

PEC2

Rashid Babiker Sánchez

7 de junio, 2020

Contents

Resumen	2
Objetivos	2
Materiales y Métodos	3
1. Muestreo y normalización por DESeq2 (código en “01 Muestreo y preprocesado.Rmd”):	3
2. Análisis descriptivo general y control de calidad (código en “02 análisis descriptivo.Rmd”) . . .	3
3. Análisis de expresión diferencial y de significación biológica (códigos en “03 ELI vs NIT.Rmd”, “04 SFI vs NIT.Rmd” y “05 ELI vs SFI.Rmd”):	3
Resultados y discusión	4
ELI vs NIT	6
ELI vs SFI	9
SFI vs NIT	12
Bibliografía	15

Resumen

Se han analizado 30 muestras de tiroides con distintos grados de infiltración: sin infiltración (NIT), con infiltración pequeña (SFI) y con gran infiltración (ELI), se ha comparado la expresión diferencial de estas tres condiciones por parejas, los resultados muestran una sobreexpresión de genes involucrados en la activación del sistema inmune en los tejidos con mayor grado de infiltración, lo que sugiere una posible infección en el paciente o una desregulación génica en estos tejidos.

El repositorio con todos los análisis realizados se puede consultar en el siguiente enlace <https://github.com/RashBabiker/PEC2.git>

Objetivos

En este estudio se persiguen 2 objetivos:

- Analizar la expresión génica de los tejidos de tiroides con distintos grados de infiltración.
- Localizar genes, rutas metabólicas y procesos diferencialmente expresados

El enunciado de la PEC indica que no se espera de este informe una interpretación biológica de los resultados.

Materiales y Métodos

A continuación, se explican los pasos seguidos para procesar y analizar las muestras proporcionadas, también se indica el nombre de los archivos de Rmarkdown del repositorio donde se puede acceder el código usado para la realización de cada tarea con una explicación mucho más detallada:

1. Muestreo y normalización por DESeq2 (código en “01 Muestreo y preprocesado.Rmd”):

Selección aleatoria de 10 muestras de cada grupo (ELI, SFI y NIT). Estas muestras se normalizan con DESeq, los datos resultantes se usan para el análisis de expresión diferencial. Para análisis descriptivos como heatmaps y PCAs se recomienda que los datos sean homocedásticos, por ello se transforman los datos normalizados con DESeq rlog.

2. Análisis descriptivo general y control de calidad (código en “02 análisis descriptivo.Rmd”)

Breve descripción visual de las muestras, para cuantificar la similitud entre replicas replicas y las diferencias entre grupos.

3. Análisis de expresión diferencial y de significación biológica (códigos en “03 ELI vs NIT.Rmd”, “04 SFI vs NIT.Rmd” y “05 ELI vs SFI.Rmd”):

A partir de los datos normalizados con DESeq2 se hacen comparaciones por parejas de los tejidos con distintos niveles de infiltración, en estas comparaciones se pone en el numerador las muestras con mayor grado de infiltración (ELI>SFI>NIT), de forma que los genes sobreexpresados serán aquellos que se expresen más en la muestra con mayor infiltración, en esta comparación.

3.1 Preparación de los datos

Según Conesa y colaboradores [1] para 10 réplicas por grupos y con un nivel de significación del 0.05, el fold change de 1,25 tiene una potencia (probabilidad de detectar expresión diferencial) del 44%, mientras que con 1.5 FC la potencia sube al 91%, por lo que usamos FC=1.5. El coeficiente de correlación elegido es 0.05 porque es el valor estándar y no se aprecia motivo para cambiarlo.

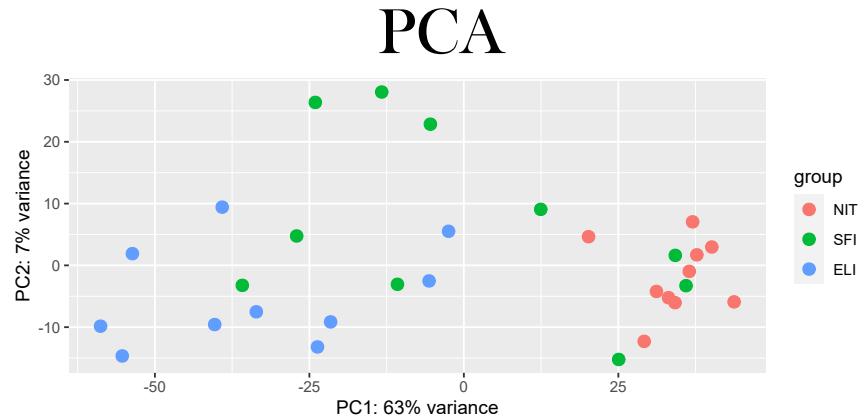
Encoger (Shrink) el fold change de los genes con pocas cuentas reduce el ruido, es util en rankings y visualización, se probaron distintos métodos para encoger, siendo el más eficaz el método ashr [2]

3.2. Anotación y Análisis de sobrerrepresentación (ORA)

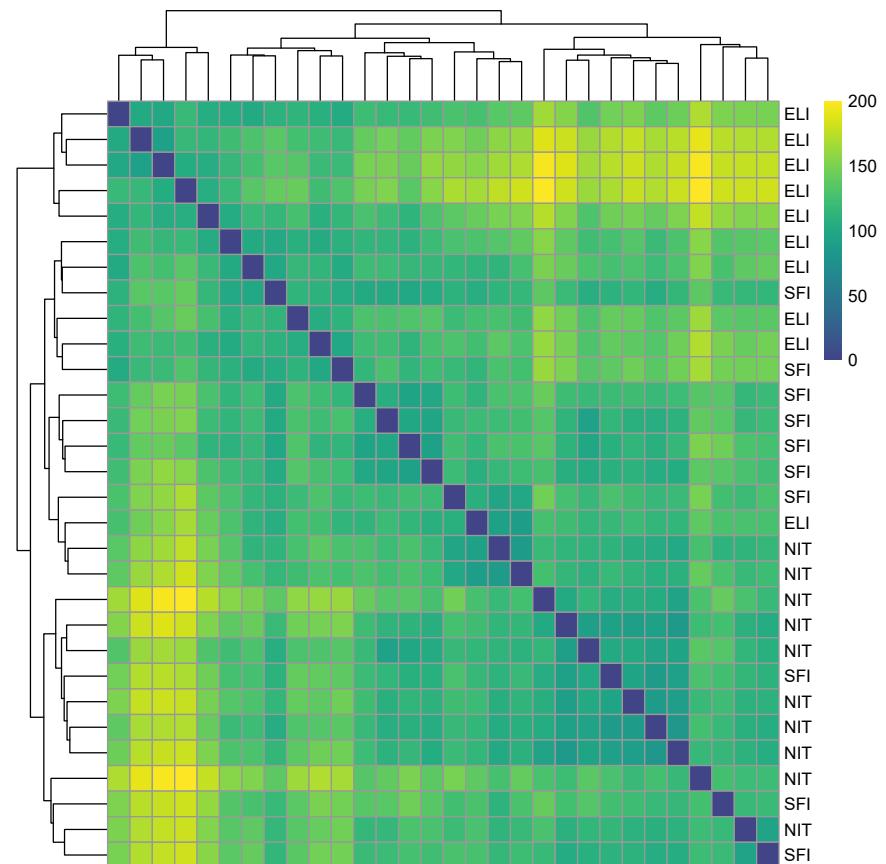
A partir del código de ENSEMBL de los genes diferencialmente expresados en las distintas comparaciones, se mapea información de otras bases de datos, de especial interés son EntrezID, GO y KEGG. El análisis de sobrerrepresentación consiste en un resumen de las rutas metabólicas de KEGG y de las categorías de GO más representadas de un conjunto de grupos. En este análisis se ha elegido mostrar solo los genes sobreexpresados en la condición con mayor filtración porque resultan de mayor interés, los procesos sobrerrepresentados en los tejidos con menor infiltración se pueden observar en la carpeta resultados del repositorio de github.

Resultados y discusión

La matriz de distancias y el PCA muestran claras diferencias en los patrones de expresión entre ELI (infiltraciones extensivas) y NIT (sin infiltraciones), pero los SFI (infiltraciones pequeñas) aparecen cerca de ambos como un punto intermedio, tiene sentido porque fenotípicamente SFI está entre ELI y NIT.

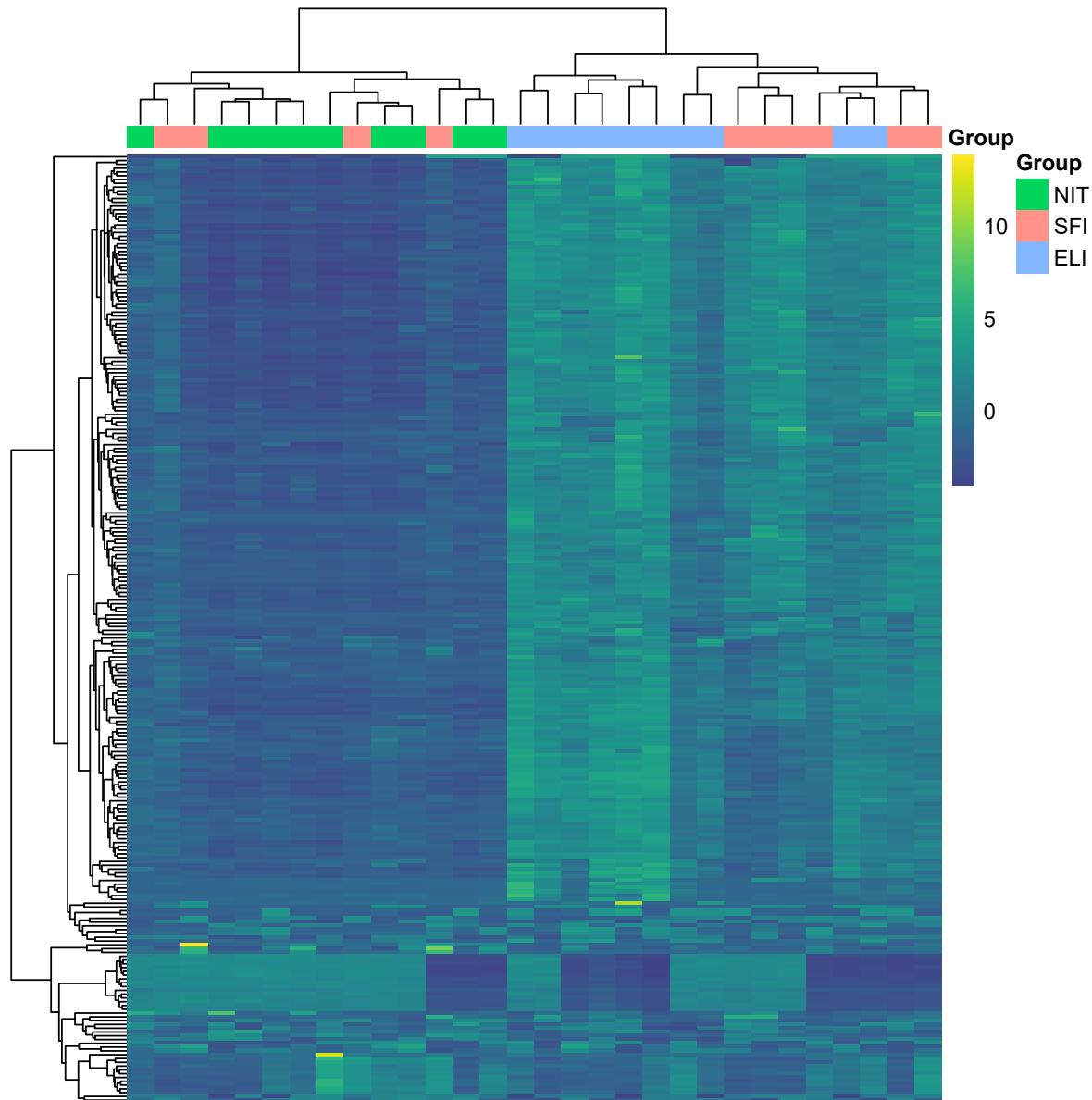


Matriz de distancias



El siguiente heatmap de los genes más variables muestra una diferenciación clara entre NIT y ELI en el clustering como en la expresión de estos genes. Si las infiltraciones no son evidentes a nivel fenotípico, el análisis de expresión de estos genes podría usarse como método de diagnóstico, o como método complementario para distinguir entre tipos de infiltración.

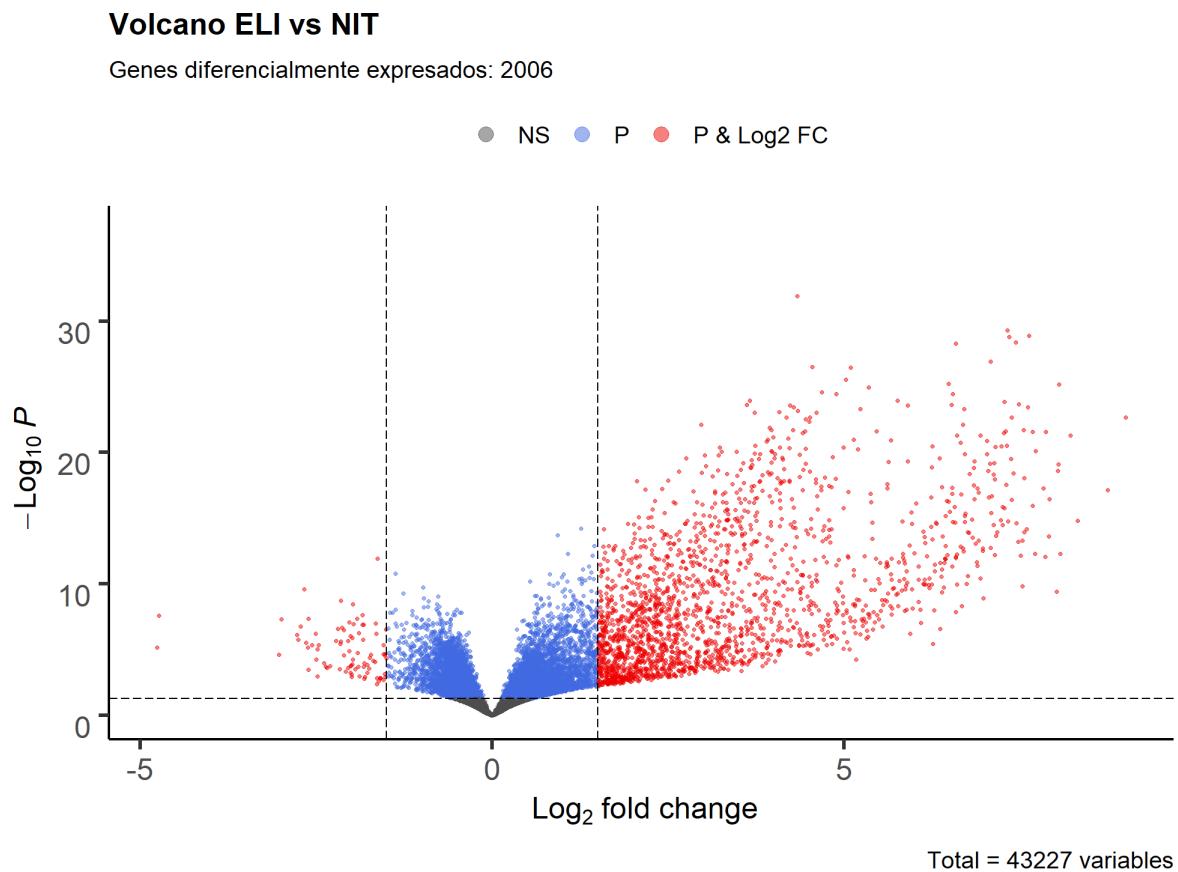
También se ven diferencia entre el grupo central de ELI, donde estos genes se expresan más que tejidos SFI y otros ELI, lo que sugiere que quizás se pudiera hacer una nueva clasificación, seguramente en ese grupo central de ELI la infiltración es mayor aún.



A continuación se pasa a analizar la expresión diferencial por parejas.

ELI vs NIT

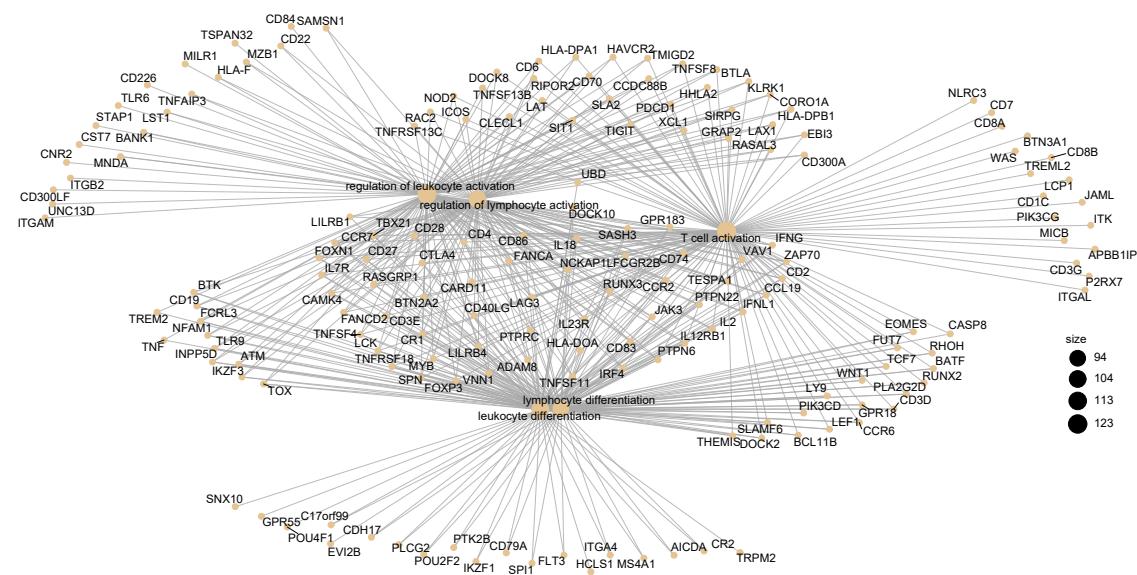
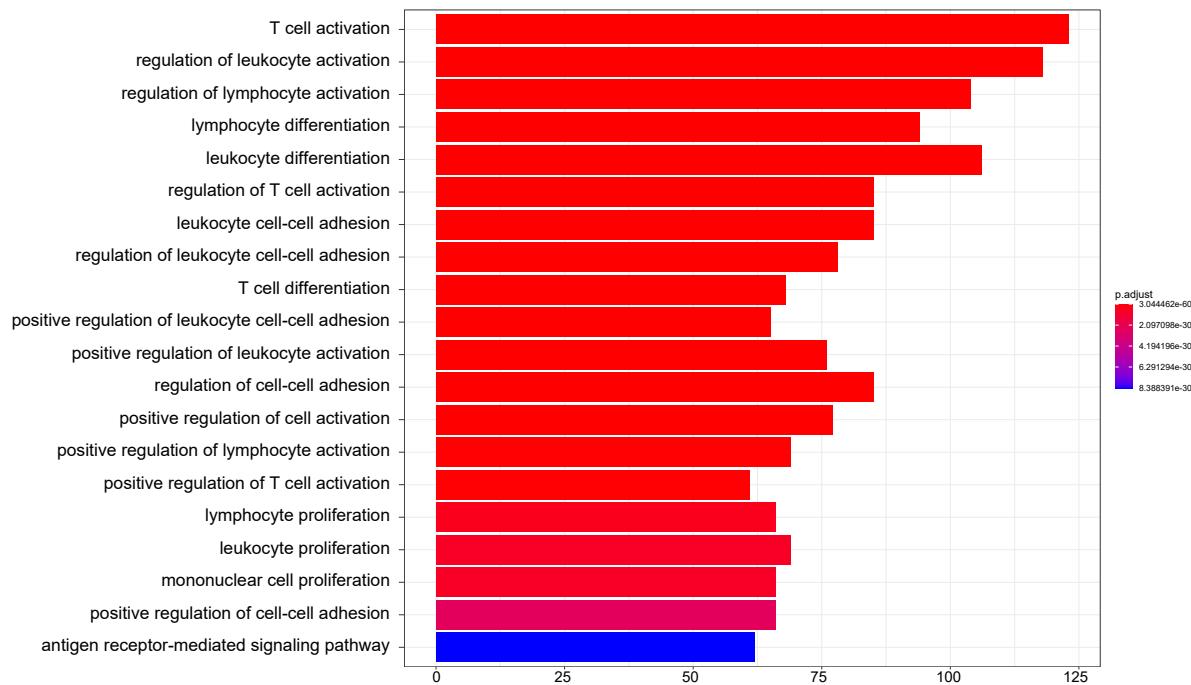
Como era de esperar, en esta comparación es donde más genes diferencialmente expresados se observan (2006, los marcados en rojo en el volcano plot). La mayoría de los genes sobreexpresados en los tejidos con mayor infiltración están relacionados con activación del sistema inmunológico, lo que sugiere una posible infección o una desregulación genética.



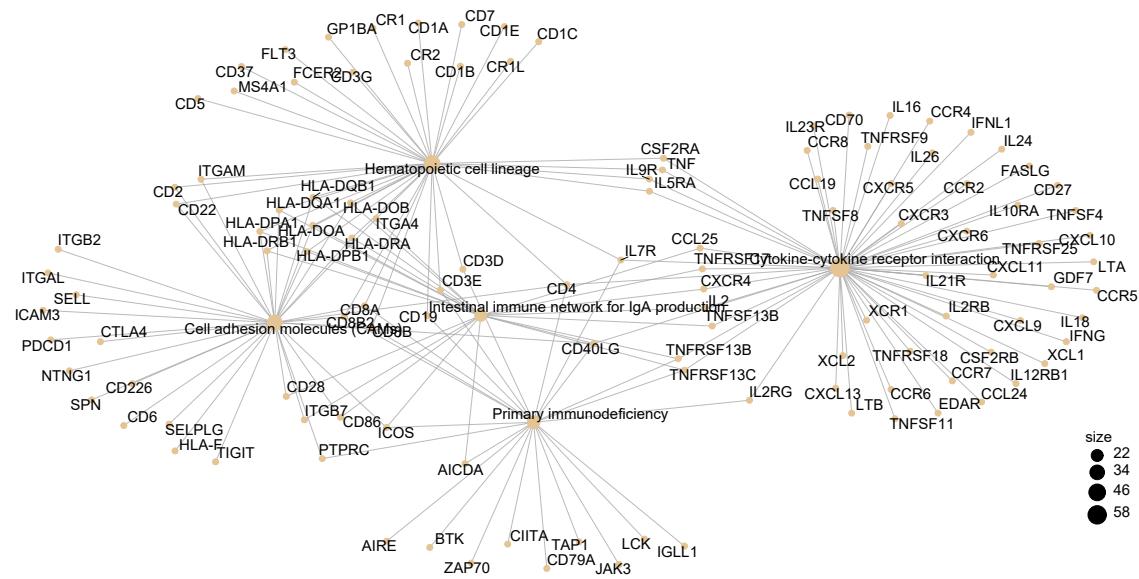
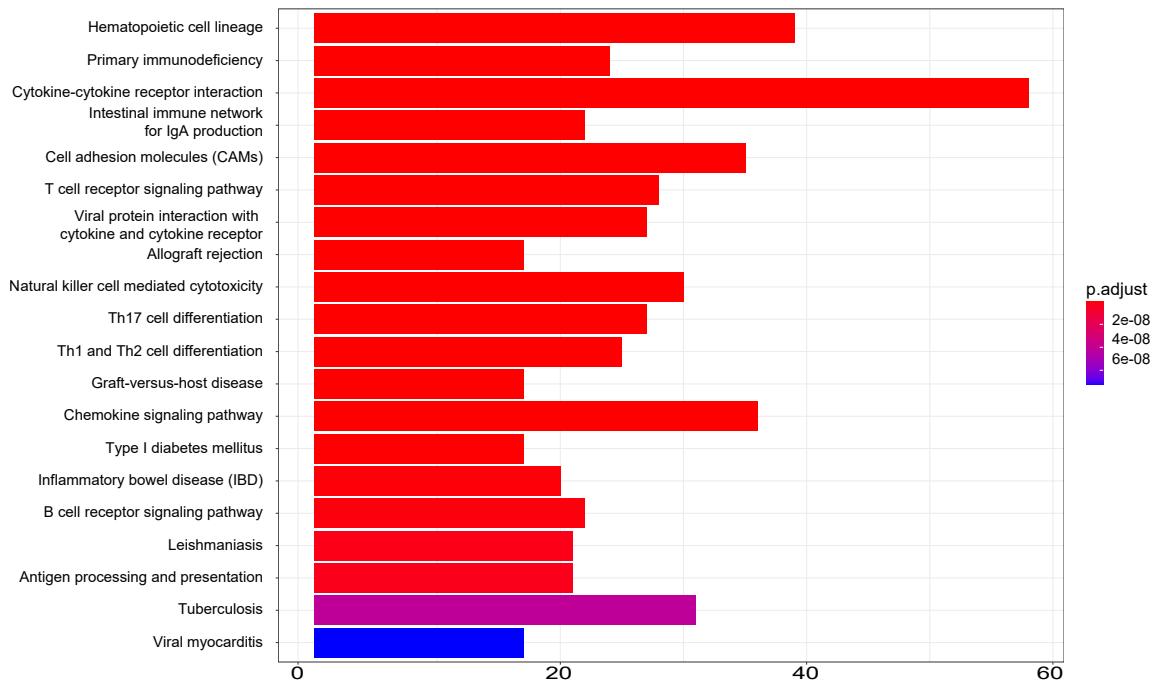
Genes más diferencialmente expresados, el archivo completo está en la carpeta resultados de github:

X	log2FoldChange	SYMBOL	ENTREZID	GO	PATH
398	ENSG00000009790.10	4.336301	TRAF3IP3	80342	GO:0005515
38	ENSG00000177455.7	7.321168	CD19	930	GO:0001923 4640
19	ENSG00000226777.3	7.626694	FAM30A	9834	NA
36	ENSG00000117322.12	7.339201	CR2	1380	GO:0001618 4610
31	ENSG00000211668.2	7.439096	NA		NA
104	ENSG00000167483.13	6.583420	NIBAN3	199786	NA
60	ENSG00000143297.14	7.076511	FCRL5	83416	GO:0005886
344	ENSG00000104894.7	4.544488	CD37	951	GO:0001772 4640
262	ENSG00000245164.2	5.093267	LINC00861	100130231	NA
272	ENSG00000089012.10	5.028000	SIRPG	55423	GO:0005515 4380
117	ENSG00000110777.7	6.480831	POU2AF1	5450	GO:0000978
6	ENSG00000211593.2	8.054668	NA		NA
233	ENSG00000026751.12	5.347433	SLAMF7	57823	GO:0002250
323	ENSG00000159753.9	4.682581	CARMIL2	146206	GO:0001726

Términos GO sobreexpresados (ELI vs NIT)

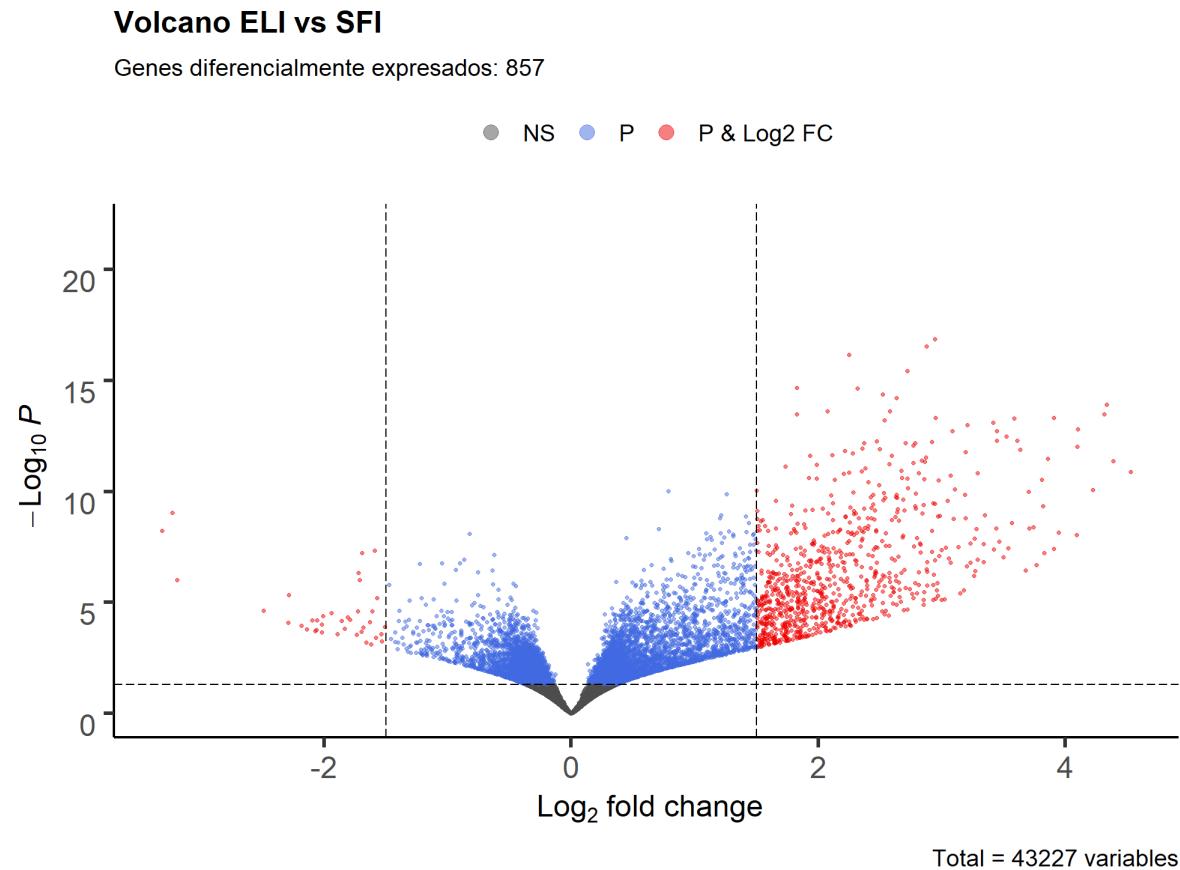


Rutas KEGG sobrexpresadas (ELI vs NIT)



ELI vs SFI

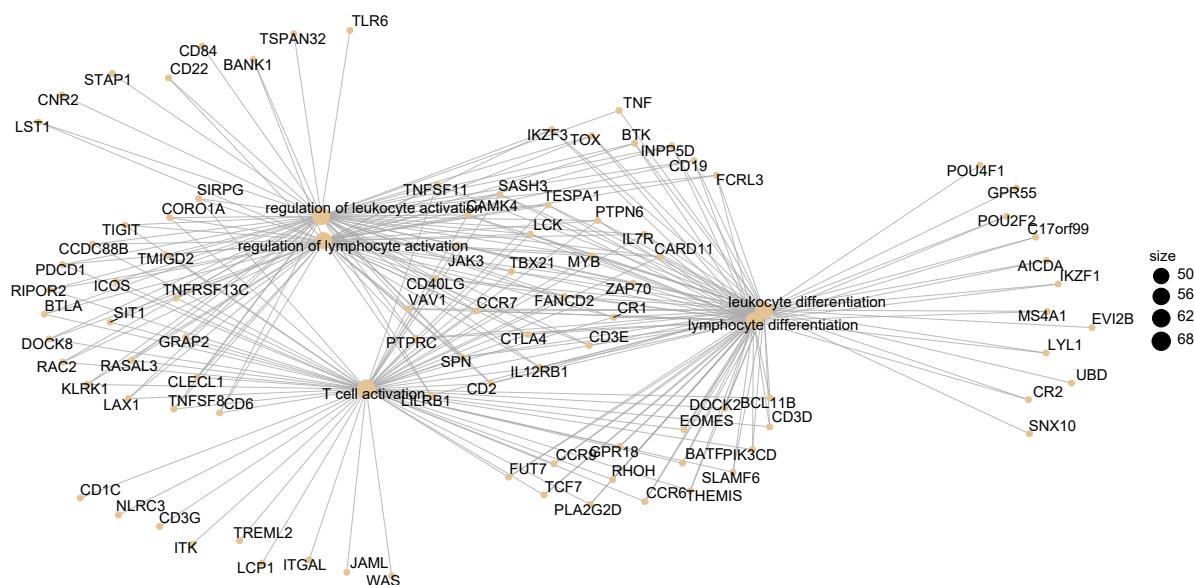
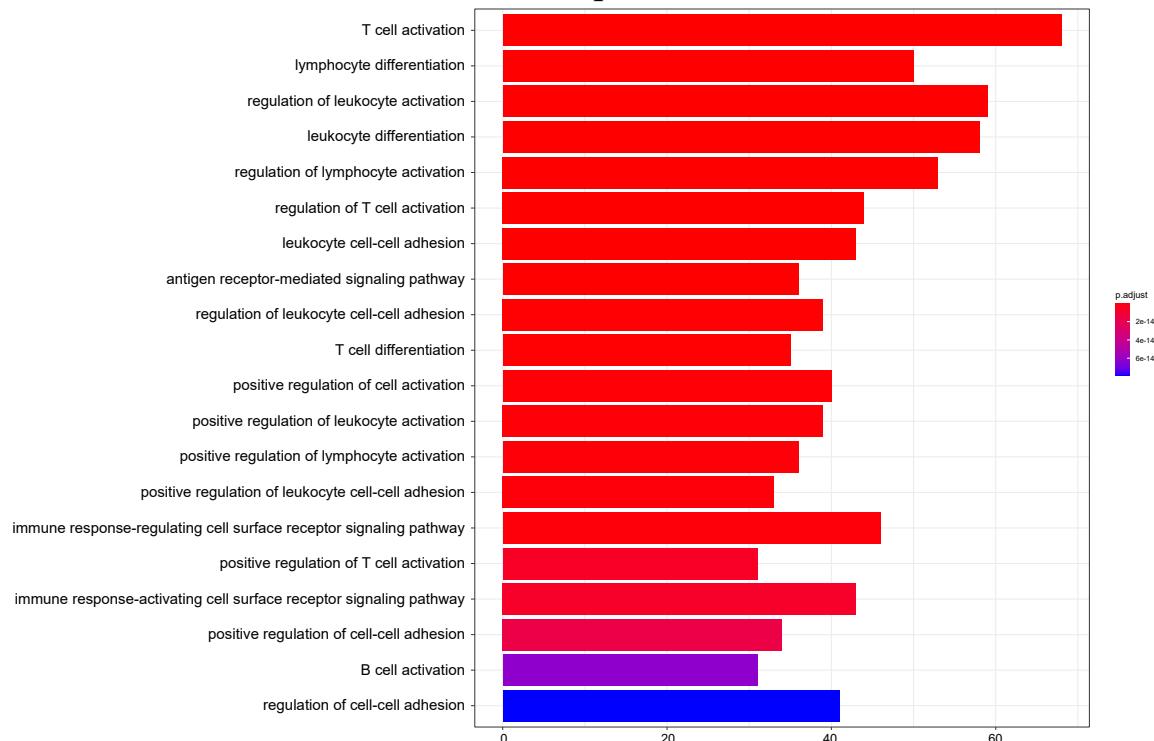
Hay una menor cantidad de genes expresados diferencialmente entre las muestras ELI y SFI (857), pero esta diferencia es bastante superior a la observada al comparar SFI y NIT, lo que indica que las muestras SFI son más similares a las NIT que a las ELI, como también se podía observar en el PCA. Las rutas con mayor sobreexpresión están relacionadas con el sistema inmunológico, al igual que en la comparación anterior.



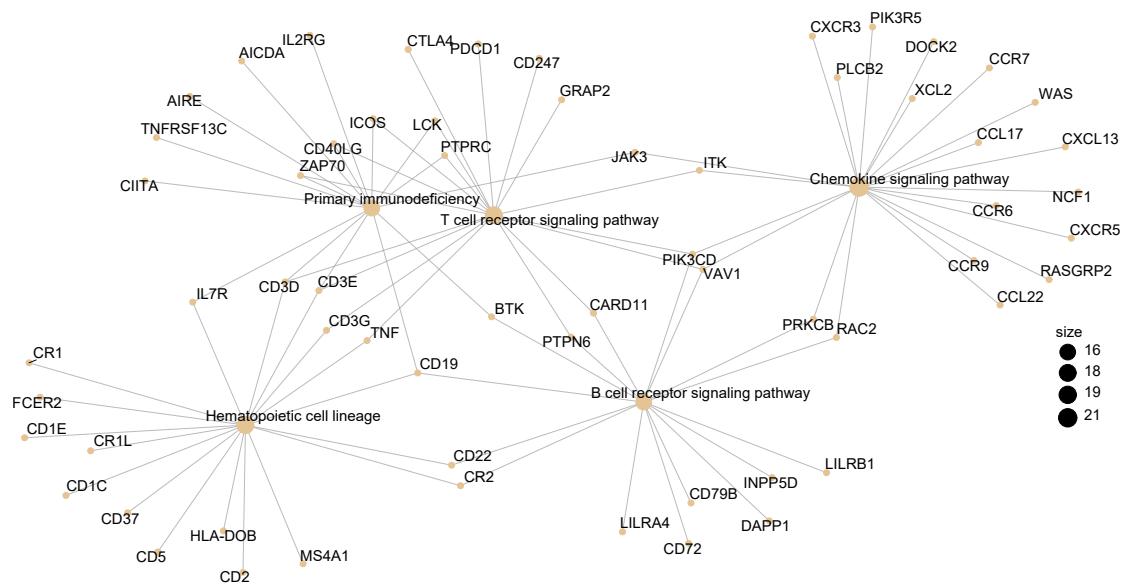
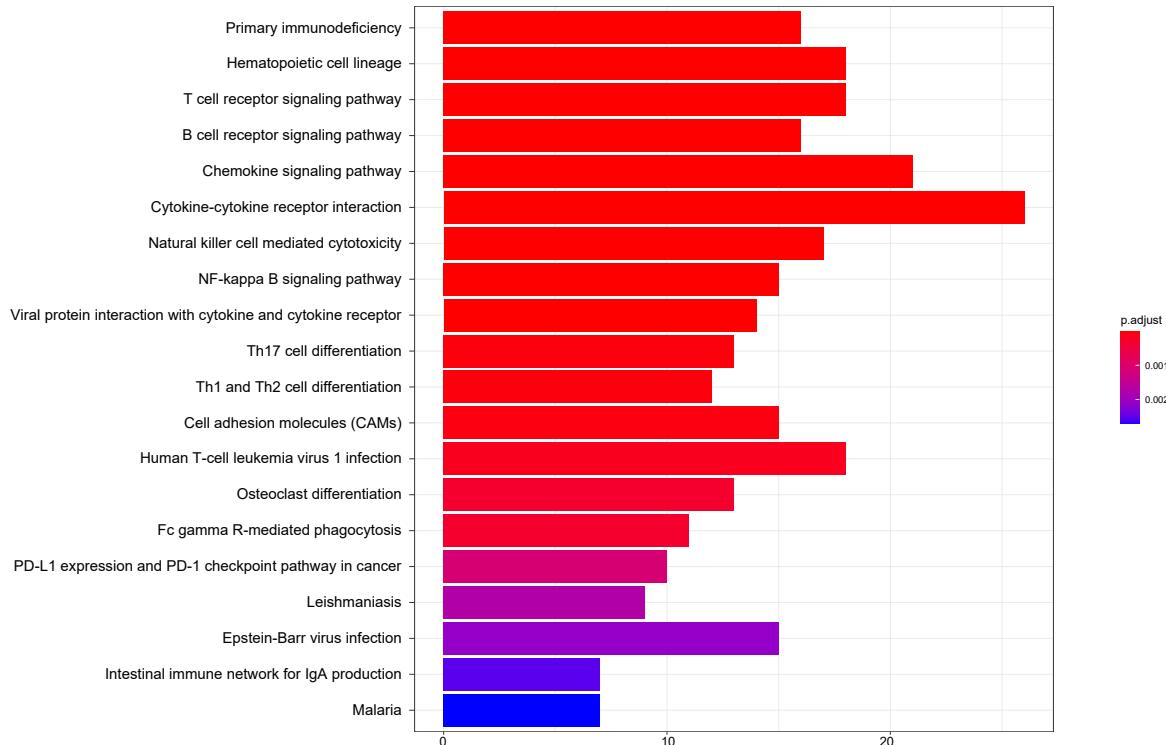
Genes más diferencialmente expresados:

X	log2FoldChange	SYMBOL	ENTREZID	GO	PATH
73 ENSG00000009790.10	2.945738	TRAF3IP3	80342	GO:0005515	
86 ENSG00000247982.2	2.877700	LINC00926	283663	NA	
282 ENSG00000175463.7	2.250444	TBC1D10C	374403	GO:0005515	
118 ENSG00000068831.14	2.720206	RASGRP2	10235	GO:0001558	4010
250 ENSG00000205744.5	2.317099	DENND1C	79958	GO:0005829	
536 ENSG00000177721.3	1.824963	ANXA2R	389289	GO:0005515	
173 ENSG00000111913.11	2.522965	RIPOR2	9750	GO:0005515	
141 ENSG00000204282.3	2.634225	TNRC6C-AS1	100131096	NA	
3 ENSG00000117322.12	4.333361	CR2	1380	GO:0001618	4610
158 ENSG00000072818.7	2.577855	ACAP1	9744	GO:0005096	4144
370 ENSG00000122986.9	2.075351	HVCN1	84329	GO:0005886	
4 ENSG00000104921.10	4.316946	FCER2	2208	GO:0002925	4640
538 ENSG00000111679.12	1.824369	PTPN6	5777	GO:0001784	4520
10 ENSG00000167483.13	3.908918	NIBAN3	199786	NA	

Términos GO sobreexpresados (ELI vs SFI)

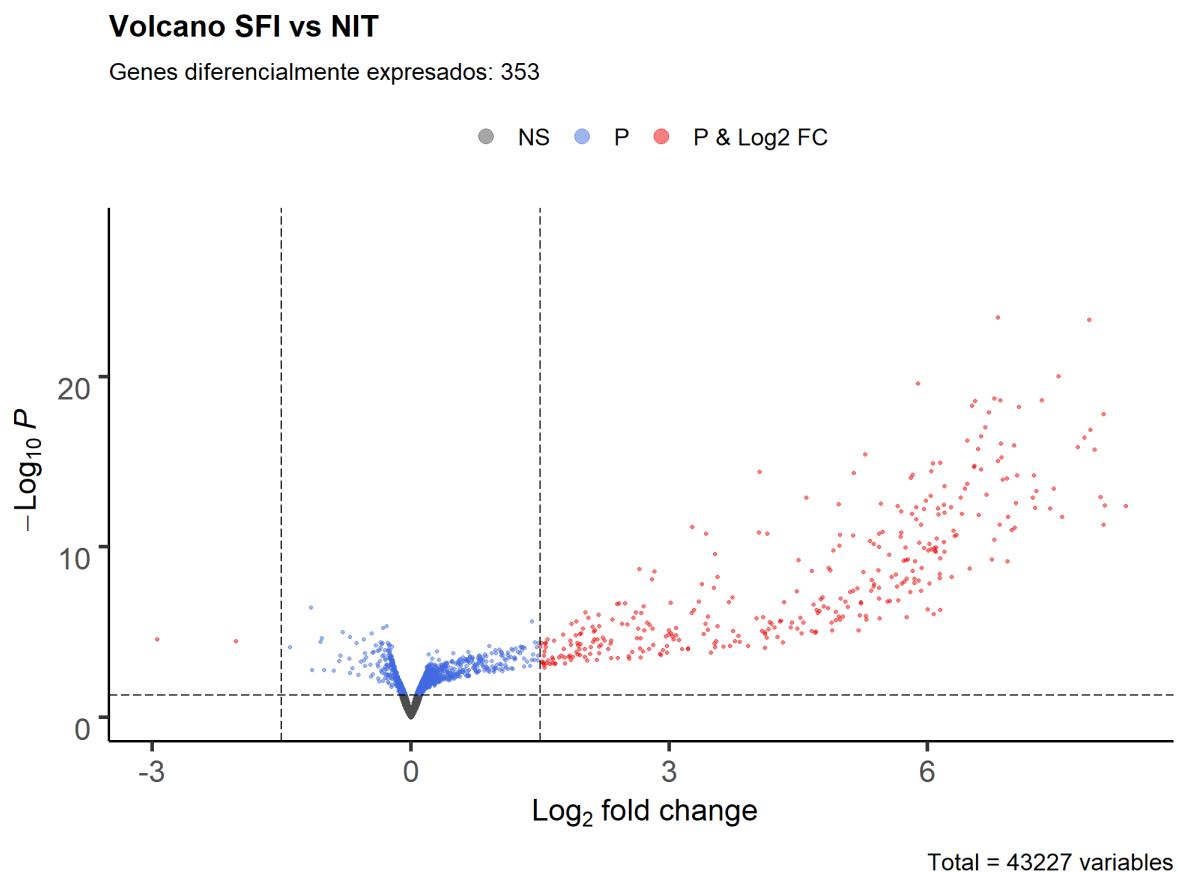


Rutas KEGG sobreexpresadas (ELI vs SFI)



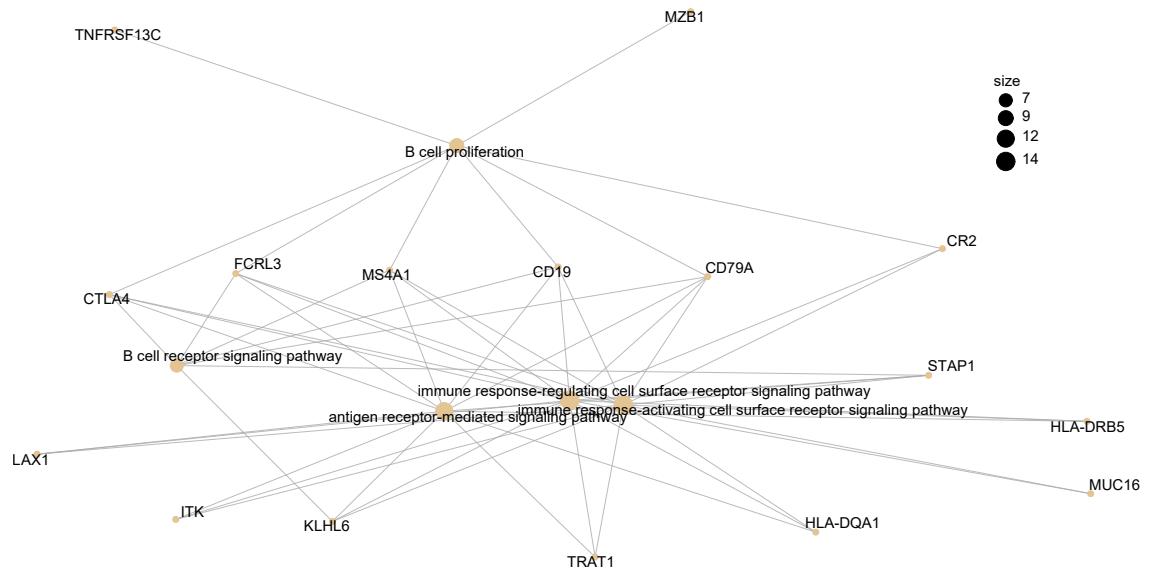
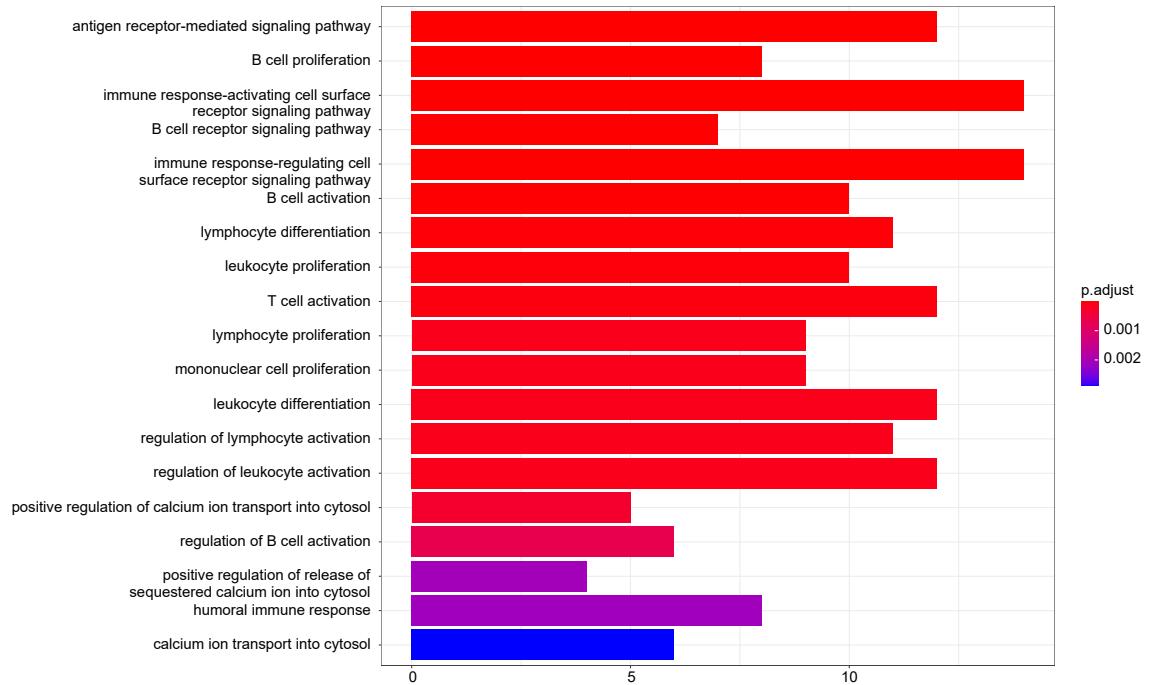
SFI vs NIT

Estas son las muestras más similares entre sí, con expresión diferencial solo en 353 genes, al igual que en las otras comparaciones, las rutas metabólicas más afectadas por la infiltración están relacionadas con la activación del sistema inmunológico.

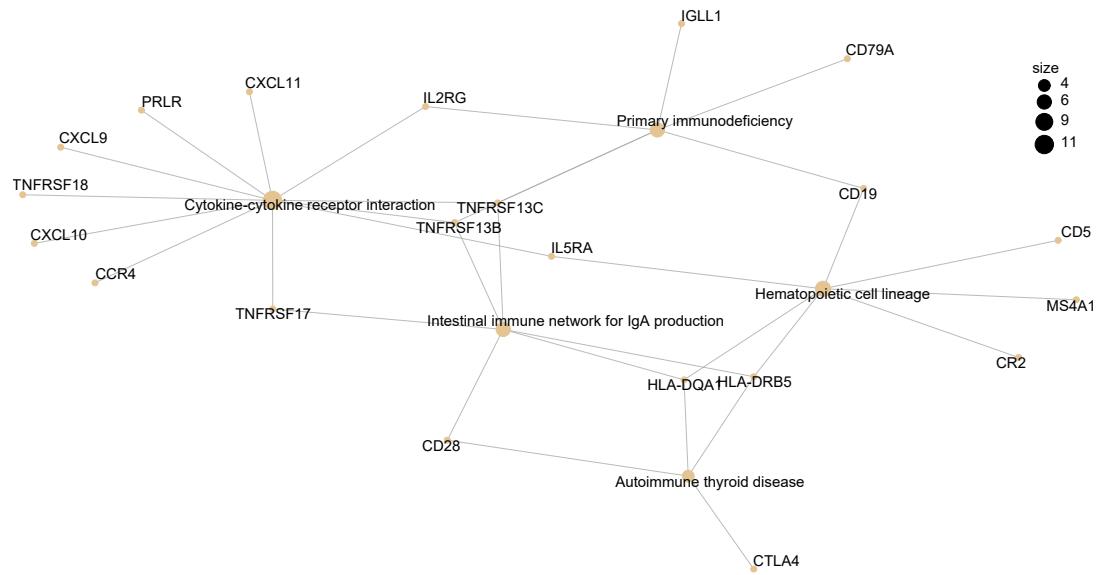
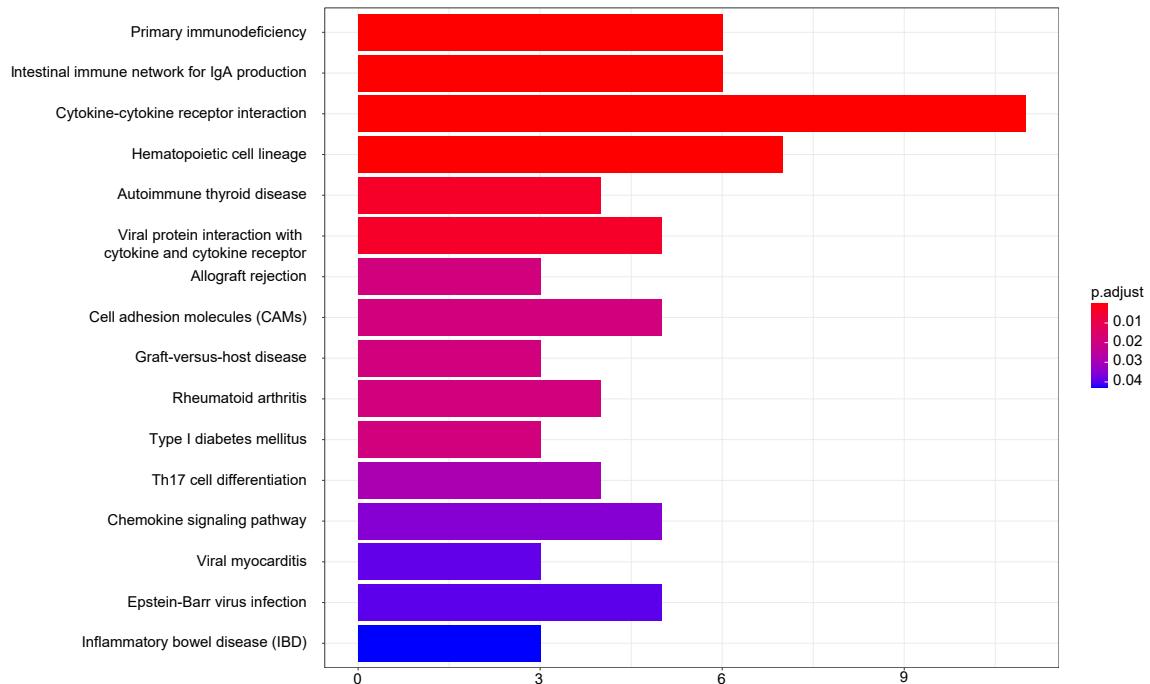


X	log2FoldChange	SYMBOL	ENTREZID	GO	PATH
8 ENSG00000211593.2	7.876016	NA		NA	
34 ENSG00000211668.2	6.817128	NA		NA	
12 ENSG00000240864.1	7.518020	NA		NA	
95 ENSG00000132465.6	5.890587	JCHAIN	3512	GO:0001895	
15 ENSG00000211965.2	7.326375	NA		NA	
32 ENSG00000211660.3	6.841566	NA		NA	
37 ENSG00000211594.2	6.773782	NA		NA	
46 ENSG00000211644.2	6.547307	NA		NA	
49 ENSG00000211896.2	6.517727	NA		NA	
20 ENSG00000211947.2	7.057434	NA		NA	
39 ENSG00000211666.2	6.711446	NA		NA	
3 ENSG00000242371.1	8.042711	NA		NA	
41 ENSG00000211945.2	6.667587	NA		NA	
7 ENSG00000222037.5	7.889544	NA		NA	

Términos GO sobreeexpresados (SFI vs NIT)



Rutas KEGG sobreexpresadas (SFI vs NIT)



Bibliografía

- [1] A. Conesa *et al.*, “A survey of best practices for RNA-seq data analysis,” *Genome Biology*, vol. 17, no. 1, pp. 1–19, 2016.
- [2] M. Stephens, “False discovery rates: a new deal,” *Biostatistics*, vol. 18, no. 2, pp. 275–294, Oct. 2016.