

# Análisis de datos ómicos - PEC2

*Leticia Rodríguez Montes*

1/6/2020

## Contents

1) Abstract . . . . .	1
2) Objetivos . . . . .	1
3) Materiales y métodos . . . . .	2
3.1) Obtención de datos . . . . .	2
3.2) Muestreo aleatorio . . . . .	2
3.3) Prefiltrado . . . . .	3
3.4) Normalización y MDS plot . . . . .	3
3.5) Estimación de la dispersión . . . . .	3
3.6) Análisis de expresión diferencial . . . . .	4
3.7) Anotación de los resultados . . . . .	4
3.8) Representación gráfica de los resultados . . . . .	4
3.9) Análisis de significación biológica . . . . .	5
4) Resultados . . . . .	5
4.1) MDS plot . . . . .	5
4.2) Dispersión . . . . .	5
4.3) Análisis de genes diferencialmente expresados . . . . .	6
4.3.1) NIT vs ELI . . . . .	7
4.3.2) SFI vs ELI . . . . .	8
4.3.2) NIT vs SFI . . . . .	9
4.3 Comparación entre los distintos contrastes . . . . .	10
4.4) Análisis de significación biológica . . . . .	11
4.4.1) NIT vs ELI . . . . .	11
4.4.2) SFI vs ELI . . . . .	13
4.4.3) NIT vs SFI . . . . .	15
5) Discusión . . . . .	17
6) Referencias . . . . .	18

## 1) Abstract

La tiroides juega un papel clave en la producción de hormonas que regulan el metabolismo del ser humano. Tras analizar las diferencias en expresión génica entre tres grupos de muestras de pacientes (sin infiltrados (NIT), con pequeños infiltrados (SFI) o con infiltrados extensos(ELI)), se observa una upregulación de genes involucrados en respuesta inmune en las muestras con mayor nivel de infiltración.

## 2) Objetivos

La tiroides es una glándula endocrina que juega un papel fundamental en la producción de hormonas que regulan el metabolismo en el ser humano. Algunos de los trastornos mejor estudiados vinculados a esta glándula son el hipotiroidismo, hipertiroidismo o algunas afecciones autoinmunes como la enfermedad de Graves o la enfermedad de Hashimoto.

El objetivo de este estudio es comparar el efecto de diferentes tipos de infiltración en la tiroides de humanos. Las muestras utilizadas proceden del conjunto de datos GTEx (Genotype-Tissue Expression). Este proyecto

pretende ser un gran recurso para que la comunidad científica pueda estudiar la regulación y expresión génica en humanos y su relación con la variación genética. El portal del proyecto contiene una enorme cantidad de muestras de origen humano de multitud de tejidos y diferentes edades, razas y sexos que han sido rigurosamente genotipadas. En este caso en particular estamos utilizando muestras de tiroides que están clasificadas atendiendo a su nivel de infiltración de la siguiente manera:

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

### 3) Materiales y métodos

La dirección del repositorio de github con todo lo necesario para reproducir el análisis es : [https://github.com/Leticia314/PEC2\\_ADO.git](https://github.com/Leticia314/PEC2_ADO.git).

#### 3.1) Obtención de datos

Los datos utilizados en este estudio proceden del portal GTEx (<https://www.gtexportal.org/home/documentationPage>). En este set de datos hay un total de 292 muestras pertenecientes a tres grupos:

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

A partir de este set de datos se tomaron 10 muestras aleatorias de cada uno de los grupos utilizando la función “sample”:

#### 3.2) Muestreo aleatorio

```
#Cargar datos
library(readr)
library(tidyverse)

targets <- read_csv("./targets.csv")
counts <- read.csv( "./counts.csv", sep = ";", header=TRUE, row.names = 1)
library("dplyr")

library(edgeR)

set.seed(1221)

samples <- c(sample(targets$Sample_Name[targets$Group=="NIT"], size=10, replace = FALSE),
            sample(targets$Sample_Name[targets$Group=="SFI"], size=10, replace = FALSE),
            sample(targets$Sample_Name[targets$Group=="ELI"], size=10, replace = FALSE))

samples_targets <- targets[targets$Sample_Name %in% samples,]
samples_targets$Sample_Name <- gsub("-", ".", samples_targets$Sample_Name)
samples <- gsub("-", ".", samples)

samples_counts <- counts[,samples]
```

Por lo tanto finalmente nuestro conjunto de datos cuenta con:

- 10 muestras de “Not infiltrated tissues” (NIT)

- 10 muestras de “Small focal infiltrates” (**SFI**)
- 10 muestras de “Extensive lymphoid infiltrates” (**ELI**)

### 3.3) Prefiltrado

Con nuestras 30 muestras ya seleccionadas se construyó un objeto DGEList (con la función DGEList del paquete “edgeR”). Teniendo en cuenta que muchos genes presentan un número de conteos muy bajos se ha decidido quitar las filas en las que los conteos por millón (cpm) son inferiores a 1 en al menos 3 muestras. De esta forma estamos eliminando genes donde la probabilidad de encontrar diferencias de expresión es muy baja, ya que en estos genes el ruido de Poisson es tan alto por el bajo número de conteos que es mucho más difícil detectar verdaderas diferencias de relevancia biológica. No obstante, de no eliminarlos, tendrían mucha influencia en el resultado final del análisis de expresión diferencial ya que aumentarían sustancialmente el número de test realizados, lo que dificulta la capacidad de alcanzar significancia estadística tras aplicar los métodos de corrección múltiple.

```
all(samples_targets$Sample_Name %in% colnames(samples_counts))

## [1] TRUE

all(samples_targets$Sample_Name == colnames(samples_counts))

## [1] FALSE

samples_counts <- samples_counts[, samples_targets$Sample_Name]
all(samples_targets$Sample_Name == colnames(samples_counts))

## [1] TRUE

dds <- DGEList(counts=samples_counts, group=samples_targets$Group)
nrow(dds)

## [1] 56202

keep <- rowSums(cpm(dds) >1) >= 3
dds <- dds[keep, ]
nrow(dds)

## [1] 18960
```

Con este prefiltrado se pasó de tener 56202 a 18960 genes.

### 3.4) Normalización y MDS plot

Posteriormente se realizó un paso de normalización para tener en cuenta las diferencias en el tamaño de las librerías utilizando la función “calcNormFactors” del paquete “edgR”. El MDS plot se realizó utilizando la función “plotMDS” del paquete “limma”.

```
dds<-calcNormFactors(dds)
```

### 3.5) Estimación de la dispersión

Para estimar las dispersiones común y gen-específica se utilizaron las funciones “estimateCommonDisp” y “estimateTagwiseDisp” del paquete “edgR”.

```
dds <- estimateCommonDisp(dds, verbose=T)
```

```
## Disp = 0.23351 , BCV = 0.4832
```

```
dds <- estimateTagwiseDisp(dds)
```

### 3.6) Análisis de expresión diferencial

Para detectar los genes diferencialmente expresados entre los diferentes grupos de muestras se utilizaron las funciones exactTest y topTags del paquete edgeR. Como método de corrección por comparaciones múltiples se utilizó el método de “Benjamini Hochberg”.

```
et_NITvsELI <- exactTest(dds, pair = c("ELI", "NIT"))
res_NITvsELI <- topTags(et_NITvsELI, n=nrow(dds$counts), adjust.method="BH")$table

et_SFIvsELI <- exactTest(dds, pair = c("ELI", "SFI"))
res_SFIvsELI <- topTags(et_SFIvsELI, n=nrow(dds$counts), adjust.method="BH")$table

et_NITvsSFI <- exactTest(dds, pair = c("SFI", "NIT"))
res_NITvsSFI <- topTags(et_NITvsSFI, n=nrow(dds$counts), adjust.method="BH")$table
```

### 3.7) Anotación de los resultados

Para anotar los resultados del análisis de expresión diferencial se utilizó el paquete “biomaRt”.

```
library("biomaRt")
human<-useMart(host="www.ensembl.org", "ENSEMBL_MART_ENSEMBL", dataset="hsapiens_gene_ensembl")
attributes<-c("ensembl_gene_id", "entrezgene_id", "hgnc_symbol")

ensembl_names1<-gsub("\\..*", "", rownames(res_NITvsELI))
genemap1<-getBM(attributes, filters="ensembl_gene_id", values=ensembl_names1, mart=human)
idx1 <-match(ensembl_names1, genemap1$ensembl_gene_id)
res_NITvsELI$entrezgene <-genemap1$entrezgene_id [ idx1 ]
res_NITvsELI$hgnc_symbol <-genemap1$hgnc_symbol [ idx1 ]

ensembl_names2<-gsub("\\..*", "", rownames(res_SFIvsELI))
genemap2<-getBM(attributes, filters="ensembl_gene_id", values=ensembl_names2, mart=human)
idx2 <-match(ensembl_names2, genemap2$ensembl_gene_id)
res_SFIvsELI$entrezgene <-genemap2$entrezgene_id [ idx2 ]
res_SFIvsELI$hgnc_symbol <-genemap2$hgnc_symbol [ idx2 ]

ensembl_names3<-gsub("\\..*", "", rownames(res_NITvsSFI))
genemap3<-getBM(attributes, filters="ensembl_gene_id", values=ensembl_names3, mart=human)
idx3 <-match(ensembl_names3, genemap3$ensembl_gene_id)
res_NITvsSFI$entrezgene <-genemap3$entrezgene_id [ idx3 ]
res_NITvsSFI$hgnc_symbol <-genemap3$hgnc_symbol [ idx3 ]
```

### 3.8) Representación gráfica de los resultados

Para presentar los datos de una forma más comprensible y resumida, que permita entender su relevancia biológica, se utilizaron diferentes paquetes como “ggplot2” o “VennDiagram”.

### 3.9) Análisis de significación biológica

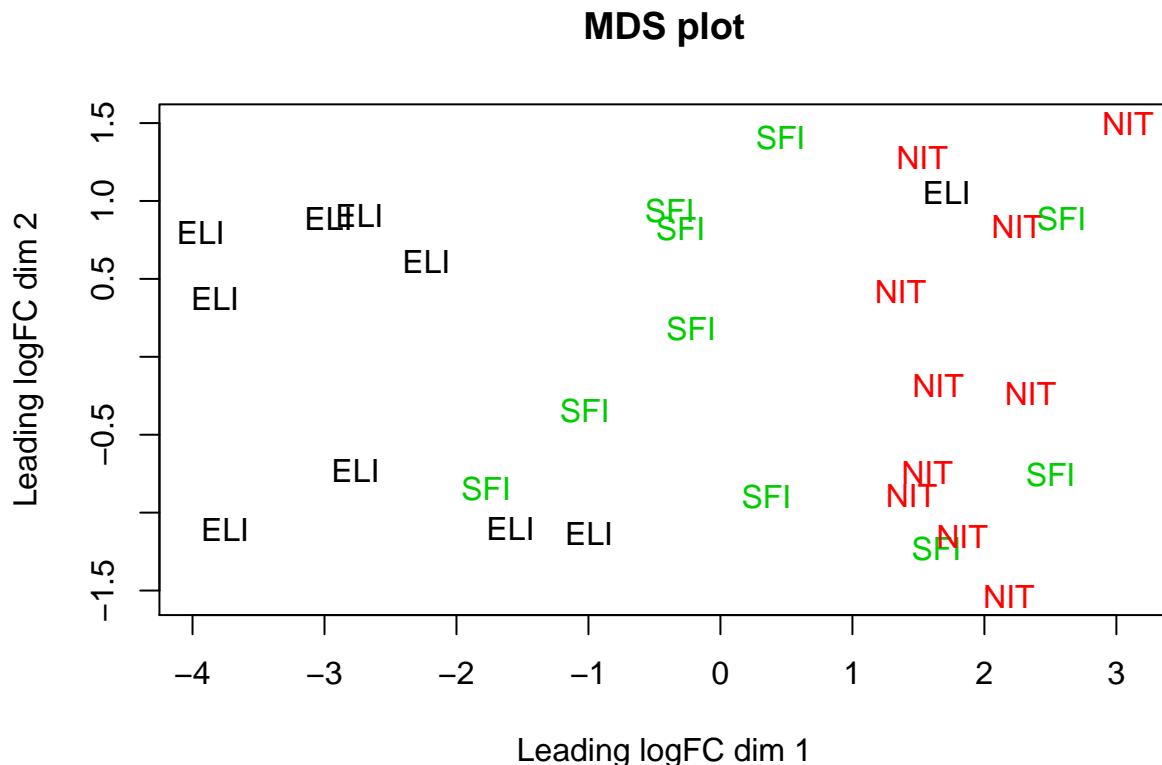
Para determinar la significación biológica de los resultados se decidió hacer un análisis de enriquecimiento. De esta forma se puede identificar si un proceso biológico o vía metabólica aparece con una frecuencia mayor o menor a la esperada en la lista de genes seleccionados que en la población total de genes. Para ello se utilizaron los paquetes “DOSE”, “pathview” y “clusterProfiler”. Para representar los resultados se utilizaron “dotplots” y “enrichment maps”, que organizan los términos GO que están enriquecidos formando una red que conecta los ejes con conjuntos de genes solapantes. De esta forma es fácil identificar módulos funcionales.

## 4) Resultados

### 4.1) MDS plot

Para tener una idea preliminar de la estructura de los datos se realizó un “MDS plot”.

```
plotMDS(dds, pch=1, col=as.numeric(dds$samples$group), labels = dds$samples$group )
title("MDS plot")
```

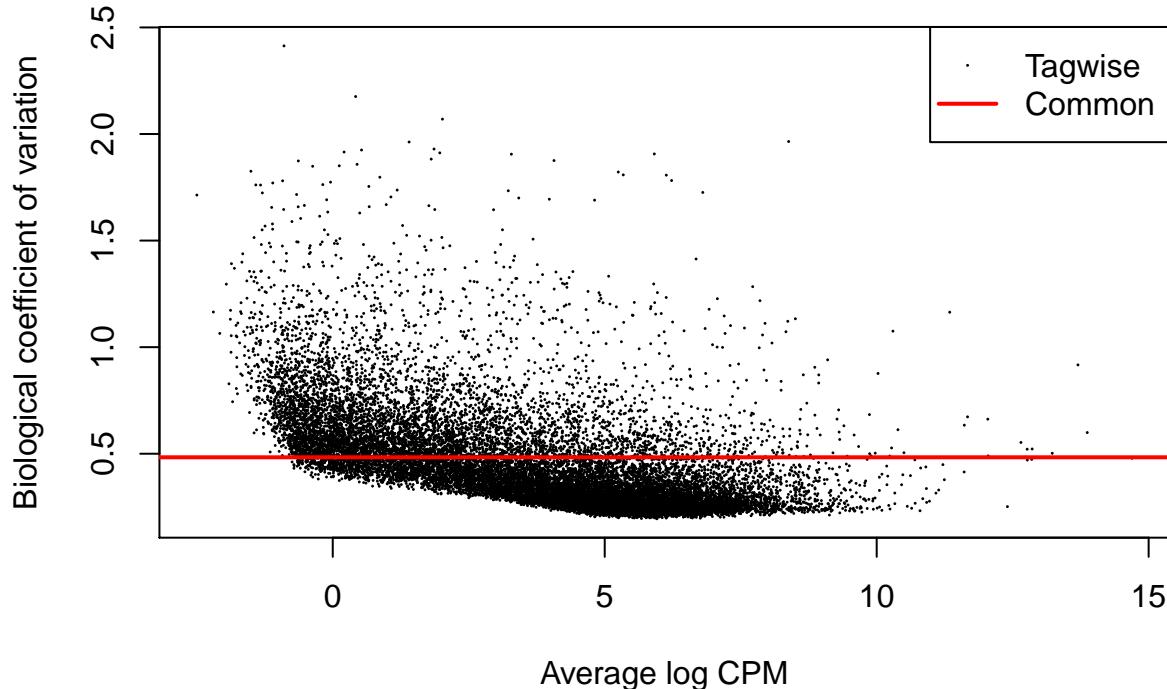


Salvo alguna excepción, las muestras parecen agruparse según el tipo de infiltración, lo cual sugiere que esta variable es una fuente de variación principal en nuestro set de datos.

### 4.2) Dispersión

La dispersión común estima el coeficiente de variación biológica (o BCV por sus siglas en inglés).

```
plotBCV(dds, col.tagwise="black")
```



En este caso la dispersión común es de Disp = 0.23351 y BCV = 0.4832. En este gráfico también se puede observar que el BCV de genes con baja expresión suele ser mucho más elevado que la dispersión común, como cabe esperar.

#### 4.3) Análisis de genes diferencialmente expresados

Estamos interesados en conocer los genes cuya expresión cambia entre las diferentes condiciones planteadas en el estudio, por lo que realizamos una selección de genes diferencialmente expresados. Se llevaron a cabo tres contrastes diferentes:

- NITvsELI
- SFIvsELI
- NITvsSFI

```
summary(decideTestsDGE(et_NITvsELI, p.value=0.05, lfc=1))
```

```
##           NIT-ELI
## Down      1753
## NotSig   16771
## Up       436
```

```
summary(decideTestsDGE(et_SFIvsELI, p.value=0.05, lfc=1))
```

```
##           SFI-ELI
## Down      1592
```

```

## NotSig    17019
## Up        349
summary(decideTestsDGE(et_NITvsSFI, p.value=0.05, lfc=1))

##          NIT-SFI
## Down      295
## NotSig   18655
## Up       10

```

Para representar los genes diferencialmente expresados en los diferentes contrastes se han realizado “volcano plots” donde :

- Los genes con log2FC mayor o menor a 1 o -1 están marcados en naranja.
- Los genes con log2FC mayor o menor a 1 o -1 y con FDR < 0.05 están marcados en verde.

#### 4.3.1) NIT vs ELI

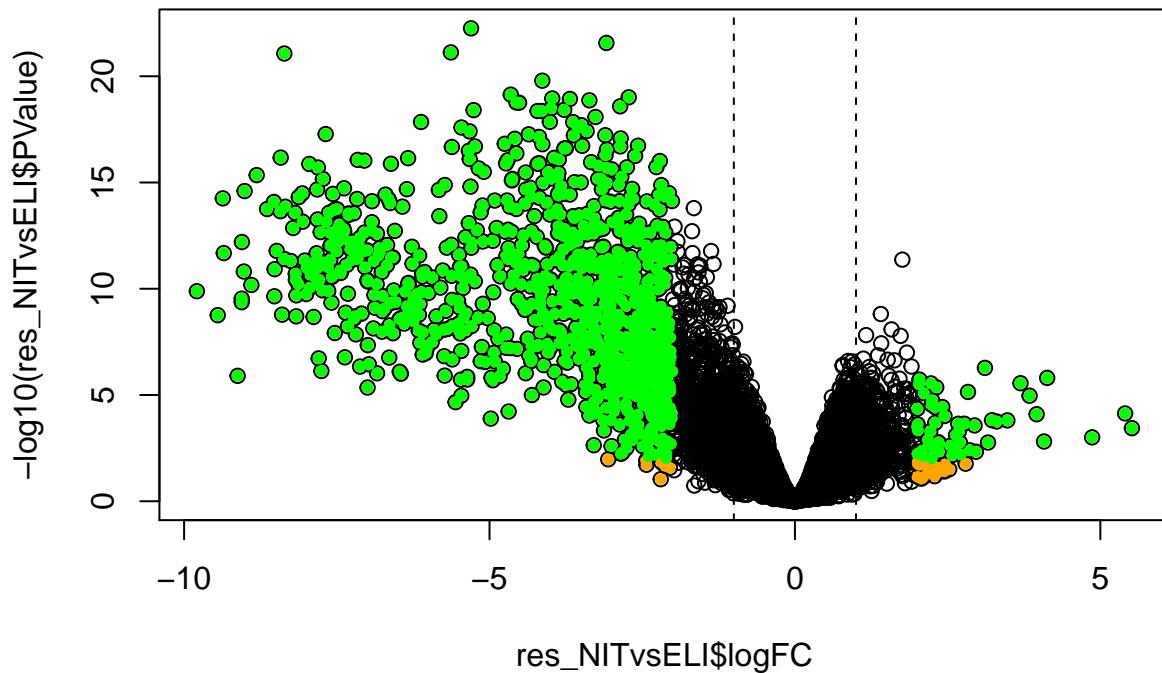
Al comparar las muestras NIT con las ELI obtenemos 1753 genes que están downregulados y 436 genes que están upregulados en NIT con respecto a ELI. Es el contraste en el que encontramos más genes diferencialmente expresados.

```

plot(res_NITvsELI$logFC, -log10(res_NITvsELI$PValue))
abline(v=c(-1,1), lty="dashed")
with(subset(res_NITvsELI, abs(logFC)>2), points(logFC, -log10(PValue),
                                                 pch=20, col="orange"))
with(subset(res_NITvsELI, FDR<.05 & abs(logFC)>2), points(logFC, -log10(PValue),
                                                 pch=20, col="green"))
title("Volcano plot NITvsELI")

```

### Volcano plot NITvsELI

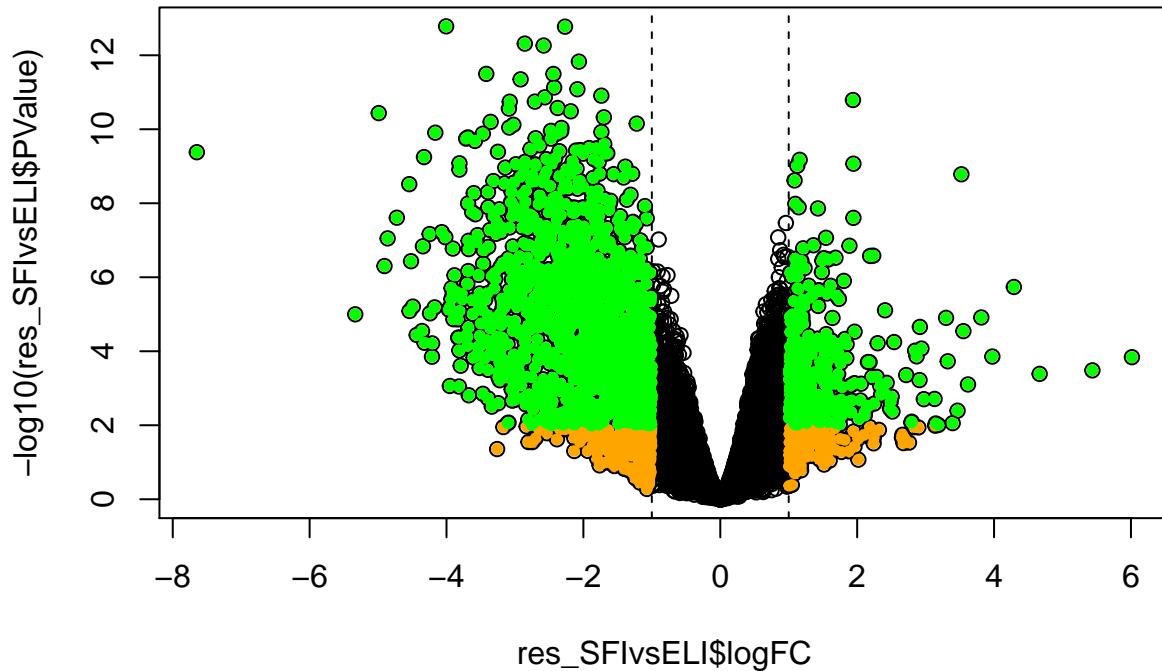


#### 4.3.2) SFI vs ELI

Al comparar las muestras SFI con las ELI obtenemos 1592 genes que están downregulados y 349 genes que están upregulados en SFI con respecto a ELI.

```
plot(res_SFIvsELI$logFC, -log10(res_SFIvsELI$PValue))
abline(v=c(-1,1), lty="dashed")
with(subset(res_SFIvsELI, abs(logFC)>1), points(logFC, -log10(PValue),
                                                 pch=20, col="orange"))
with(subset(res_SFIvsELI, FDR<.05 & abs(logFC)>1), points(logFC, -log10(PValue),
                                                 pch=20, col="green"))
title("Volcano plot SFIvsELI")
```

### Volcano plot SFIvsELI

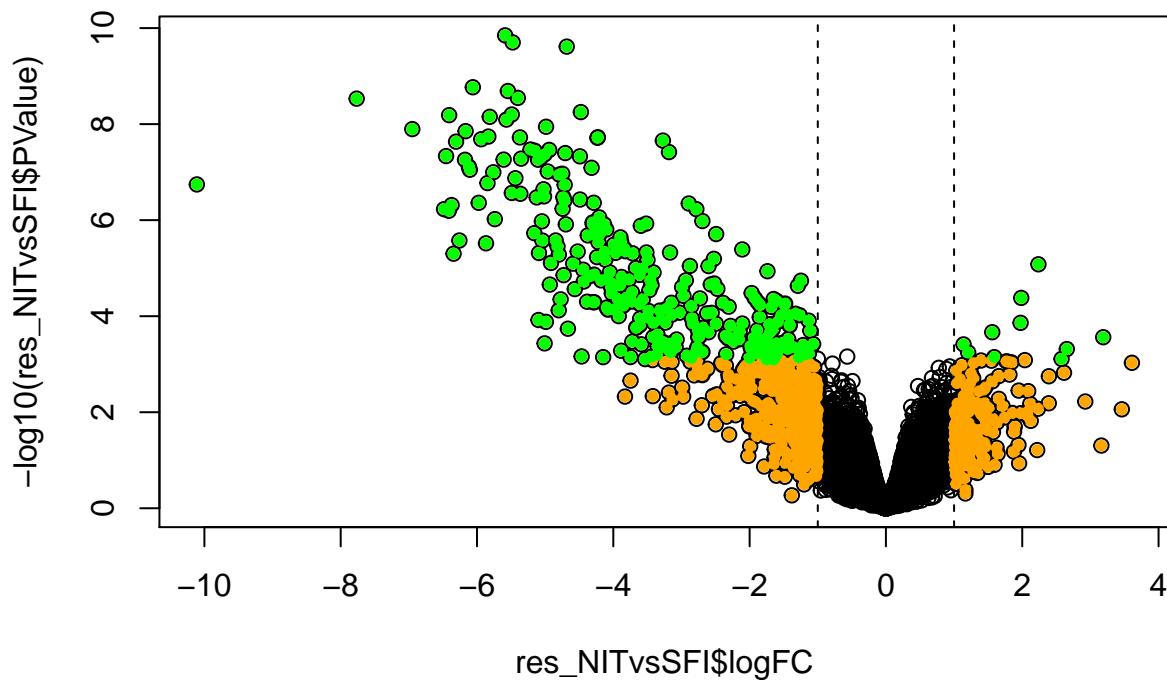


#### 4.3.2) NIT vs SFI

Al comparar las muestras NIT con las SFI obtenemos 295 genes que están downregulados y 10 genes que están upregulados en NIT con respecto a SFI. Es el contraste en el que obtenemos un menor número de genes diferencialmente expresados.

```
plot(res_NITvsSFI$logFC, -log10(res_NITvsSFI$PValue))
abline(v=c(-1,1), lty="dashed")
with(subset(res_NITvsSFI, abs(logFC)>1), points(logFC, -log10(PValue),
                                                 pch=20, col="orange"))
with(subset(res_NITvsSFI, FDR<.05 & abs(logFC)>1), points(logFC, -log10(PValue),
                                                 pch=20, col="green"))
title("Volcano plot NITvsSFI")
```

## Volcano plot NITvsSFI



### 4.3 Comparación entre los distintos contrastes

Para comprobar si los genes que cambian significativamente su nivel de expresión entre estas tres condiciones coinciden entre sí se decidió hacer un diagrama de Venn.

```

sig_NITvsELI2<- subset(res_NITvsELI, FDR<0.05 & abs(logFC)>1)
sig_NITvsELI_names <- rownames(sig_NITvsELI2)
sig_SFIvsELI2<- subset(res_SFIvsELI, FDR<0.05 & abs(logFC)>1)
sig_SFIvsELI_names <- rownames(sig_SFIvsELI2)
sig_NITvsSFI2<- subset(res_NITvsSFI, FDR<0.05 & abs(logFC)>1)
sig_NITvsSFI_names <- rownames(sig_NITvsSFI2)

common <- intersect(intersect(rownames(sig_NITvsSFI2), rownames(sig_NITvsELI2)), rownames(sig_SFIvsELI2))

comb <- unique(c(sig_NITvsELI_names,sig_NITvsSFI_names, sig_SFIvsELI_names))

# Comparing comb with the above two
NITvsELI <- comb %in% sig_NITvsELI_names
SFIvsELI <- comb %in% sig_SFIvsELI_names
NITvsSFI <- comb %in% sig_NITvsSFI_names

# Generating venn counts to plot venn diagram
table_venn <- cbind(NITvsELI, NITvsSFI, SFIvsELI)
counts_venn <- vennCounts(table_venn)
vennDiagram(counts_venn, cex = 1, names = c("NITvsELI", "NITvsSFI", "SFIvsELI"),

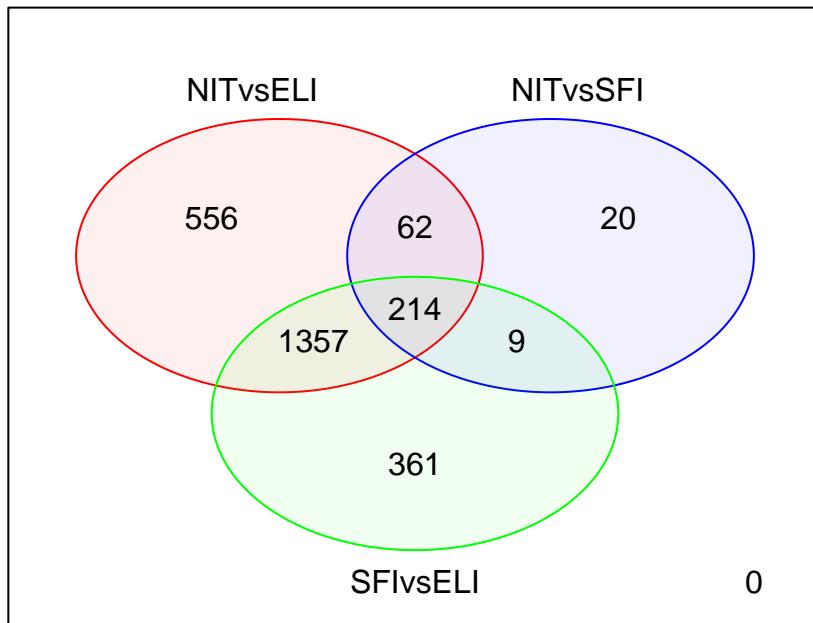
```

```

circle.col = c("red", "blue", "green"))
title("Genes en común entre todas las comparaciones
      con FDR < 0.05 and abs(log2FC) > 1")

```

## Genes en común entre todas las comparaciones con FDR < 0.05 and abs(log2FC) > 1



241 de los genes diferencialmente expresados son los mismos en todos los contrastes.

### 4.4) Análisis de significación biológica

Estamos interesados en saber a qué rutas metabólicas o en qué procesos biológicos están involucrados esos genes diferencialmente expresados. Como podemos observar, en todos los contrastes los términos GO más destacados están relacionados con la respuesta inmunológica: “respuesta inmune adaptativa”, “activación de linfocitos”, “proliferación de linfocitos”...

#### 4.4.1) NIT vs ELI

```

library(org.Hs.eg.db)
library(DOSE)
library(pathview)
library(clusterProfiler)
library(AnnotationHub)
library(ensemblDb)
library(tidyverse)

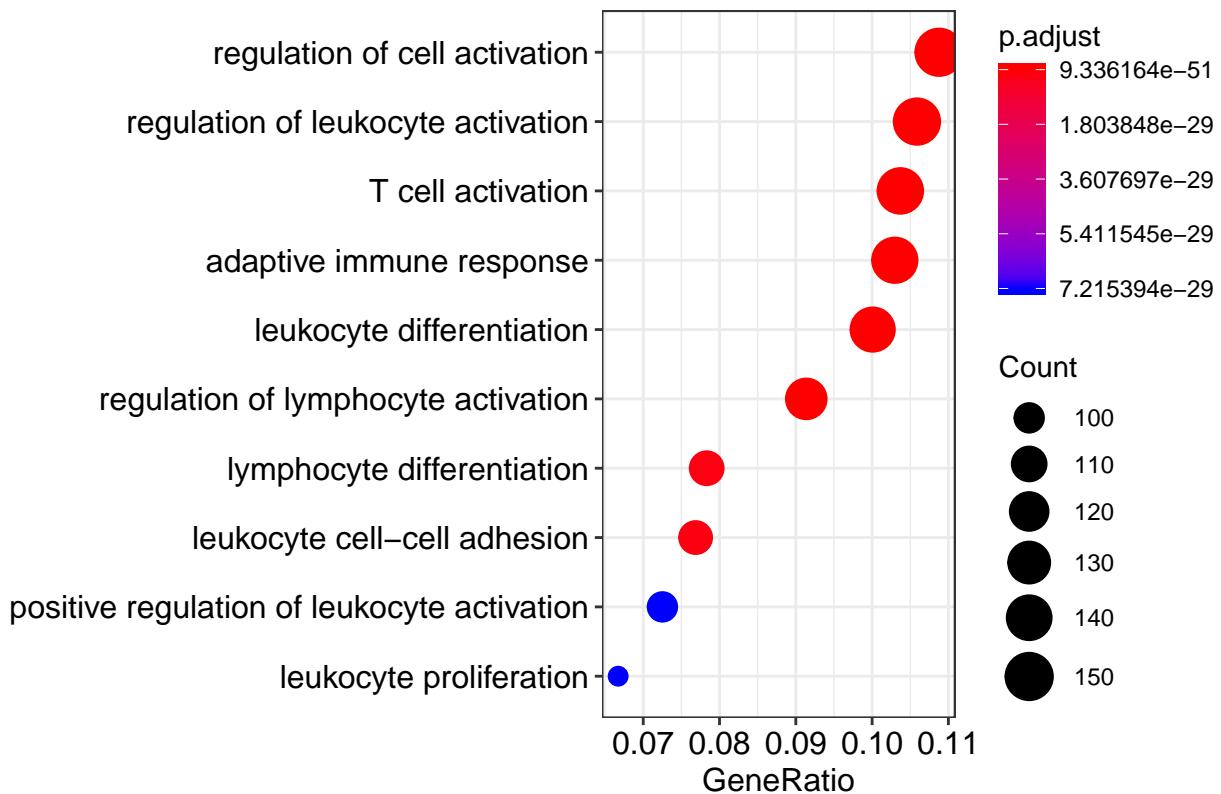
organism = "org.Hs.eg.db"

```

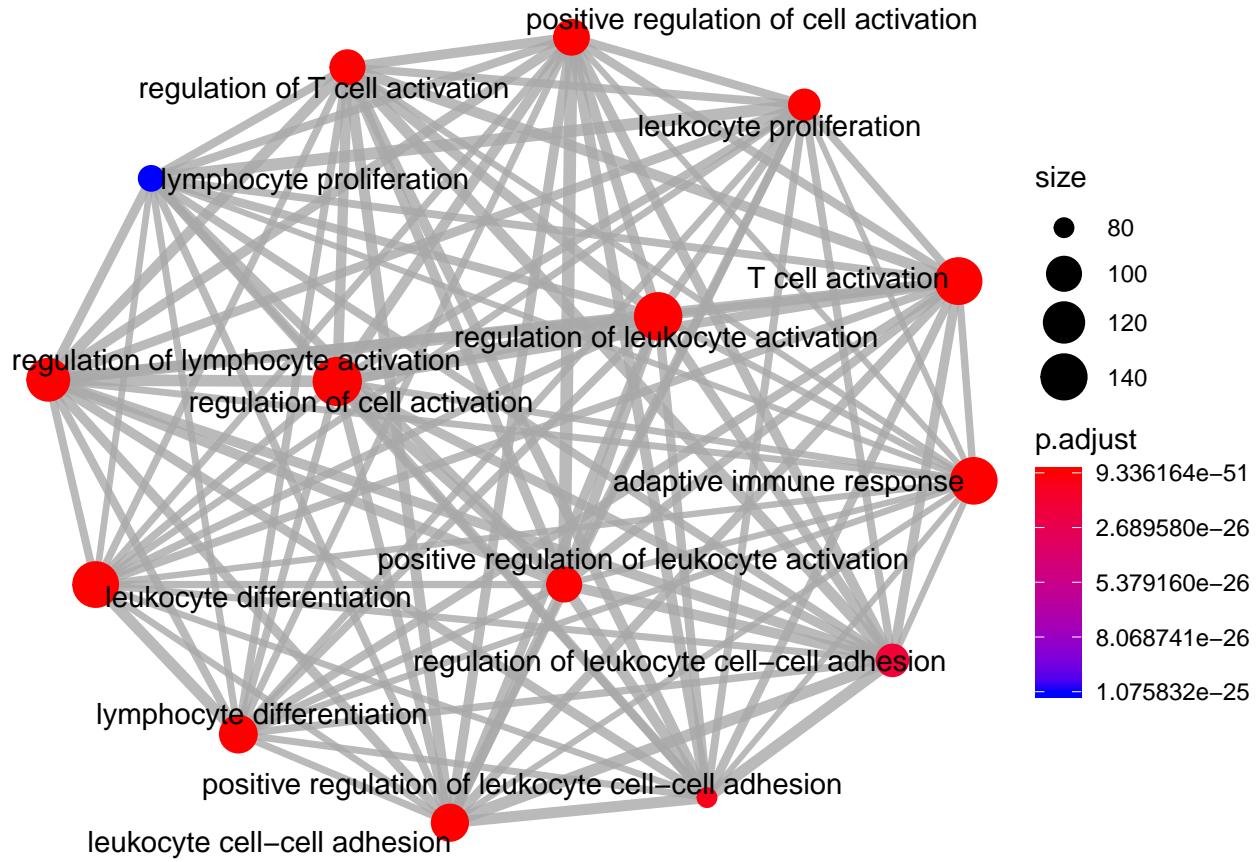
```

original_gene_list1 <- res_NITvsELI$logFC
names(original_gene_list1) <- gsub("\.\.*","", rownames(res_NITvsELI))
gene_list1<-na.omit(original_gene_list1)
gene_list1 <- sort(gene_list1, decreasing = TRUE)
sig_genes1 <- subset(res_NITvsELI, FDR < 0.05)
genes1 <- sig_genes1$logFC
names(genes1) <- gsub("\\..*","", rownames(sig_genes1))
genes1 <- na.omit(genes1)
genes1 <- names(genes1)[abs(genes1) > 1]
go_enrich1 <- enrichGO(gene = genes1,
                       universe = names(gene_list1),
                       OrgDb = organism,
                       keyType = 'ENSEMBL',
                       readable = T,
                       ont = "BP",
                       pvalueCutoff = 0.05,
                       qvalueCutoff = 0.10)
dotplot(go_enrich1)

```



```
emapplot(go_enrich1, showCategory = 15)
```



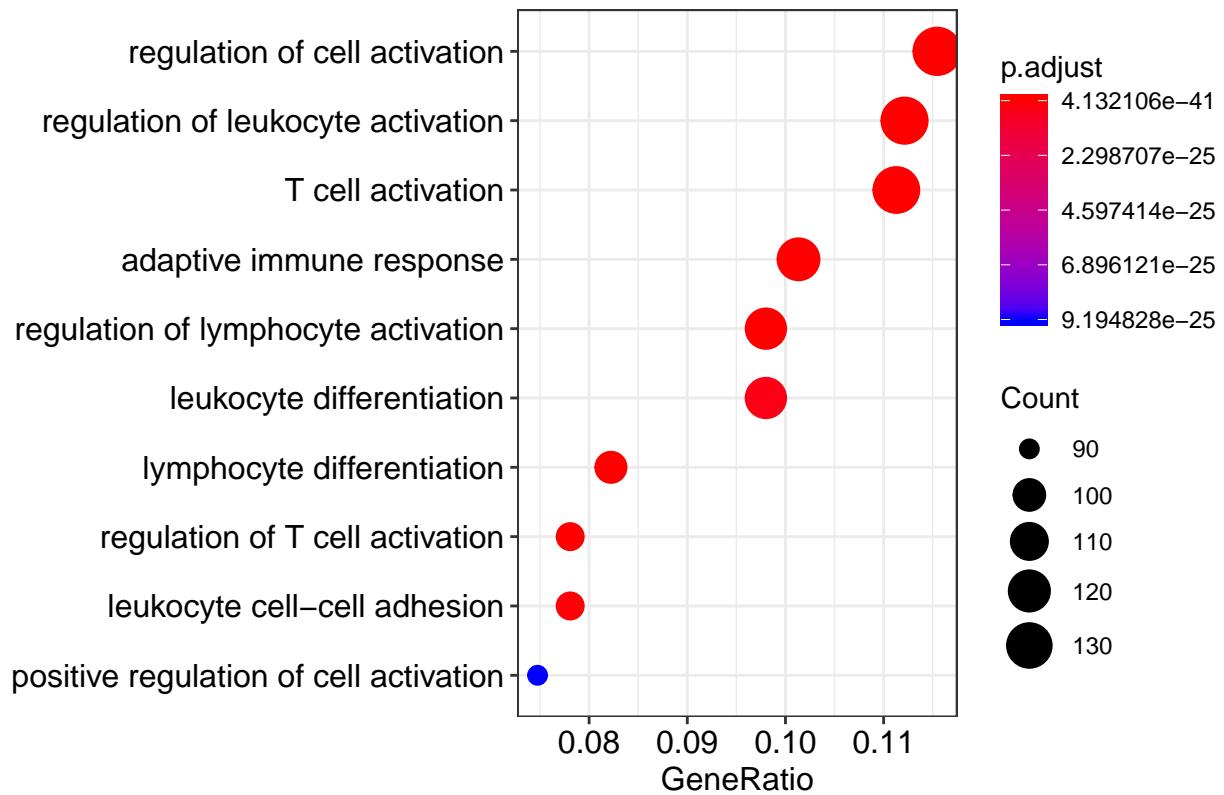
#### 4.4.2) SFI vs ELI

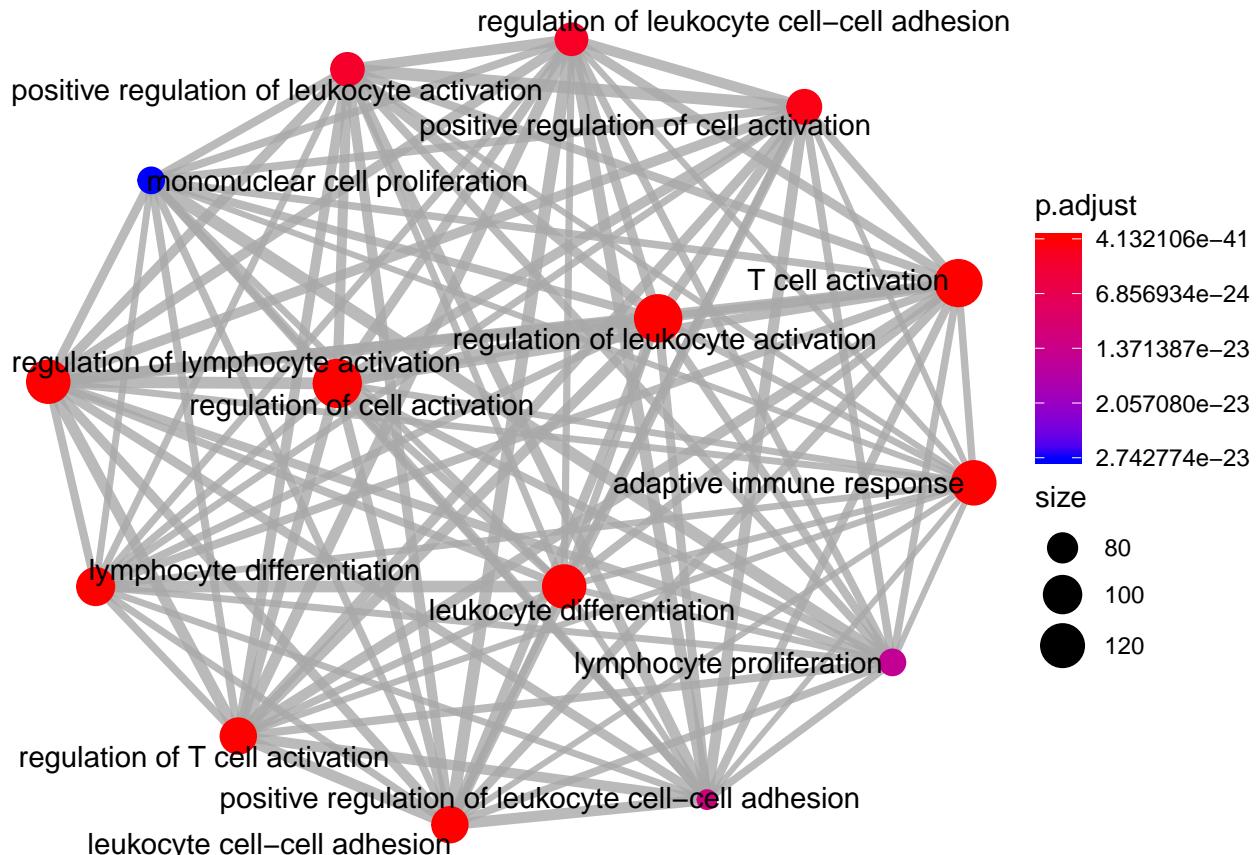
```

original_gene_list2 <- res_SFIvsELI$logFC
names(original_gene_list2) <- gsub("\\..*", "", rownames(res_SFIvsELI))
gene_list2<-na.omit(original_gene_list2)
gene_list2 <- sort(gene_list2, decreasing = TRUE)
sig_genes2 <- subset(res_SFIvsELI, FDR < 0.05)
genes2 <- sig_genes2$logFC
names(genes2) <- gsub("\\..*", "", rownames(sig_genes2))
genes2 <- na.omit(genes2)
genes2 <- names(genes2)[abs(genes2) > 1]
go_enrich2 <- enrichGO(gene = genes2,
                        universe = names(gene_list2),
                        OrgDb = organism,
                        keyType = 'ENSEMBL',
                        readable = T,
                        ont = "BP",
                        pvalueCutoff = 0.05,
                        qvalueCutoff = 0.10)
dotplot(go_enrich2)

## wrong orderBy parameter; set to default `orderBy = "x"`

```





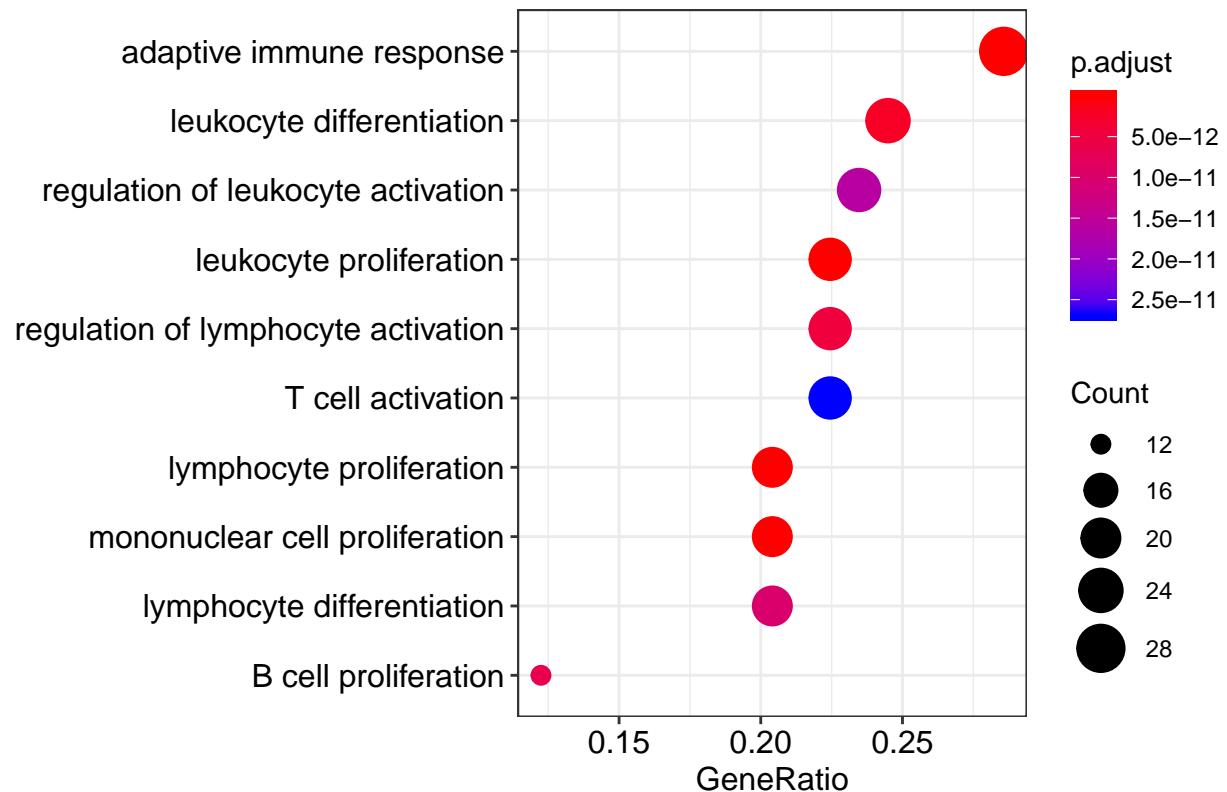
#### 4.4.3) NIT vs SFI

```

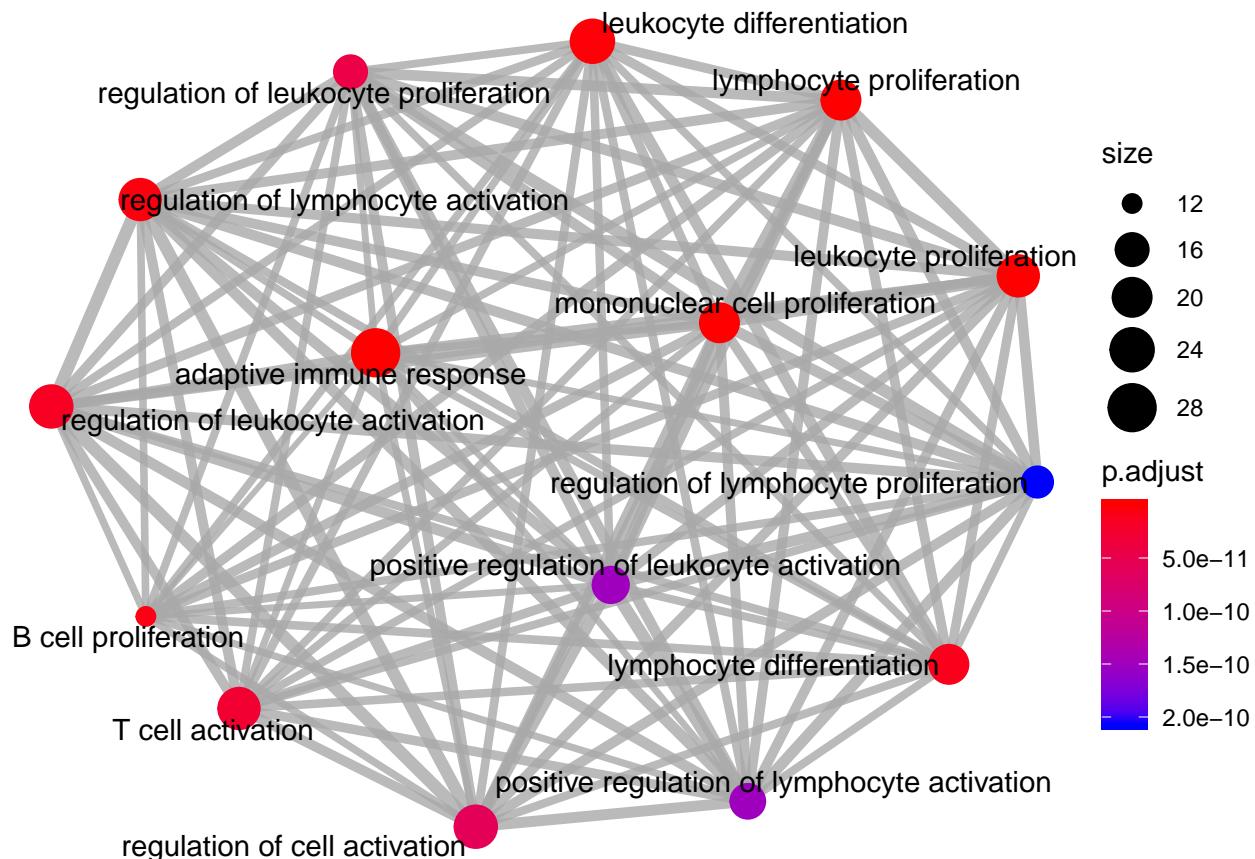
original_gene_list3 <- res_NITvsSFI$logFC
names(original_gene_list3) <- gsub("\\..*", "", rownames(res_NITvsSFI))
gene_list3<-na.omit(original_gene_list3)
gene_list3 <- sort(gene_list3, decreasing = TRUE)
sig_genes3 <- subset(res_NITvsSFI, FDR < 0.05)
genes3 <- sig_genes3$logFC
names(genes3) <- gsub("\\..*", "", rownames(sig_genes3))
genes3 <- na.omit(genes3)
genes3 <- names(genes3)[abs(genes3) > 1]
go_enrich3 <- enrichGO(gene = genes3,
                        universe = names(gene_list3),
                        OrgDb = organism,
                        keyType = 'ENSEMBL',
                        readable = T,
                        ont = "BP",
                        pvalueCutoff = 0.05,
                        qvalueCutoff = 0.10)
dotplot(go_enrich3)

## wrong orderBy parameter; set to default `orderBy = "x"`

```



```
emappplot(go_enrich3, showCategory = 15)
```



## 5) Discusión

Tras el análisis realizado en las muestras de tiroides con distinto grado de infiltración se han realizado las siguientes observaciones:

- Las muestras sin infiltraciones o con pequeños infiltrados locales son las más similares entre sí (menos genes diferencialmente expresados entre estos dos grupos). No obstante hay considerablemente más genes diferencialmente expresados entre estos dos grupos y las muestras con extensa infiltración.
- Los términos GO que aparecen destacados en cualquiera de los contrastes se refieren a procesos inmunes e incluyen conceptos como respuesta inmune o proliferación, diferenciación, regulación y activación de células inmunes como linfocitos.
- Estas observaciones se corresponden con lo que cabría esperar ya que al aumentar el grado de infiltración de la tiroides habría más células del sistema inmune en el tejido, y en esta población de células es donde se expresan a niveles muy altos este tipo de genes relacionados con la respuesta inmunitaria.
- Sería interesante repetir el mismo estudio utilizando técnicas de célula única (single-cell RNA sequencing) para poder determinar si los cambios que observamos entre los diferentes grupos se deben solo a un cambio en la composición celular del tejido (en ELI hay más linfocitos que en NIT) o si el perfil transcripcional del resto de las células de la glándula tiroidea también cambia entre los diferentes grupos (Hwang et al., 2018).
- Asimismo el análisis de expresión diferencial podría hacerse algo más complejo utilizando modelos lineales donde se incluyesen otras variables que potencialmente podrían estar también relacionadas con el nivel de infiltración, como la edad o el sexo (Marderstein et al, 2020).

## 6) Referencias

- GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet.* 2013;45(6):580-585. doi:10.1038/ng.2653.
- Hwang, B., Lee, J.H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* 50, 96 (2018). <https://doi.org/10.1038/s12276-018-0071-8>
- Marderstein, A.R., Uppal, M., Verma, A. et al. Demographic and genetic factors influence the abundance of infiltrating immune cells in human tissues. *Nat Commun* 11, 2213 (2020). <https://doi.org/10.1038/s41467-020-16097-9>.