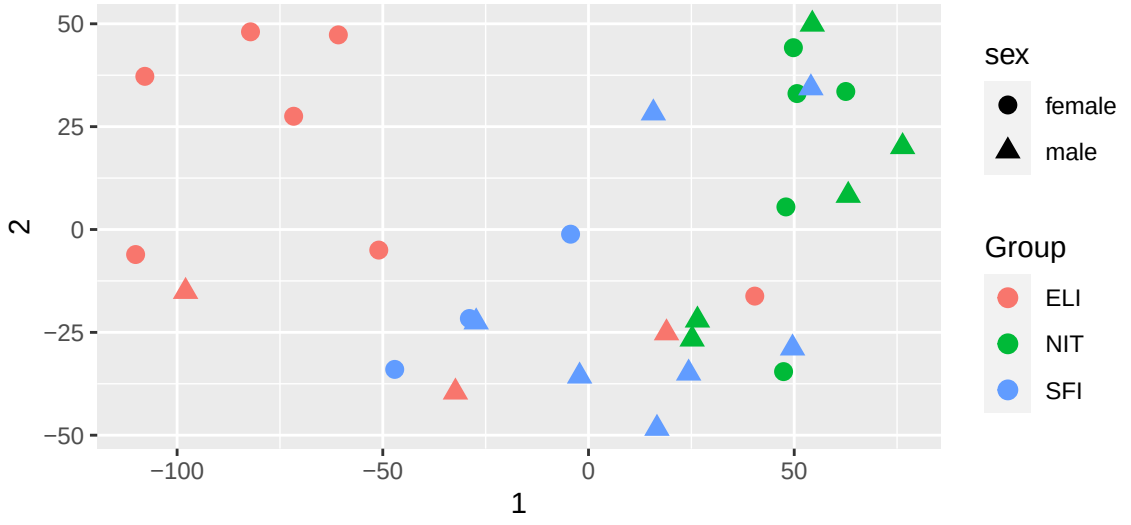

Figure 4: Visualización de las distancias con una matriz basada en MDS usando los valores transformados de VST

En el gráfico de la Figura 4, se muestra la distribución de los grupos empleando el Classical multidimensional scaling (MDS) para la matriz de distancias de los datos transformados de VST. El gráfico presenta un aspecto muy parecido al plot de PCA anterior (Figura 3), aunque en éste, además, se ha añadido la diferenciación por sexo, pero observamos que no hay un patrón claro entre las muestras de hombre o mujer.

Identificación de genes diferencialmente expresados

El parámetro `log2FoldChange` estima el cambio de expresión del gen debido al tratamiento, en este caso a los distintos métodos de infiltración. El parámetro `padj` corresponde al p-valor ajustado por el método Benjamini-Hochberg (BH). Con el fin de determinar que genes son signifcativos, podemos determinar un *threshold* para estos parámetros.

Con el fin de ser lo restrictivos, establecemos un *threshold* de 0.05 para el p-valor ajustado, mientras que dejamos el *threshold* de `log2FoldChange` por defecto.

Podemos realizar tres contrastes que comparen 2 a 2, los 3 grupos de infiltración en tiroides.

a) SFI-NIT

```
summary(SFIvsNIT)
```

```
##
```

```
## out of 43520 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)       : 588, 1.4%
## LFC < 0 (down)     : 72, 0.17%
## outliers [1]       : 0, 0%
## low counts [2]     : 16880, 39%
## (mean count < 3)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

En este caso observamos que los porcentajes de `log2FoldChange` para el resumen del análisis son bastante bajos, de modo que esperamos que la diferencia de tratamiento tenga poca incidencia en la expresión génica. Como hemos visto en la distribución de las Figuras 3 y 4, las muestras de SFI presentan mayor variabilidad y en consecuencia, se reduce el impacto en la expresión génica entre las dos condiciones comparadas. En este caso, aplicando los criterios establecidos, obtenemos 660 genes diferencialmente expresados y se descartarían 25985 genes.

Table 4: Genes diferencialmente expresados (downregulated) comparando las muestras de SFI y NIT (padj < 0.05). Tabla ordenada por log2FoldChange.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSG00000115602	623.62163	-2.41019	0.65178	-3.69786	0.00022	0.01410
ENSG00000249138	13.46636	-2.39713	0.64709	-3.70451	0.00021	0.01376
ENSG00000170439	156.44821	-2.31964	0.50838	-4.56276	0.00001	0.00078
ENSG00000164326	79.26190	-2.29781	0.68003	-3.37899	0.00073	0.03495
ENSG00000185186	56.11319	-2.29477	0.67659	-3.39169	0.00069	0.03415

Table 5: Genes diferencialmente expresados (upregulated) comparando las muestras de SFI y NIT (padj < 0.05). Tabla ordenada por log2FoldChange en orden inverso.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSG00000211658	143.48962	7.48569	1.32166	5.66387	0	1e-05
ENSG00000211667	50.70898	7.24720	1.15794	6.25868	0	0e+00
ENSG00000211904	170.49491	7.20944	1.09925	6.55849	0	0e+00
ENSG00000211942	200.90884	6.84151	1.23844	5.52428	0	2e-05
ENSG00000211676	36.95513	6.83726	1.14339	5.97980	0	0e+00

En las Tablas 4 y 5 se destacan los 5 genes que en condiciones de infiltración SFI o NIT presentan un expresión más significativa, bien sea upregulated o downregulated. El contraste de significancia presentan un p-valor ajustado menor a 0.05 y se encuentran ordenadas de acuerdo al `log2FoldChange`. La primera columna de estas tablas indica el código del gen en GenCode.

b) ELI-NIT

```
summary(ELIvsNIT)
```

```
##
## out of 43520 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)       : 3394, 7.8%
## LFC < 0 (down)     : 1312, 3%
```

```
## outliers [1]       : 0, 0%
## low counts [2]     : 12661, 29%
## (mean count < 1)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

En este caso observamos que los porcentajes de `log2FoldChange` para el resumen del análisis son algo mayor que en en caso anterior, especialmente en la expresión de genes upregulated, de modo, que esperamos mayor número de genes significativos. En este caso, en las Figuras 3 y 4 veíamos que las muestras de estos grupos se encontraban en posiciones opuestas, de modo que podemos intuir que el método de infiltración NIT y ELI provocan mayores cambios en la expresión génica. En consecuencia, aplicando los criterios establecidos, obtenemos 4706 genes diferencialmente expresados y se descartarían 26158 genes.

Table 6: Genes diferencialmente expresados (downregulated) comparando las muestras de ELI y NIT (padj < 0.05). Tabla ordenada por log2FoldChange.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSG00000250606	71.17927	-3.59914	1.14753	-3.13643	0.00171	0.01608
ENSG00000174417	3.11053	-3.53171	1.06424	-3.31854	0.00090	0.00970
ENSG00000232135	1.40549	-3.39745	1.27207	-2.67081	0.00757	0.04972
ENSG00000108688	1.33062	-3.24021	1.14074	-2.84044	0.00451	0.03377
ENSG00000233517	3.67694	-3.13938	0.97805	-3.20985	0.00133	0.01313

Table 7: Genes diferencialmente expresados (upregulated) comparando las muestras de ELI y NIT (padj < 0.05). Tabla ordenada por log2FoldChange en orden inverso.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSG00000211658	143.48962	9.65645	1.32093	7.31034	0	0
ENSG00000170054	55.73579	9.44326	1.36753	6.90533	0	0
ENSG00000223350	194.16760	9.42116	1.16084	8.11582	0	0
ENSG00000156234	820.91683	9.34286	0.99497	9.39013	0	0
ENSG00000128438	511.88078	9.18444	0.88577	10.36887	0	0

En las Tablas 6 y 7 se destacan los 5 genes que en condiciones de infiltración ELI o NIT presentan un expresión más significativa, bien sea upregulated o downregulated. El contraste de significancia presentan un p-valor ajustado menor a 0.05 y se encuentran ordenadas de acuerdo al `log2FoldChange`. La primera columna de estas tablas indica el código del gen en GenCode.

c) ELI-SFI

```
summary(ELIvsSFI)
```

```
##
## out of 43520 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)       : 1324, 3%
## LFC < 0 (down)     : 295, 0.68%
## outliers [1]       : 0, 0%
## low counts [2]     : 14349, 33%
## (mean count < 2)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

En este caso observamos que los porcentajes de `log2FoldChange` para el resumen del análisis son igualmente bajos como en la primera comparación, de modo que esperamos que la diferencia de tratamiento tenga poca incidencia en la expresión génica. Como hemos visto en la distribución de las Figuras 3 y 4, las muestras de SFI presentan mayor variabilidad y en consecuencia, se reduce el impacto en la expresión génica entre las dos condiciones comparadas. En este caso, aplicando los criterios establecidos, obtenemos 1619 genes diferencialmente expresados y se descartarían 27557 genes.

Table 8: Genes diferencialmente expresados (downregulated) comparando las muestras de ELI y SFI (padj < 0.05). Tabla ordenada por log2FoldChange.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSG00000110680	347.86197	-4.97878	1.45119	-3.43082	6e-04	0.01696
ENSG00000197978	6.60268	-4.79194	0.97995	-4.88998	0e+00	0.00019
ENSG00000233517	3.67694	-4.07249	0.96178	-4.23434	2e-05	0.00168
ENSG00000251676	4.07278	-3.35918	0.90474	-3.71288	2e-04	0.00787
ENSG00000054803	70.61294	-3.25433	0.78849	-4.12729	4e-05	0.00231

Table 9: Genes diferencialmente expresados (upregulated) comparando las muestras de ELI y SFI (padj < 0.05). Tabla ordenada por log2FoldChange en orden inverso.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSG00000160505	19.01557	5.83391	0.98319	5.93365	0.00000	0.00000
ENSG00000170054	55.73579	5.54831	1.22946	4.51279	0.00001	0.00071
ENSG00000261030	2.35732	5.30586	1.22824	4.31989	0.00002	0.00130
ENSG00000224187	6.41677	5.22752	1.73466	3.01357	0.00258	0.04756
ENSG00000211912	3.11298	5.15344	1.47487	3.49417	0.00048	0.01420

En las Tablas 8 y 9 se destacan los 5 genes que en condiciones de infiltración ELI o SFI presentan un expresión más significativa, bien sea upregulated o downregulated. El contraste de significancia presentan un p-valor ajustado menor a 0.05 y se encuentran ordenadas de acuerdo al `log2FoldChange`. La primera columna de estas tablas indica el código del gen en GenCode.

Anotación de los resultados

- **Plot del gen más significativo entre tratamientos**

En las gráficas de plotCounts (Figuras 5, 6 y 7) se visibilizan los valores de *counts* para el gen más significativo, es decir, para el gen con el p-valor ajustado menor en la comparación de los grupos. En los 3 casos, destacan dos aspectos: a) los valores de ELI son más altos que en los otros dos grupos y los de NIT tienden a ser los más bajos y b) las mayores diferencias de valores se observan entre ELI y NIT, dado que SFI suele tener valores intermedios.

- **Gene Clustering**

En la Figura 8 se muestra la expresión de los 20 genes con mayor variabilidad entre las muestras que se encuentran anotadas con el grupo al que pertenecen. Se observa que una gran mayoría de genes, 14 de los 20, presentan una expresion *downregulated* especialmente en muestas del grupo NIT, mientras que para el resto se encuentran generalmente *upregulated*. En cambio, el primer cluster de genes (5 primeras filas) se expresan de forma diferencial en parte de las muestras NIT y SFI.

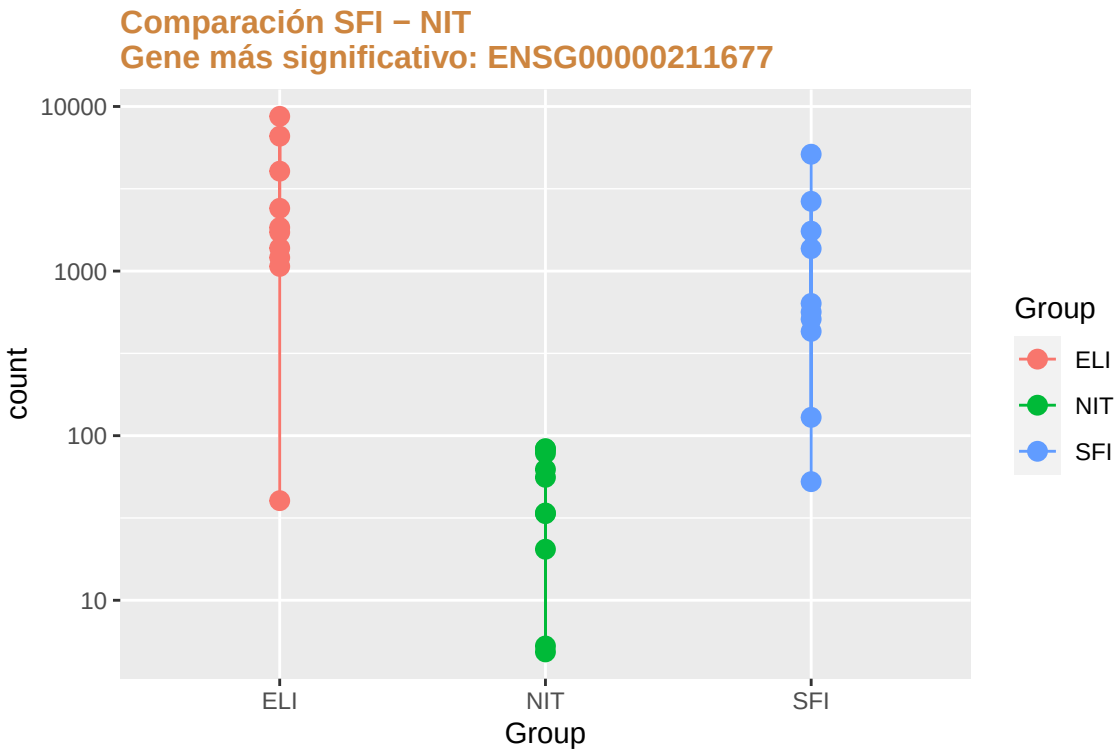

Figure 5: plotCounts del gen maś significativo en cada una de las comparaciones de los tratamientos

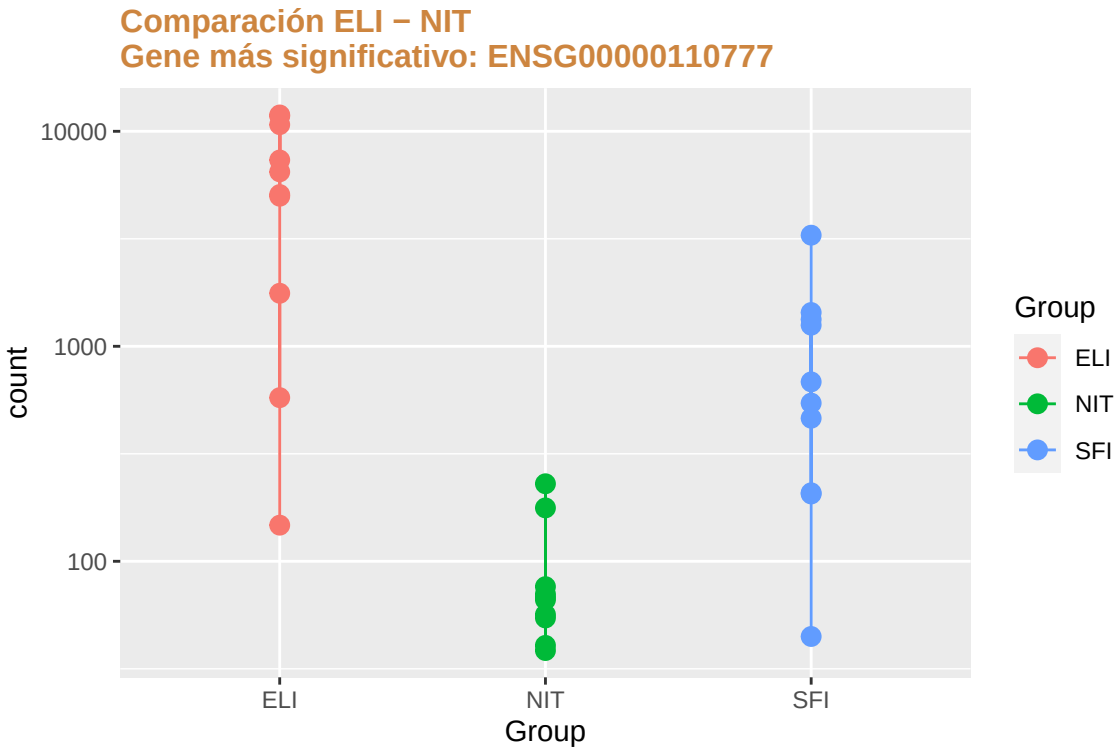

Figure 6: plotCounts del gen maś significativo en cada una de las comparaciones de los tratamientos

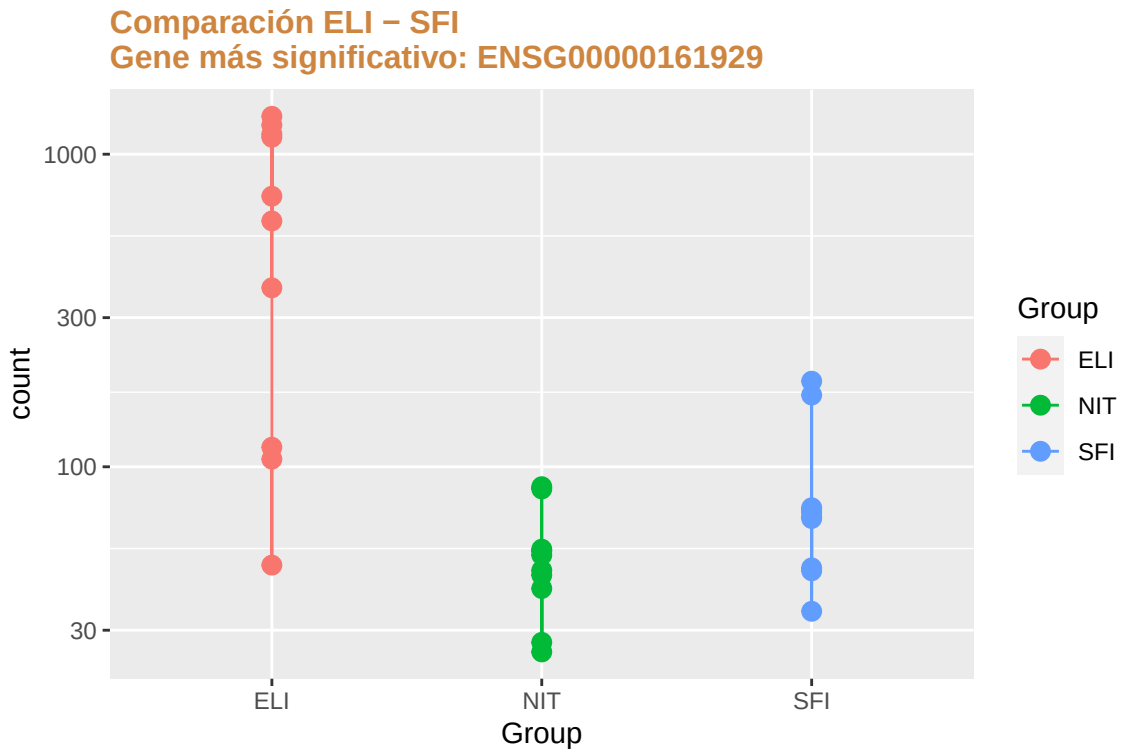

Figure 7: plotCounts del gen maś significativo en cada una de las comparaciones de los tratamientos

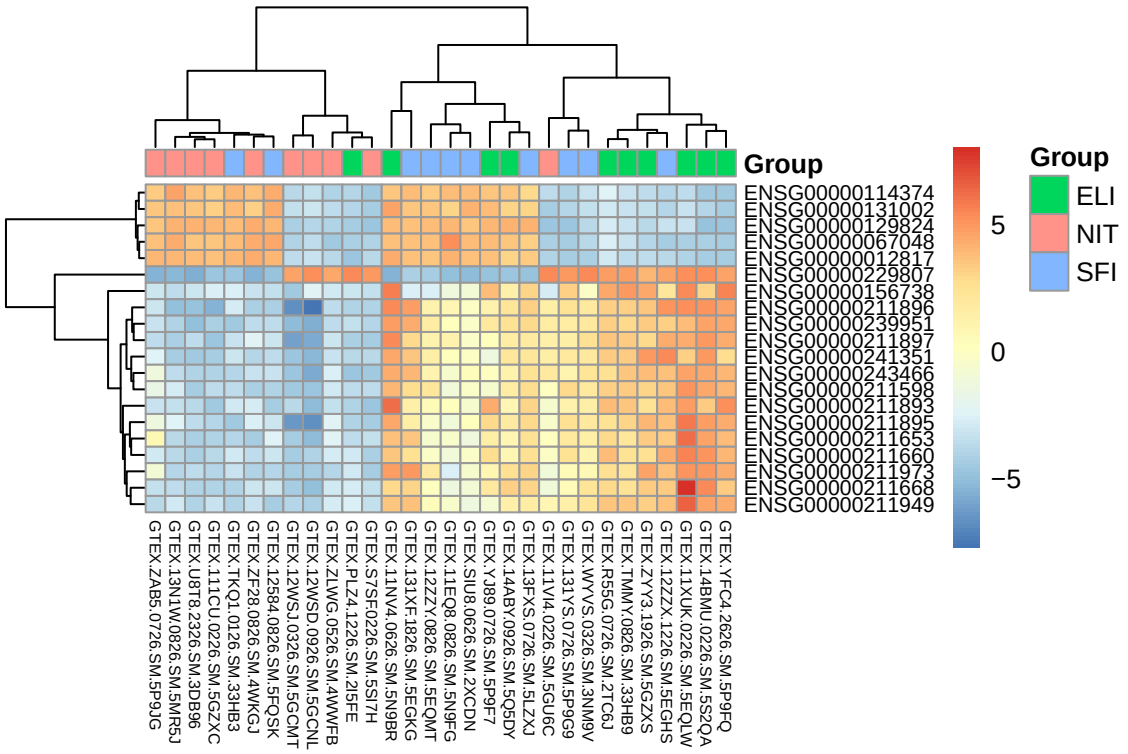

Figure 8: Cluster de los genes con mayor variabilidad entre las muestras

Análisis de significación biológicas

Una vez obtenemos la lista de genes expresados diferencialmente entre dos condiciones, debemos interpretar su relevancia biológica, es decir, conocer en que rutas metabólicas están implicados para conocer su función. Este análisis lo realizamos con la ayuda del paquete `clusterProfiler` que nos permite conocer dichas rutas de acuerdo al `Entrez ID` o al `Symbol` de cada gen.

En las siguientes Tablas 10, 11 y 12 se han anotado el `Symbol`y el `Entrez ID` para cada uno de los genes significativamente diferenciados entre las distintas comparaciones. Aunque lamentablemente, algunos códigos no se encuentran en la base de datos de Bioconductor consultada (`org.Hs.eg.db` y `EnsDb.Hsapiens.v86`) probablemente dado que son transcritos.

a) SFI-NIT

Table 10: Genes diferencialmente expresados comparando las muestras de SFI y NIT (padj < 0.05)

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol	entrez
ENSG00000211677	1423.0473	4.85682	0.62787	7.73542	0	0	IGLC2	NA
ENSG00000128438	511.8808	6.33730	0.88628	7.15043	0	0	TBC1D27	NA
ENSG00000170476	1696.3966	4.87931	0.69367	7.03408	0	0	MZB1	51237
ENSG00000211899	10346.3590	4.71479	0.68705	6.86241	0	0	IGHM	NA
ENSG00000143297	1938.0791	4.77007	0.69963	6.81799	0	0	FCRL5	83416

En las Figura 9 y 10, observamos que gran parte de los genes diferenciados se encuentran involucrados en la unión a antígeno y al receptor de la immunoglobulina.

b) ELI-NIT

Table 11: Genes diferencialmente expresados comparando las muestras de ELI y NIT (padj < 0.05)

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol	entrez
ENSG00000110777	2371.8056	6.12554	0.56233	10.89322	0	0	POU2AF1	5450
ENSG00000205056	339.7410	7.28254	0.66638	10.92852	0	0	RP11-693J15.5	100507616
ENSG00000083454	1269.6090	6.42334	0.59305	10.83095	0	0	P2RX5	5026
ENSG00000156738	6391.7116	8.25474	0.77052	10.71322	0	0	MS4A1	931
ENSG00000132704	797.1009	8.64578	0.81008	10.67275	0	0	FCRL2	79368

En las Figura 11 y 12, observamos que gran parte de los genes diferenciados se encuentran involucrados en la unión a antígeno y al receptor de la immunoglobulina. Además en este caso, también destacan con p-valores más altos pero significativos, los genes implicados en *cell adhesion molecule binding*, *small GTPase binding* y *Ras GTPase binding*.

c) ELI-SFI

Table 12: Genes diferencialmente expresados comparando las muestras de ELI y SFI (padj < 0.05)

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol	entrez
ENSG00000161929	272.9008	3.02864	0.43093	7.02808	0	0	SCIMP	388325
ENSG00000169679	141.1118	2.75568	0.42261	6.52060	0	0	BUB1	699
ENSG00000111913	1628.3603	2.55475	0.39304	6.50004	0	0	FAM65B	9750
ENSG00000157456	120.1114	2.42791	0.38156	6.36313	0	0	CCNB2	9133
ENSG00000170006	391.1272	2.37179	0.37763	6.28078	0	0	TMEM154	201799

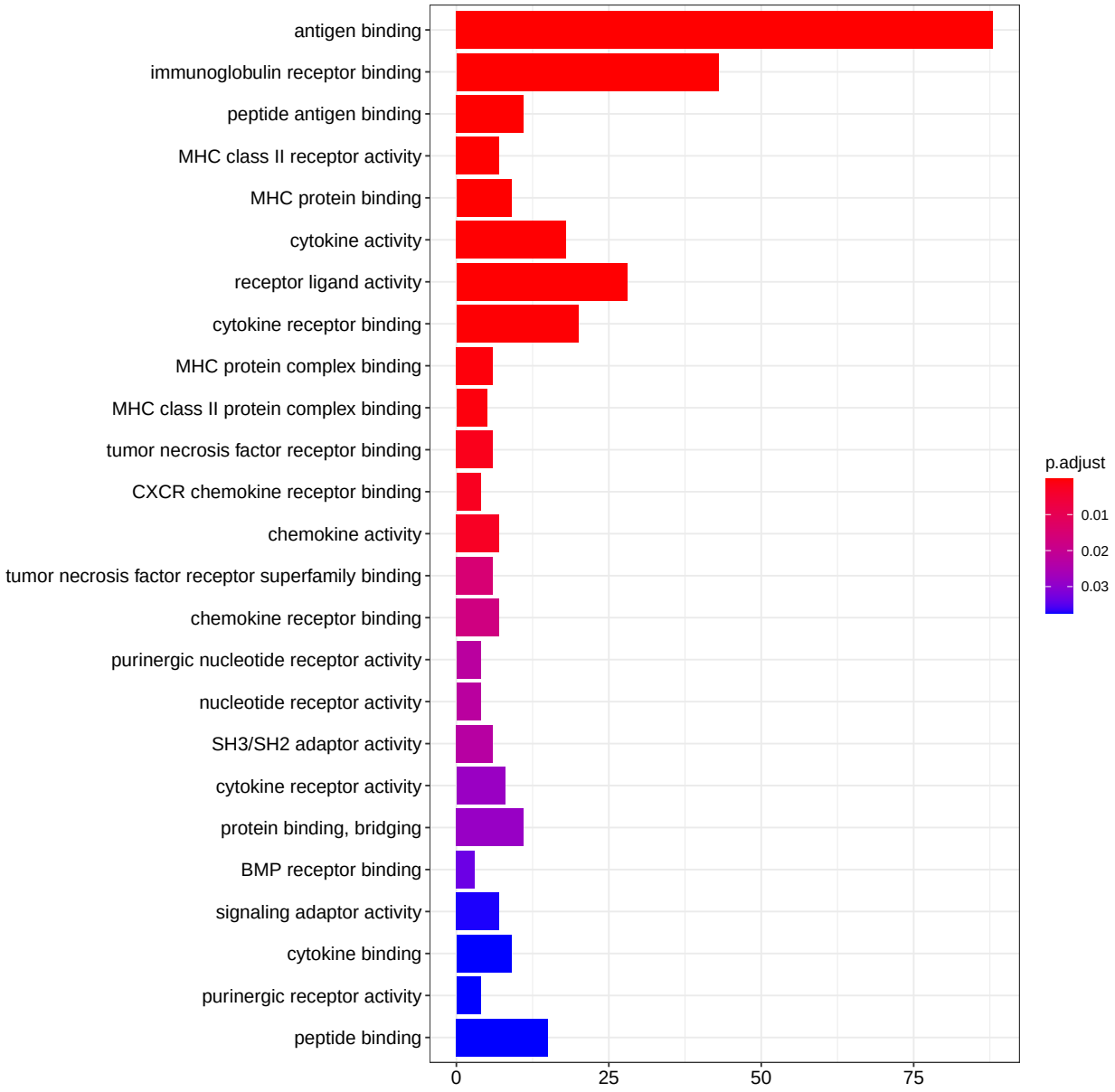

Figure 9: Barplot para los términos de EnrichmentGO para el análisis del tipo de infiltración SFI vs NIT en la expresión génica

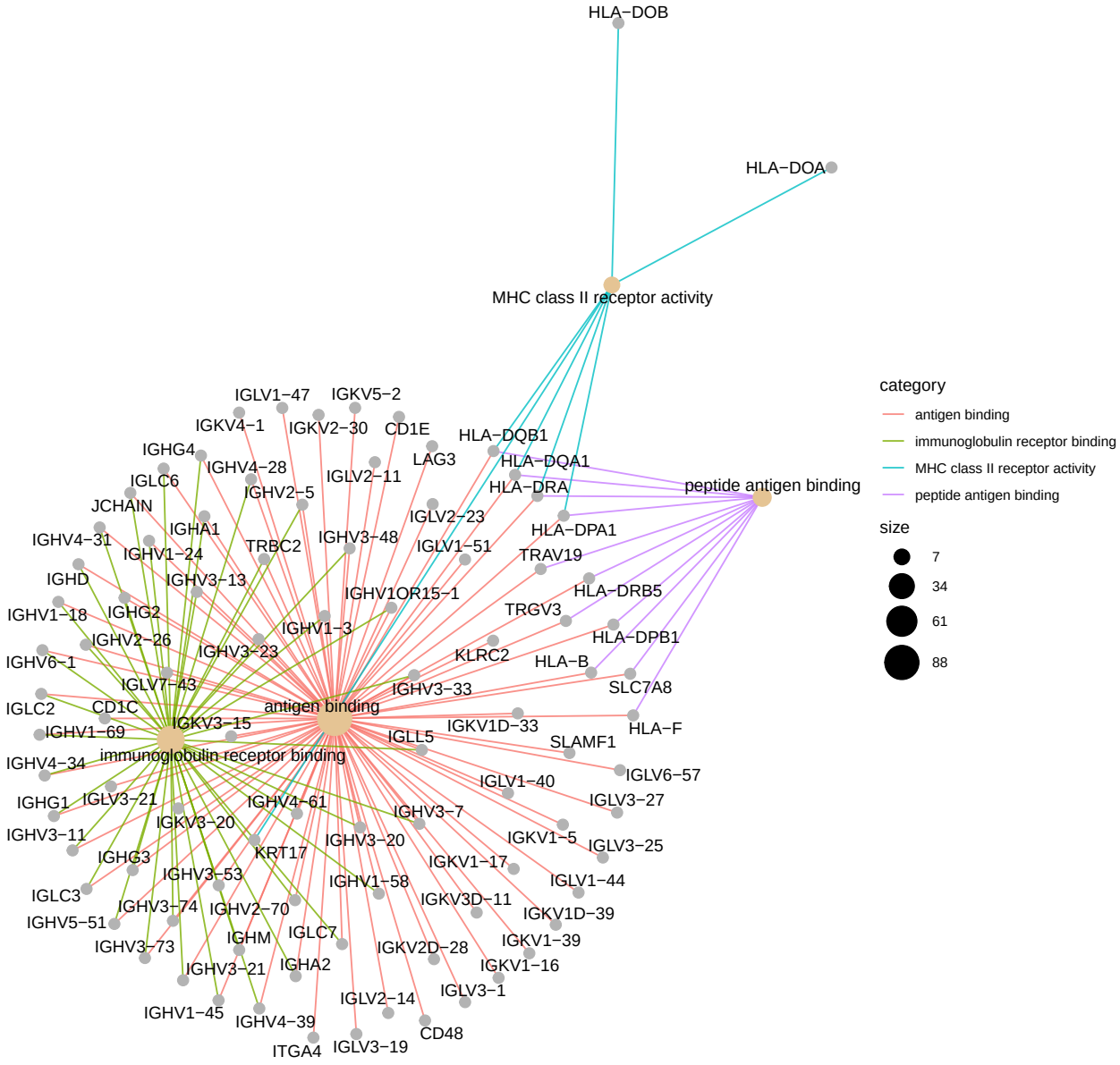

Figure 10: Red para los términos de EnrichmentGO para el análisis del tipo de infiltración SFI vs NIT en la expresión génica

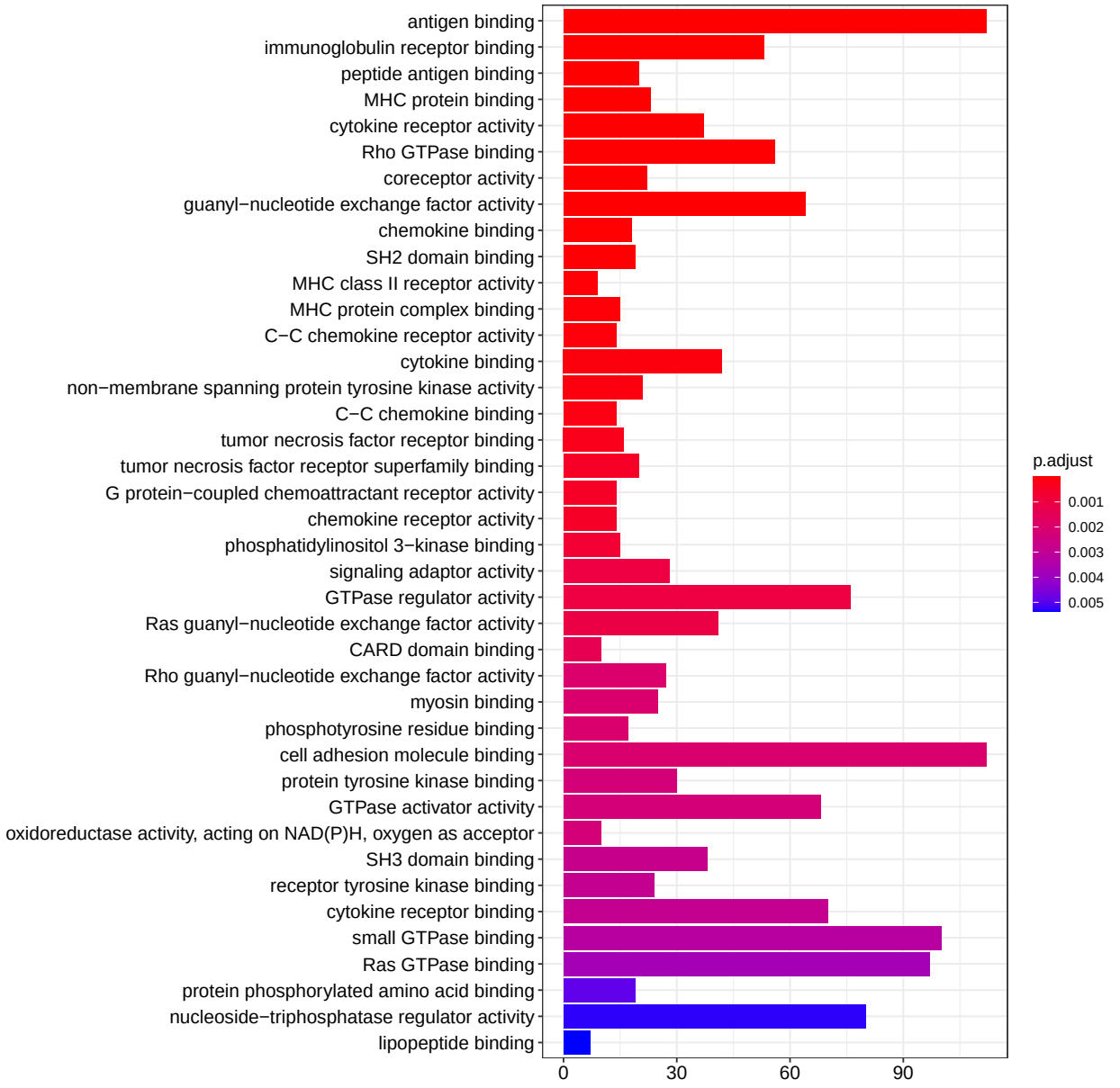

Figure 11: Barplot para los términos de EnrichmentGO para el análisis del tipo de infiltración ELI vs NIT en la expresión génica

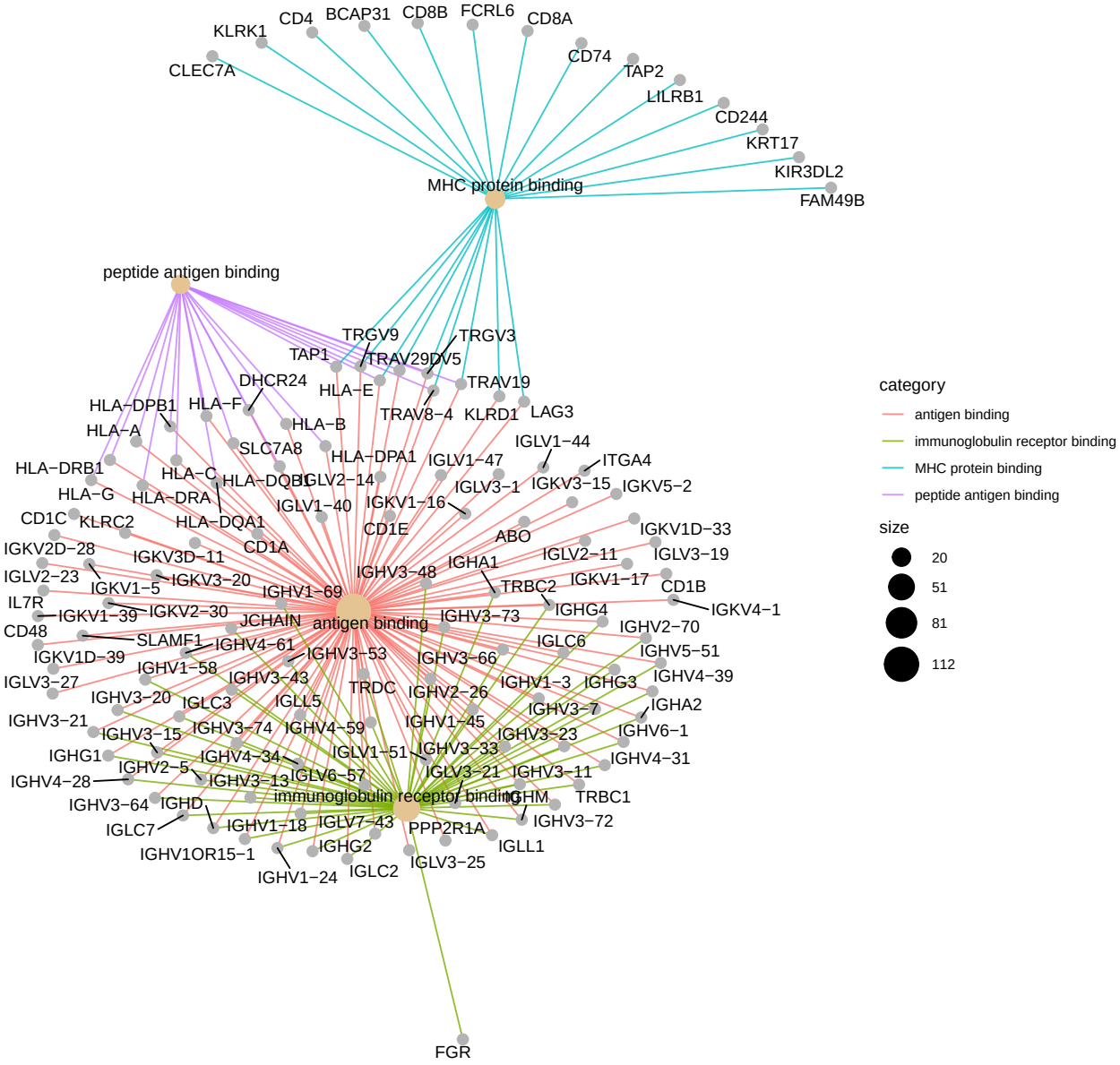

Figure 12: Red para los términos de EnrichmentGO para el análisis del tipo de infiltración ELI vs NIT en la expresión génica

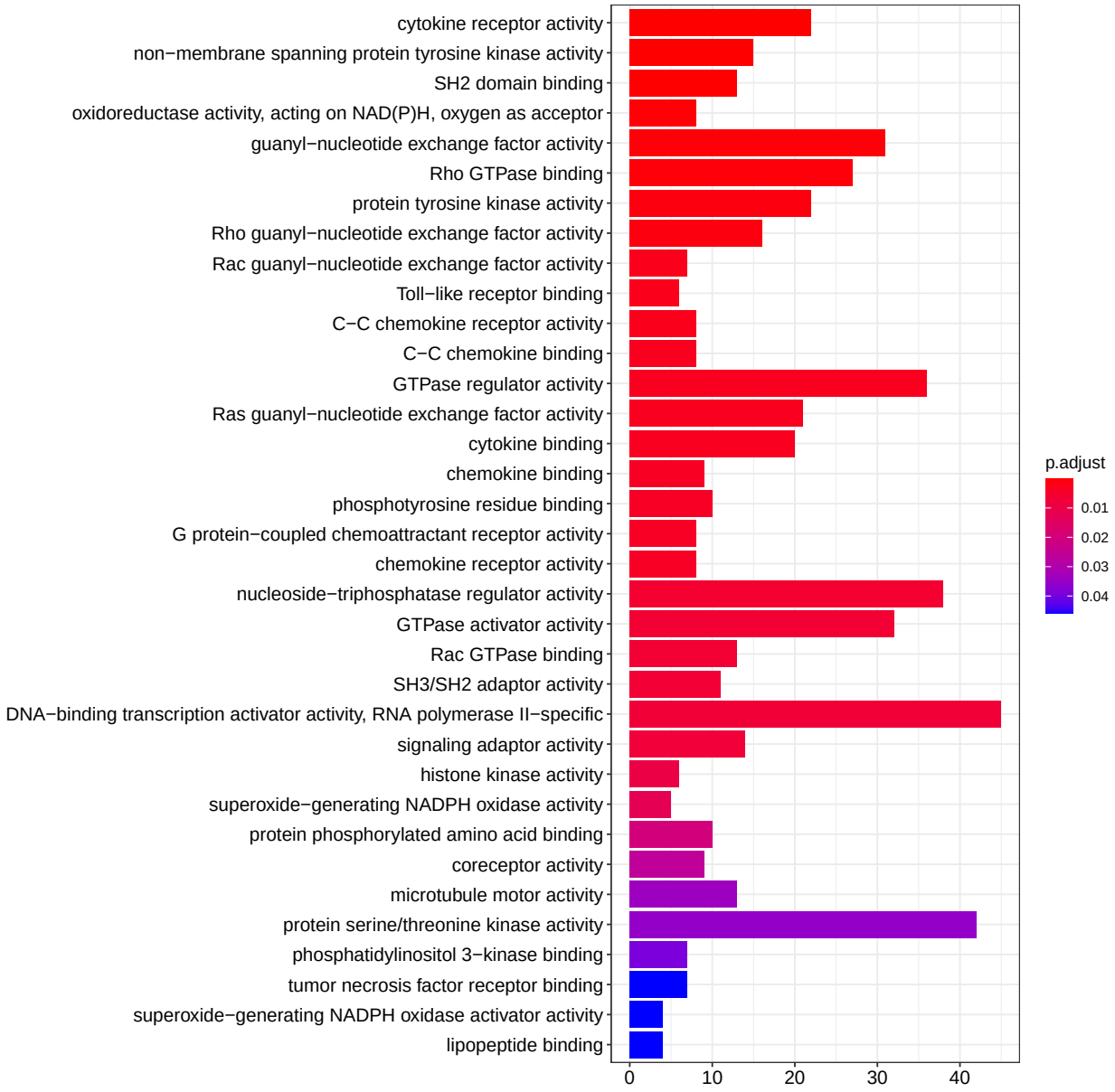

Figure 13: Barplot para los términos de EnrichmentGO para el análisis del tipo de infiltración ELI vs SFI en la expresión génica

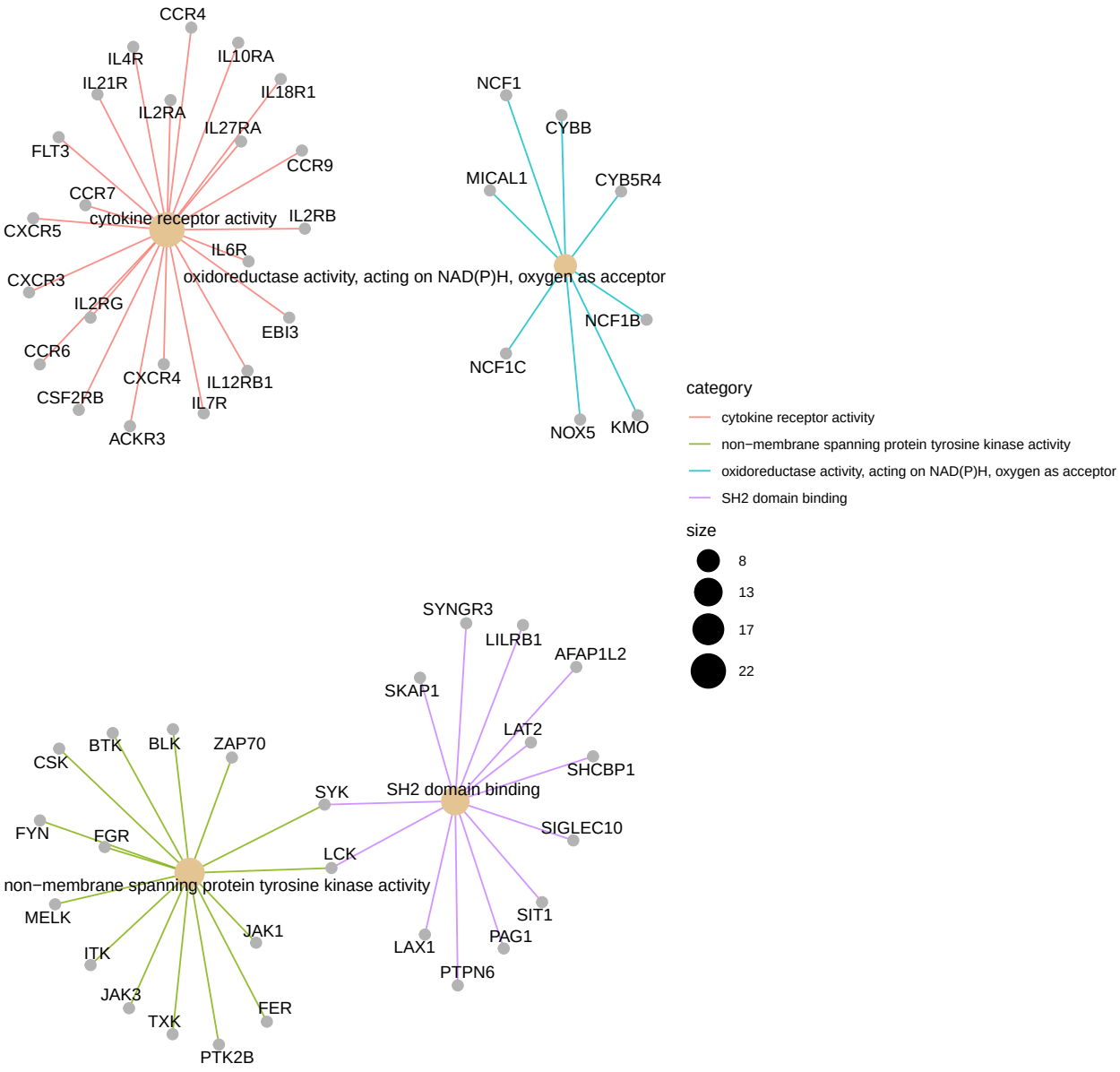

Figure 14: Red para los términos de EnrichmentGO para el análisis del tipo de infiltración ELI vs SFI en la expresión génica

En las Figura 13 y 14, observamos que en este caso, la función de los genes significativos se encuentra más repartida, destacando en este caso a diferencia de las comparaciones anteriores, *nucleoside-triphosphatase regulator activity*, *DNA-binding transcription activator activity, RNA polymerase II-specific* y *protein serine/threonine kinase activity*.

INFORME DEL ANÁLISIS

Identificación de la expresión transcripcional en muestras de tiroides

Abstract

El método de infiltración en tiroides se ha observado que puede afectar a la expresión genética. Para ello, se ha hecho un contaje de los tránscritos de 292 muestras, y se ha realizado un análisis de un subset de 30 de éstas distribuidas en los tres métodos de infiltración (*i.e.*, SFI, NIT, ELI). Tras el análisis de la expresión de genes, se ha obtenido que hay una transcripción significativa principalmente entre los grupos NIT y ELI. Además, se ha observado que los genes se encuentra implicados en rutas metabólicas de unión a antígeno y al receptor de la immunoglobulina.

Introducción y Objetivo

El objetivo principal de este estudio es **analizar el efecto en la expresión genética debido a la infiltración mediante métodos en tiroides.**

No obstante, para una mejor contextualización de los resultados, sería necesario mayor información del estudio en el que se han recogido los datos analizados.

Materiales y Métodos

- Diseño del estudio

El estudio presenta un factor con 3 niveles, correspondiente a 3 métodos de infiltración en tiroides: a) *Extensive lymphoid infiltrates* (ELI); b) *Not infiltrated tissues* (NIT); c) *Small focal infiltrates* (SFI). La muestra original de participantes del estudio se ha reducido a una muestra de 30 sujetos, distribuyendo 10 réplicas para cada uno de los grupos. Además se han anotado otras características como el sexo, el experimento o un código identificativo.

Para cada una de las muestras, se ha realizado un contaje de la expresión transcripcional de una serie de genes implicados en tiroides, en total se han analizado 56202 genes.

- Diseño computacional

EL análisis de los datos se ha llevado a cabo empleando la versión 3.6.1 de R, así como funciones y paquetes pertenecientes al proyecto Bioconductor destinados a un análisis RNAseq. Con el fin de optimizar el protocolo, se han seguido los tutoriales presentados recientemente para dicho análisis [1] [2] [3] [4].

Resultados y Discusión

El análisis de los datos se ha llevado a cabo en primer lugar analizando la calidad de los datos. De modo que el primer paso es realizar un filtraje no específico eliminando aquellos genes para los que no se han obtenido contajes. De este filto, obtenemos 43525 genes que potencialmente pueden ser significativos. A continuación, se realiza una normalización de los datos usando el método *variance stabilizing transformation* (VST).

Seguidamente, se ha realizado un contraste en la que se ha especificado que los grupos a comparar son los tres métodos de infiltración, comparados dos a dos (SFI-NIT, ELI-NIT y ELI-SFI). Como resultado del test de significancia y aplicando cutoff para el p-valor ajustado con el método Benjamini and Hochberg menor a 0.05, se obtienen una serie de genes significativos para cada una de las comparaciones. Por último, se han

analizado los términos de enrichment de GO database para agrupar los genes significativos en función de la ruta metabólica en la se ven implicada dichos genes.

De acuerdo con los resultados, se observa que hay mayores diferencias en la expresión de los grupos ELI y NIT, en cambio, las muestras de SFI presentan mayor variabilidad. Por lo tanto, se han obtenido 660, 4706 y 1619 genes significativos para las comparaciones SFI-NIT, ELI-NIT y ELI-SFI, respectivamente. La mayor parte de los genes significativos se encuentran involucrados en funciones muy diversas que incluyen la unión a antígeno y al receptor de la immunoglobulina, cell adhesion molecule binding, small GTPase binding, Ras GTPase binding, nucleoside-triphosphatase regulator activity o DNA-binding transcription activator activity.

Conclusión

Tras este informe, se ha conseguido analizar el contaje de genes transcritos aplicando un protocolo de RNAseq con el fin de valorar la respuesta génica en muestras con diferentes métodos de infiltración en tiroides. Se ha comprendido la importancia de cada paso del análisis y se han obtenido múltiples figuras que nos dan una imagen de la calidad de los datos, de la significancia de los test de comparación y de la participación de genes sigficativos en rutas metabólicas. No obstante, se requiere mayor información del estudio con el fin de comprender el contexto de éste. Así como también se requiere una mejor comprensión de toda la información extraída de cada gráfico y un análisis más detallado de cada uno de los genes significativos.

References

[1] Michael I. Love, Simon Anders, Huber W. Analyzing rna-seq data with deseq2 2020. http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html.

[2] Ricardo Gonzalo Sanz, Sánchez-Pla A. Analyzing rna-seq data with deseq2 2020. https://github.com/ASPteaching/Omics_data_analysis-Case_study_2-RNA-seq.

[3] Rainer J. Generating an using ensembl based annotation packages 2020. https://www.bioconductor.org/packages/release/bioc/vignettes/ensembldb/inst/doc/ensembldb.html#8_using_ensdb_objects_in_the_annotationdbi_framework.

[4] Michael I. Love, Simon Anders, Vladislav Kim, Huber W. RNA-seq workflow: Gene-level exploratory analysis and differential expression 2019. http://master.bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html#aligning-reads-to-a-reference-genome.