

Análisis de datos de expresión (RNA-seq) en muestras de infiltración de tiroides

Análisis Datos Ómicos - PEC 2

Jorge Vallejo Ortega

9 de September, 2020

Índice

1 Abstract	2
2 Objetivos	2
3 Materiales y métodos	2
3.1 Datos de origen	2
3.2 Diseño experimental	2
3.3 Procedimiento seguido en el análisis	3
4 Identificación y anotación de genes diferencialmente expresados	12
4.1 Anotación de las listas de genes	13
4.2 Cuantificación de genes diferencialmente expresados	15
5 Comparaciones múltiples	17
5.1 Diagrama de Venn	18
6 Significatividad biológica	18
6.1 Test de sobrerepresentación de términos GO	18
6.2 Visualización de resultados: Gráficos de barras	20
7 Discusión	21
8 Apéndice A: Código	21
9 Apéndice B: Reproducibilidad	21
Referencias	23

1 Abstract

Análisis de expresión génica diferencial de un estudio obtenido del proyecto [Genotype-Tissue Expression \(GTEx\)](#). Partimos de datos de expresión obtenidos por **RNA-seq**, pertenecientes a un análisis de muestras tiroides pertenecientes a tres grupos: NIT (*not infiltrated tissues*, tejidos no infiltrados), SFI (*small focal infiltrates*, infiltrados focales pequeños), y ELI (*extensive lymphoid infiltrates*, infiltrados linfoides amplios). Como resultado hemos obtenido **listados de genes** con expresión diferencial entre los diferentes grupos, y **listados de términos GO** sobrerepresentados en dichos listados.

2 Objetivos

Los objetivos de este análisis son

- i) averiguar los cambios de expresión génica en la tiroides entre situación de salud (NIT) y situación patológica (SFI y ELI),
- ii) explorar los cambios en la expresión génica entre tejidos de tiroides en situación patológica más severa (ELI) y menos severa (SFI), y finalmente
- iii) examinar las rutas moleculares implicadas.

3 Materiales y métodos

3.1 Datos de origen

Originalmente, los datos brutos fueron obtenidos por RNA-seq y proceden del proyecto [Genotype-Tissue Expression \(GTEx\)](#). Para este análisis, sin embargo, he trabajado con datos ya pre-procesados en forma de dos archivos con formato CSV. Estos archivos (counts.csv y targets.csv) se pueden encontrar en mi perfil de GitHub siguiendo [este enlace](#).

El archivo counts.csv contiene el número de cuentas detectadas para cada gen en cada muestra del estudio. Es una tabla en la que cada observación (fila) corresponde a un gen reconocido por el proyecto Ensembl, y cada variable (columna) a una de las muestras de tejido del estudio. A partir de los datos de este archivo podemos analizar la expresión de cada gen en cada muestra.

El archivo targets.csv contiene información relevante acerca de cada muestra. Cada observación (fila) de la tabla corresponde a una muestra, y cada variable (columna) corresponde a una característica de la muestra como el código de identificación, el tejido de origen, el grupo experimental al que pertenece, o el sexo del sujeto del que procede la muestra.

3.2 Diseño experimental

El total de muestras es de 292; divididas en 236 muestras del grupo NIT, 42 muestras del grupo SFI y 14 muestras del grupo ELI. Para este informe se nos pidió seleccionar 10 muestras, aleatoriamente, de cada grupo para trabajar sobre un total de 30 muestras.

Debido a que los perfiles de expresión génica pueden variar según el sexo (Naqvi et al. 2019), he seleccionado las muestras de tal forma que, dentro de cada grupo experimental, la cantidad de muestras provenientes de hembras sea la misma que de varones (ver sección “[Obtención de los datos en bruto](#)”).

Como hay tres grupos, el análisis de expresión diferencial lo he efectuado sobre tres comparaciones: SFI-NIT, ELI-NIT y ELI-SFI.

3.3 Procedimiento seguido en el análisis

Los pasos seguidos para realizar el presente análisis han sido los siguientes:

1. Obtención de los datos de expresión en bruto.
2. Control de calidad de los datos.
3. Normalización.
4. Filtraje no específico.
5. Identificación de genes diferencialmente expresados.
6. Anotación de los resultados.
7. Comparación entre comparaciones.
8. Análisis del enriquecimiento de rutas.

3.3.1 Obtención de los datos en bruto

Como se ha señalado anteriormente, hemos recibido dos archivos con formato CSV conteniendo los datos pre-procesados de 292 muestras, de las cuales debíamos seleccionar al azar 10 de cada grupo experimental (en total 30 muestras).

Para seleccionar las muestras al azar, he dividido primero el total de muestras por grupos experimentales (NIT, SFI y ELI), y cada grupo lo he dividido por sexos (varón y hembra). De cada uno de esos grupos (NIT-varones, NIT-hembras, SFI-varones, etc.) he elegido al azar 5 muestras para un total de 30 muestras sobre las que realizar el informe.

Table 1: Distribución de la cantidad de muestras en este informe.

	varones	hembras
NIT	5	5
SFI	5	5
ELI	5	5

Todo el trabajo de selección y procesado de datos, y los análisis estadísticos, han sido realizados con el lenguaje de programación R. El entorno de desarrollo ha sido RStudio. Muchos de los paquetes usados como extensión de R han sido obtenidos de Bioconductor, con especial relevancia para este informe del paquete DESeq2. El listado completo de los paquetes usados y sus versiones se puede leer en el [apéndice B](#). El código utilizado para la manipulación de datos, y la composición de este informe, así como los archivos con los datos; puede consultarse en mi página de Github: https://github.com/jorgevallejo/RNAseq_analysis_PEC2

3.3.2 Control de calidad de los datos

Con el control de calidad pretendemos averiguar si los datos de alguna de las muestras presentan defectos o sesgos que desaconsejen usarlos, antes de continuar con el análisis.

En este caso hemos usado examinado los datos de expresión mediante diferentes representaciones gráficas, en busca de anomalías.

3.3.2.1 Distribución de los datos en bruto y normalización

Como la distribución de las cuentas presenta una alta asimetría, hemos transformado los datos de cuentas en pseudocuentas ($\log_2(\text{cuentas} + 1)$) para aproximar la distribución a la curva normal. Esta transformación ayudará a la visualización de los datos.

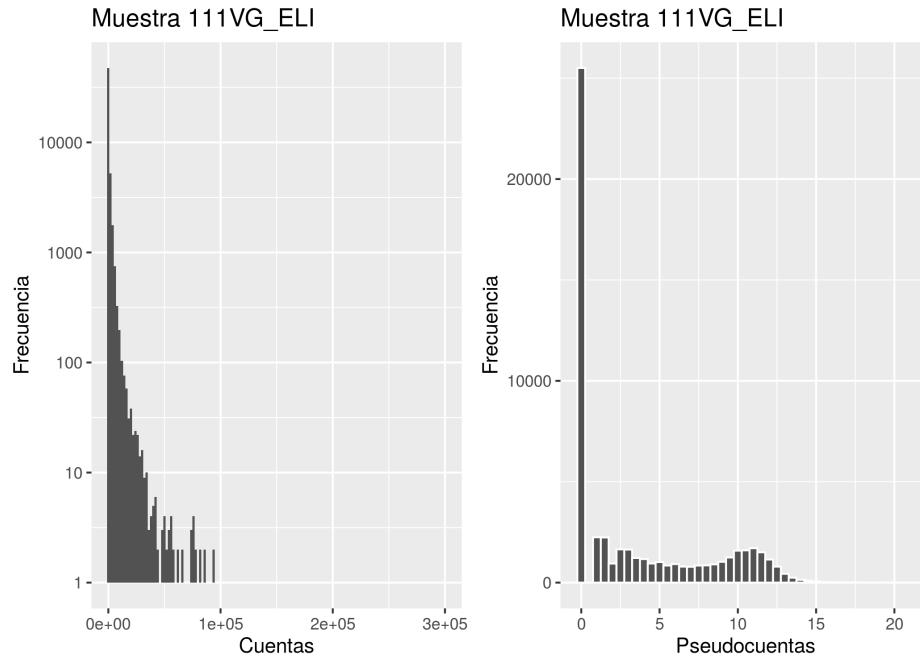


Figure 1: (izq.) Las cuentas de los genes en la muestra presentan una distribución fuertemente asimétrica hacia números bajos. La transformación en pseudocuentas (der.) produce un perfil de frecuencias más parecido a la distribución normal.

3.3.2.2 Distribución de pseudocuentas por muestra (diagramas de cajas)

Con los diagramas de cajas podemos observar y comparar la distribución de las pseudocuentas en las diferentes muestras. Esto comprobar si alguna de las muestras presenta una distribución radicalmente distinta del resto, lo que sería señal de algún error o problema con esa muestra:

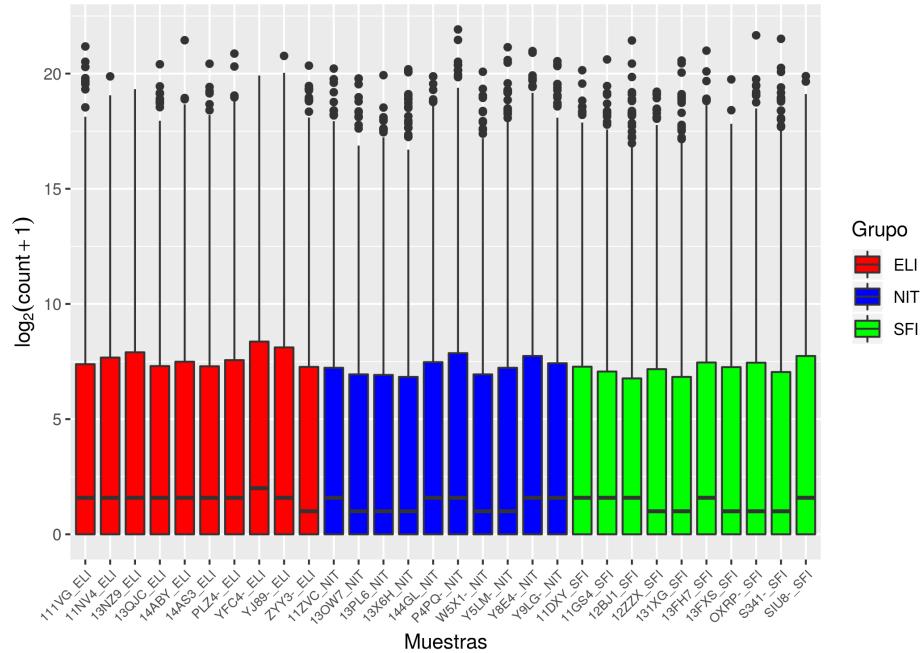


Figure 2: Diagramas de cajas construidos a partir de las pseudocuentas de los datos. Cada color representa un grupo experimental. La distribución de datos es similar en todas las muestras.

3.3.2.3 Distribución de pseudocuentas por muestra (histogramas de densidad)

Los histogramas nos ofrecen una perspectiva diferente sobre la distribución de las pseudocuentas, más detallada que los diagramas de cajas.

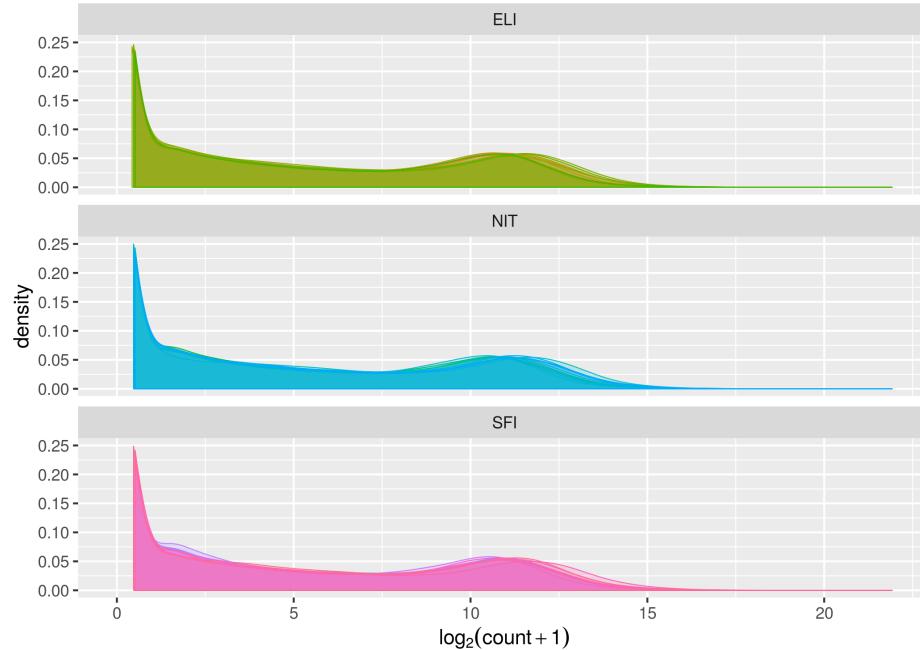


Figure 3: Histogramas suavizados de densidades de pseudocuentas en cada muestra, separados por grupo experimental.

En general, el perfil de los histogramas es muy parecido para todas las muestras y entre los diferentes grupos experimentales.

3.3.2.4 Mapa de calor

Los mapas de calor ayudan a explorar los parecidos y diferencias entre muestras. Mediante colores que representan las distancias entre muestras, y agrupando jerárquicamente muestras según su similaridad.

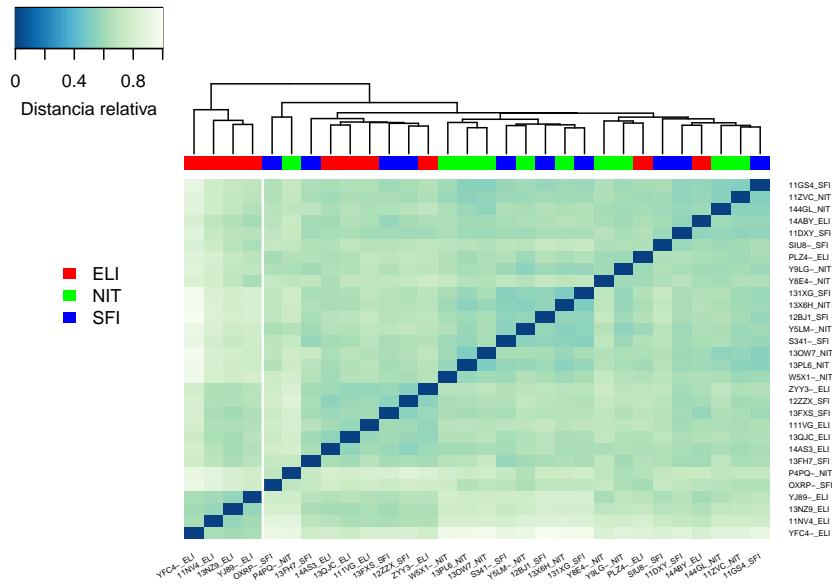


Figure 4: Mapa de calor de la distancia relativa entre muestras. El dendrograma en la parte superior agrupa las muestras jerárquicamente por proximidad.

Por lo que podemos ver en el mapa de calor, la mayor parte de las muestras del grupo ELI (infiltraciones amplias) se agrupan entre ellas (aunque en dos grupos diferentes), mientras que el resto de agrupaciones son mezclas de muestras de los grupos SFI (infiltraciones pequeñas) y NIT (sin infiltraciones).

3.3.2.5 Análisis de componentes principales de las muestras

Éste tipo de gráfica se puede usar para reconocer las características clave en sets de datos multidimensionales. Esto es, aquellas características que explican la mayor parte de la variabilidad en los datos.

Al aplicar el análisis sobre nuestros datos, esperamos que las muestras se agrupen por grupo experimental y, probablemente, por sexo.

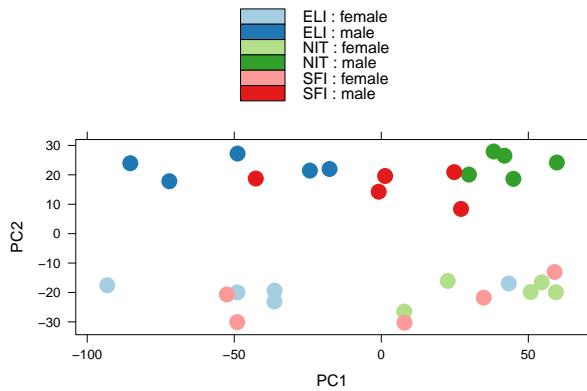


Figure 5: Análisis de componentes principales. PC1: Fuente principal de variabilidad entre muestras. Parece corresponder al grupo experimental. PC2: Segunda mayor fuente de variabilidad entre muestras. Separa claramente las muestras de hembras y varones.

En el resultado de la gráfica vemos que la separación entre grupos no es clara. Las muestras del grupo ELI en su mayoría se reúnen a la izquierda del gráfico, las muestras NIT a la derecha, y las muestras SFI se entremezclan con ambos grupos definiendo una posición más o menos central. La separación entre sexos sin embargo, es clara, con las muestras procedentes de varones en la zona superior y las de hembras en la inferior.

3.3.3 Filtraje no específico

Una primera medida para disminuir el ruido estadístico, y agilizar la velocidad de los cálculos, consiste en eliminar todos aquellos genes con ninguna o muy poca expresión. Podemos eliminarlos con relativa seguridad porque, si no se expresan en las muestras de ninguno de los dos grupos, tampoco nos están ofreciendo ninguna información respecto a expresión diferencial.

Eliminamos del dataset aquellos genes con menos de dos cuentas en total (teniendo en cuenta todas las muestras).

Total inicial de genes en el set de datos: 56 202 genes.

Total genes después de filtrar: 43 621 genes.

Genes filtrados: 12 581

3.3.4 Transformación de los datos: normalización

Las estrategias de normalización sirven para disminuir diferencias entre muestras provocadas por sesgos técnicos. Uno de los sesgos más comunes en RNA-seq son las diferencias en el número de lecturas en cada muestra, y para corregirlo hemos elegido el método conocido como *expresión de logaritmo relativo* (RLE por sus siglas en inglés).

Table 2: Factores de normalización aplicados a cada muestra (redondeados a 4 dígitos).

Muestras	Factores
111VG_ELI	0.8775
11NV4_ELI	1.0081
13NZ9_ELI	1.2457
13QJC_ELI	0.8719
14ABY_ELI	1.1514
14AS3_ELI	0.8305
PLZ4-_ELI	1.166
YFC4-_ELI	1.5881
YJ89-_ELI	1.4468
ZYY3-_ELI	0.8465
11ZVC_NIT	0.9131
13OW7_NIT	0.7631
13PL6_NIT	0.8322
13X6H_NIT	0.7983
144GL_NIT	1.0391
P4PQ-_NIT	1.5468
W5X1-_NIT	0.7082
Y5LM-_NIT	1.1009
Y8E4-_NIT	1.2545
Y9LG-_NIT	1.1996
11DXY_SFI	0.938
11GS4_SFI	0.8835
12BJ1_SFI	0.8431
12ZZX_SFI	0.7249
131XG_SFI	0.8268
13FH7_SFI	1.1227
13FXS_SFI	0.8794
OXRP-_SFI	1.4208
S341-_SFI	0.9164
SIU8-_SFI	1.1849

Obsérvese que ninguno de los factores se aleja excesivamente de la unidad, lo que apunta a que ninguna de las muestras estudiadas tiene una profundidad de lectura muy alejada del resto.

3.3.5 Transformación de los datos: estabilización de la varianza

Por lo general, el análisis de datos multidimensional es muy sensible a cambios en la varianza. Por eso hemos transformado los datos para que la medias de cuentas de los diferentes genes presenten aproximadamente la misma varianza.

Debido a que el tamaño del dataset empieza a entrar en un rango mediano (30 muestras), hemos utilizado el método de transformación estabilizadora de la varianza (VST).



Figure 6: La mayor variabilidad en la expresión génica se encuentra en aquellos genes con menos cuentas. La transformación estabilizadora de varianza (VST) elimina dicha variabilidad.

3.3.6 Distancia entre muestras

Con los datos transformados (normalización y estabilización de varianza), volvemos a examinar gráficamente la relación entre muestras.

3.3.6.1 Mapa de calor

Mediante un mapa de calor podemos qué muestras se parecen más entre ellas. Esperamos que muestras del mismo grupo experimental se agrupen entre sí.

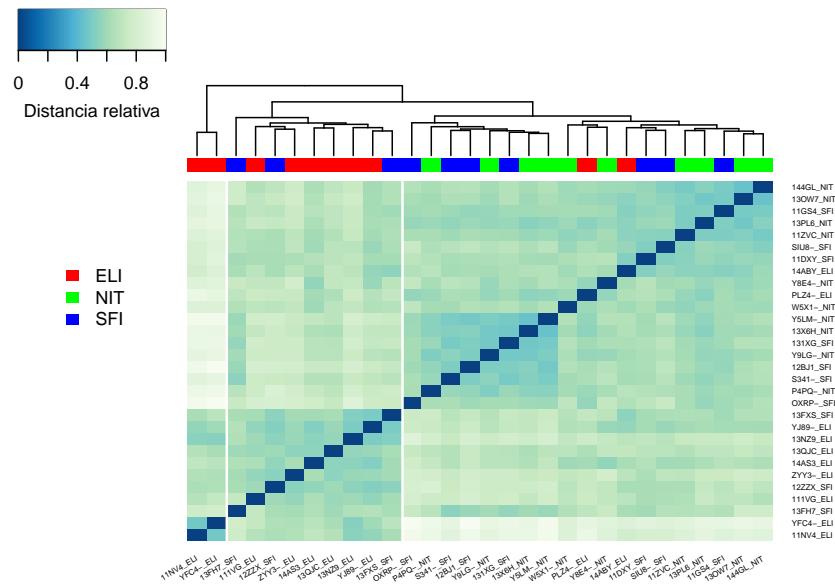


Figure 7: Mapa de calor de la distancia relativa entre muestras, a partir de datos transformados. El dendrograma en la parte superior agrupa las muestras jerárquicamente por proximidad.

En este nuevo mapa de calor de distancia entre muestras, con datos ya transformados mediante normalización y estabilización de la varianza, vemos un panorama similar el del mapa anterior. La mayor parte de las muestras del grupo ELI (infiltraciones amplias) se reúnen en dos grupos diferentes, mientras que el resto de agrupaciones son mezclas de muestras de los grupos SFI (infiltraciones pequeñas) y NIT (sin infiltraciones).

3.3.7 Análisis de componentes principales

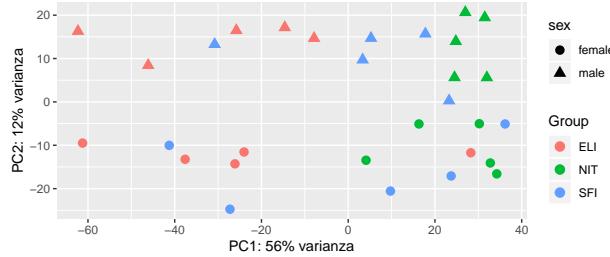


Figure 8: Gráfica de análisis de componentes principales. Cada marca representa una muestra, codificada según el grupo experimental al que pertenece (color) y el sexo (forma).

Al igual que observamos con los datos sin transformar, la mayor parte de la varianza parece deberse al grupo experimental, con las muestras del grupo ELI en su mayoría reunidas a la izquierda del gráfico, las muestras NIT a la derecha, y las muestras SFI entremezcladas con ambos grupos formando una suerte de puente. La aportación a la varianza por sexos sigue siendo clara, con las muestras procedentes de varones en la zona superior y las de hembras en la inferior.

4 Identificación y anotación de genes diferencialmente expresados

Después de explorar los datos para comprender su estructura, realizar análisis de calidad, normalizar los datos de expresión, estabilizar la varianza, y filtrar los genes menos informativos; llega el momento de analizar qué genes se expresan de forma diferente entre los grupos experimentales.

Para el análisis de este ensayo, consideramos una matriz de diseño siguiendo un **modelo lineal** de dos factores (*grupo experimental* y *sexo*). El factor *sexo* con dos niveles: *varón* y *hembra*. El factor *grupo experimental*, con tres niveles (*NIT*, *SFI* y *ELI*), se ha considerado como **factor principal**.

Así, la **fórmula** para el modelo lineal toma el siguiente aspecto:

$$\text{cuentas} = \text{sexo} + \text{grupo experimental}$$

El **análisis estadístico** se ha llevado a cabo mediante la función `DESeq()` del paquete *DESeq2* de Bioconductor (Love, Huber, and Anders 2014), que realiza una estimación de los factores de tamaño, estimación de la dispersión, y ajuste al modelo lineal generalizado (GLM) usando la distribución binomial negativa.

Como resultado del análisis estadístico obtenemos, para cada comparación entre grupos, los **cambios relativos de expresión** en cada gen (log2 fold change) y el **p-valor ajustado** que nos informa acerca de la significatividad estadística de ese cambio.

Como ejemplo, mostramos aquí los seis genes con menor p-valor ajustado de la comparación SFI-NIT:

Table 3: Ejemplo de genes con expresión diferencial entre los grupos SFI y NIT, ordenados por menor p-valor ajustado. Incluye el valor log2FC y el código ENSEMBL del gen.

	log2FC	p-valor ajustado
ENSG00000105369.5	4.66	2e-04
ENSG00000211673.2	5.64	2e-04
ENSG00000152137.2	1.31	3e-04
ENSG00000128438.6	4.60	3e-04
ENSG00000211899.3	4.00	3e-04
ENSG00000211934.2	4.86	3e-04

4.1 Anotación de las listas de genes

En las listas que hemos generado los genes están identificados por un código Ensembl. Sin embargo, a nosotros humanos nos resulta más cómodo trabajar con los símbolos o nombres de cada gen. Es por eso que hemos procedido a anotar las listas para incluir el símbolo de cada gen y su código Entrez.

No debemos de dejar de tener en cuenta, sin embargo, que las bases de datos Ensembl y Entrez utilizan criterios diferentes para definir qué es un gen, y tienen requisitos diferentes para incluir un gen en la base de datos. Esto tiene como consecuencia que muchos de los códigos de Ensembl no tengan su equivalente en códigos Entrez.

Ejemplos de cada listado de genes, ordenados por p-valor ajustado (decreciente), incluyendo el símbolo del gen y su nombre:

Table 4: SFI vs.NIT. Ejemplo de genes con expresión diferencial entre los grupos SFI y NIT, ordenados por menor p-valor ajustado. Incluye el símbolo y el nombre del gen correspondientes a la base de datos Entrez. NA = casos en los que no hay correspondencia entre el código ENSEMBL y la base de datos Entrez.

	Símbolo	Nombre	p-valor ajustado	log2FC
ENSG00000105369.5	CD79A	CD79a molecule	2e-04	4.66
ENSG00000211673.2	NA	NA	2e-04	5.64
ENSG00000152137.2	HSPB8	heat shock protein family B (small) member 8	3e-04	1.31
ENSG00000128438.6	NA	NA	3e-04	4.60
ENSG00000211899.3	NA	NA	3e-04	4.00
ENSG00000211934.2	NA	NA	3e-04	4.86

Table 5: ELI vs. NIT. Ejemplo de genes con expresión diferencial entre los grupos ELI y NIT, ordenados por menor p-valor ajustado.

	Símbolo	Nombre	p-valor ajustado	log2FC
ENSG00000136573.8	BLK	BLK proto-oncogene, Src family tyrosine kinase	2.00e-17	6.45
ENSG00000156738.13	MS4A1	membrane spanning 4-domains A1	2.80e-16	7.30
ENSG00000177455.7	CD19	CD19 molecule	2.80e-16	7.68
ENSG00000083454.17	P2RX5	purinergic receptor P2X 5	2.80e-16	5.64
ENSG00000035720.3	STAP1	signal transducing adaptor family member 1	1.48e-15	6.49
ENSG00000128438.6	NA	NA	2.04e-15	7.48

Table 6: ELI vs. SFI. Ejemplo de genes con expresión diferencial entre los grupos ELI y SFI, ordenados por menor p-valor ajustado.

	Símbolo	Nombre	p-valor ajustado	log2FC
ENSG00000145386.5	CCNA2	cyclin A2	2.0e-06	1.65
ENSG00000170006.7	TMEM154	transmembrane protein 154	6.0e-06	2.37
ENSG00000242142.1	NA	NA	7.0e-06	3.27
ENSG00000272906.1	NA	NA	9.0e-06	1.61
ENSG00000272144.1	NA	NA	1.5e-05	1.99
ENSG00000089225.15	TBX5	T-box transcription factor 5	1.5e-05	-4.60

Las listas completas de genes con expresión diferencial entre los diferentes grupos pueden descargarse como archivos en formato CSV desde estos enlaces:

[SFI vs NIT](#)

[ELI vs NIT](#) [ELI vs SFI](#)

4.2 Cuantificación de genes diferencialmente expresados

Si consideramos aceptable una fracción de falsos positivos del 10%, podemos considerar como significativos todos aquellos genes con un p-valor ajustado por debajo de 0.1, y calcular cuántos genes con expresión diferencial estadísticamente significativa detectamos en cada una de las comparaciones; además los podemos dividir entre genes regulados al alza (log2FC positivo) y regulados a la baja (log2FC negativo).

Table 7: Tabla resumen del número de genes con expresión diferencial estadísticamente significativa (FDR 10%) en cada una de las comparaciones, tanto a la alta como a la baja.

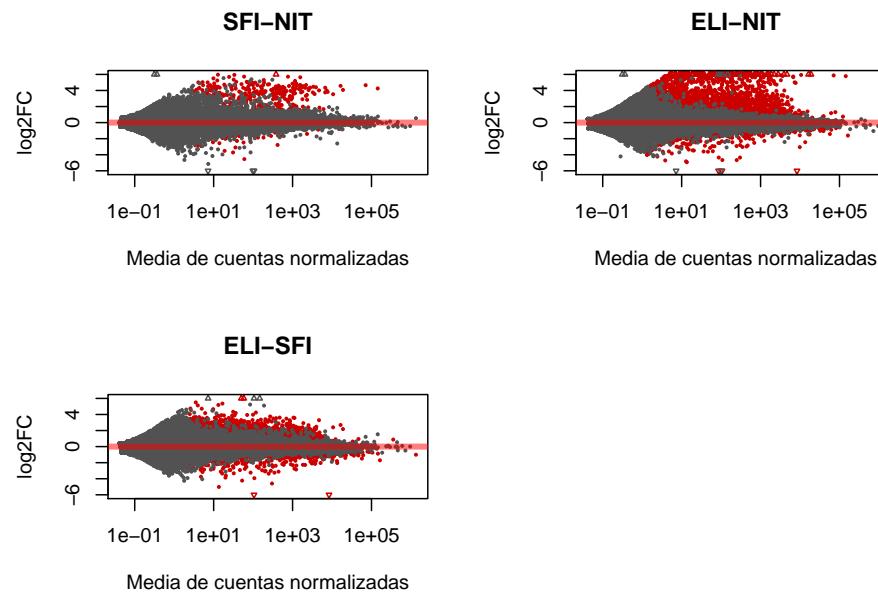
	SFI-NIT	ELI-NIT	ELI-SFI
Regulación al alza	434	3 172	3 384
Regulación a la baja	246	1 336	2 366
Total	680	4 508	5 750

4.2.1 Gráficas de resultados

4.2.1.1 Gráficas MA

Una forma más visual de explorar la cantidad de genes con expresión diferencial estadísticamente significativa es mediante gráficas MA comparando la expresión diferencial (log-fold changes), frente a la fuerza de expresión (media de cuentas):

\begin{figure}



\caption{Gráficas mostrando expresión diferencial (log-fold changes) frente a fuerza de expresión (media de cuentas normalizadas). Los puntos rojos representan los genes con un p-valor ajustado inferior a 0.1 (FDR 10%).} \end{figure}

Tanto en la tabla como en las gráficas podemos ver que en las comparaciones del grupo control (NIT) con los grupos patológicos (SFI y ELI), los genes regulados al alza (con más expresión en tejido patológico que

en tejido normal) son mayoría, aproximadamente el doble. En la comparación ELI-NIT la diferencia de proporción entre genes regulados al alza y a la baja no es tan grande.

Otro detalle interesante es que el total de genes diferencialmente expresados es mucho menor en la comparación SFI-NIT que en las otras dos comparaciones. Como ya vimos en el análisis de componentes principales y en el mapa de calor, las muestras NIT y SFI presentan perfiles de expresión génica más parecidos entre sí que con las muestras ELI.

4.2.1.2 Histograma de p-valores no ajustados

Para comprobar que no estamos introduciendo errores durante el análisis representamos, de cada comparación, un histograma de los p-valores no ajustados de cada gen. El perfil esperado es el de un pico de genes cercanos al valor 0, y el resto de genes repartidos por el abanico de probabilidades. Desviaciones de ese perfil indicarían algún fallo en los datos o en nuestro tratamiento de estos.

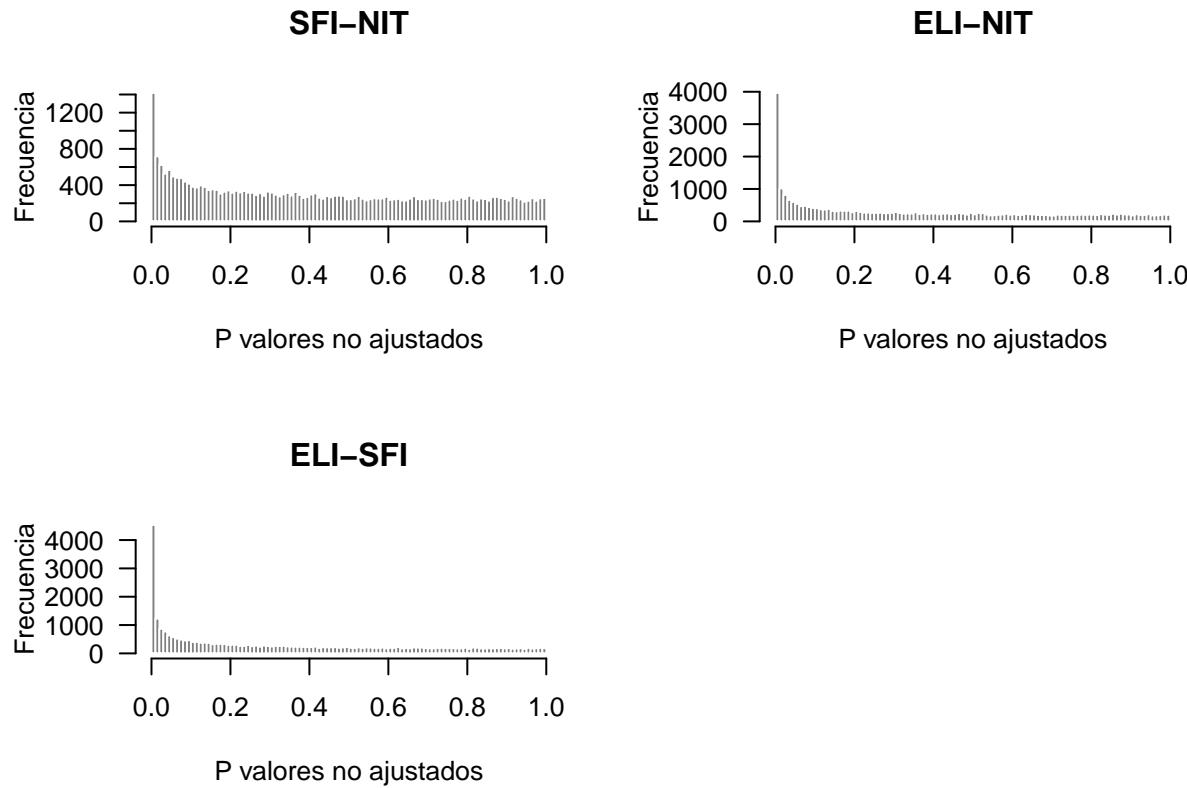


Figure 9: Histogramas de p-valores no ajustados de cada gen. El perfil observado es el esperado; un pico de genes cercanos al valor 0, y el resto de genes repartidos en el abanico de probabilidades.

4.2.1.3 Agrupación jerárquica de los genes más variables

Volvemos una vez más al mapa de calor, esta vez para explorar no sólo las relaciones entre las muestras, sino también entre los genes. Para generar esta figura hemos tenido en cuenta los 20 genes con mayor varibilidad entre muestras.

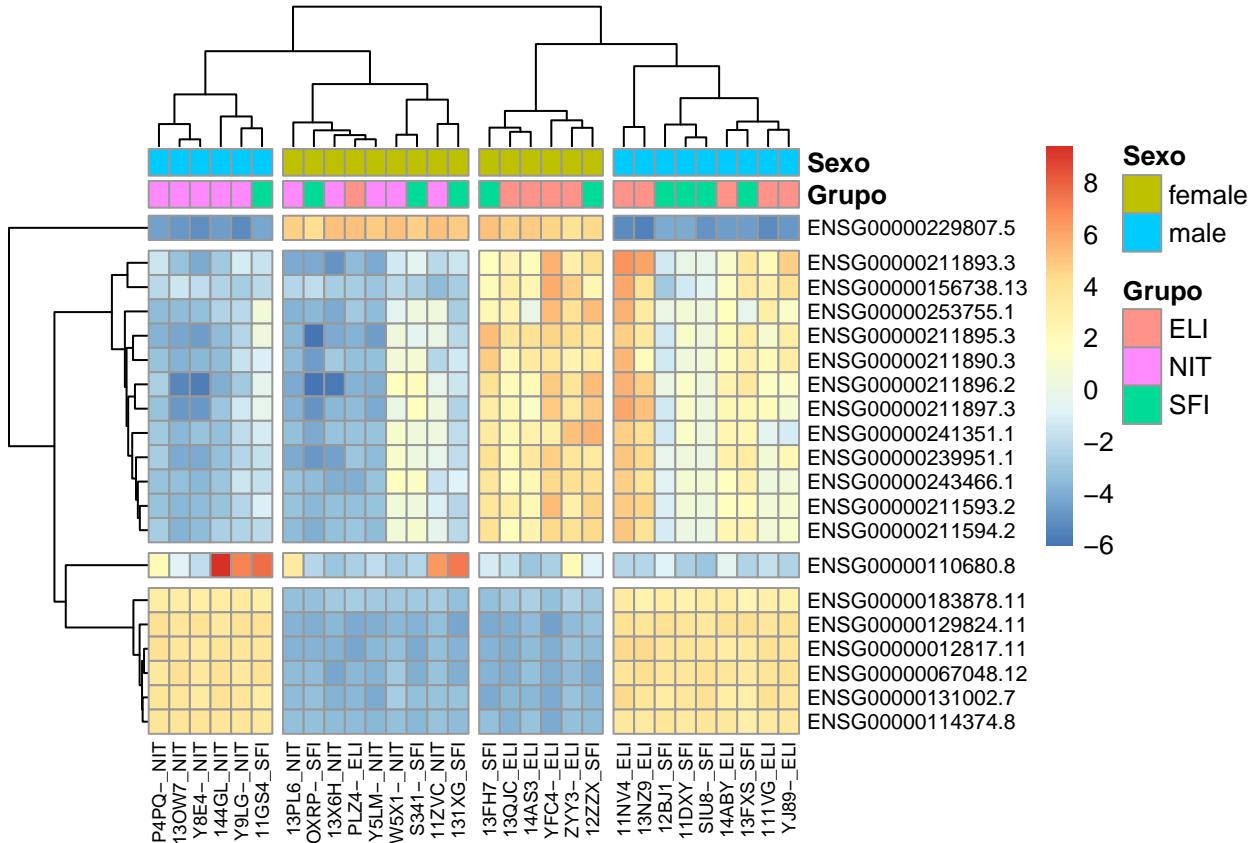


Figure 10: Mapa de calor de los 20 genes con mayor variación entre muestras. Cada columna representa una muestra, y cada fila representa uno de los genes.

Vemos que las muestras se reúnen en dos grandes grupos; uno de ellos el grupo control (NIT) junto con algunas muestras SFI, y el otro grupo una mezcla de muestras ELI y SFI. Resulta muy interesante que esas dos grandes agrupaciones de muestras estén divididas a su vez en sub-agrupaciones por sexo.

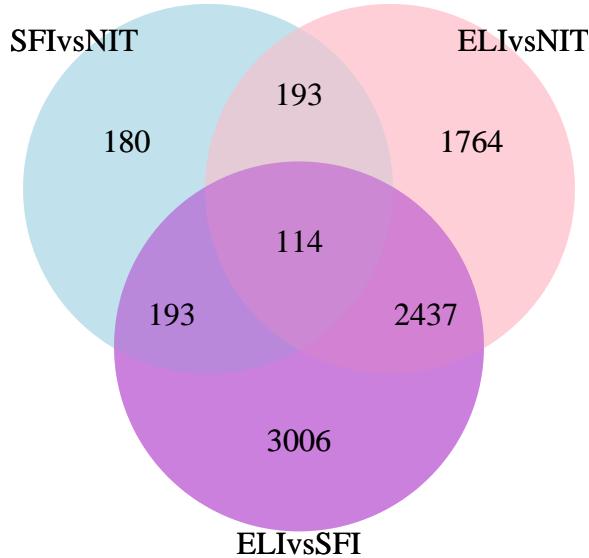
En cuanto a las agrupaciones de genes, vemos que esta lista de los 20 genes con mayor varianza entre muestras está muy influenciada por el sexo de la muestra y que dos de sus agrupaciones (la del extremo superior y la del extremo inferior) presentan un patrón de expresión diferencial dependiente principalmente del sexo.

5 Comparaciones múltiples

Debido a que los grupos experimentales corresponden a grupo control (NIT), síntomas suaves (SFI), y síntomas severos (ELI); puede ser de gran interés saber qué genes presentan expresión diferencial, por ejemplo, en las comparaciones del grupo control frente a los grupos con síntomas, y qué genes presentan expresión diferencial sólo en las comparaciones entre grupos con síntomas pero no cuando se compara con el grupo control.

5.1 Diagrama de Venn

Una de las formas más cómodas de presentar esta información son los diagramas de Venn, donde el solapamiento entre los círculos representa los genes en común con expresión diferencial entre las comparaciones:



6 Significatividad biológica

Una vez tenemos nuestras listas de genes anotadas, un herramienta más para interpretar los resultados del estudio es el examen de la significatividad biológica. En este informe, lo que hemos hecho es, a partir de las listas de genes con comportamiento diferencial, comprobar si existen funciones, procesos biológicos o rutas moleculares que aparezcan con más frecuencia en estas listas que en el resto de genes analizados.

Como listas de genes hemos utilizado las siguientes:

SFIvsNIT - lista de genes con comportamiento diferencial en la comparación entre los grupos NIT y SFI.

ELIvsNIT - lista de genes con comportamiento diferencial en la comparación entre los grupos NIT y SFI.

ELIvsSFI - lista de genes con comportamiento diferencial en la comparación entre los grupos NIT y SFI.

Universo - lista de todos los genes con código Entrez en la [base de datos org.Hs.eg.db](#).

6.1 Test de sobrerepresentación de términos GO

El análisis estadístico lo hemos realizado con la función `enrichGO()` del paquete `clusterProfiler` para el lenguaje R. Esta función devuelve un listado de términos GO estadísticamente más representados en nuestra lista de genes de interés respecto de una lista de referencia (en este caso la lista que hemos llamado *Universo*).

Para los efectos de este informe nos hemos limitado a explorar la sobrerepresentación referida a los **procesos biológicos**; dejando a parte por motivos de tiempo y espacio el estudio de sobrerepresentación en los términos referidos a funciones moleculares y componentes celulares.

Sí hemos incluido, sin embargo, todas las comparaciones entre grupos (SFI vs NIT, ELI vs NIT y ELI vs SFI).

El resultado del test es un listado de términos GO (junto con su descripción y estadísticos de test) sobrerepresentados en nuestra lista, respecto a la lista de referencia. Los tres listados se pueden descargar completos en formato CSV desde estos enlaces:

[GO SFI vs NIT](#)

GO ELI vs NIT

GO ELI vs SFI

Table 8: Ejemplo de términos GO sobrerepresentados en la lista de genes con expresión diferencial para la comparación SFI vs NIT

ID	Description	BgRatio
GO:0051480	regulation of cytosolic calcium ion concentration	357/18670
GO:0030098	lymphocyte differentiation	353/18670
GO:0006874	cellular calcium ion homeostasis	458/18670
GO:0055074	calcium ion homeostasis	471/18670
GO:0007204	positive regulation of cytosolic calcium ion concentration	319/18670
GO:0072503	cellular divalent inorganic cation homeostasis	493/18670

6.2 Visualización de resultados: Gráficos de barras

Una forma habitual de representar los resultados de sobrerepresentación de términos GO es mediante gráficos de barras ordenados por el p-valor ajustado en orden creciente. En los siguientes gráficas se muestran, para cada comparación, los primeros veinte términos GO. El color de cada barra representa el p-valor ajustado, y su longitud el número de genes asociados al término GO:

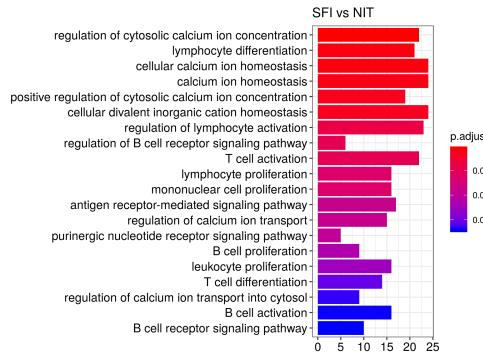


Figure 11: Gráfico de barras para la comparación SFIT vs NIT.

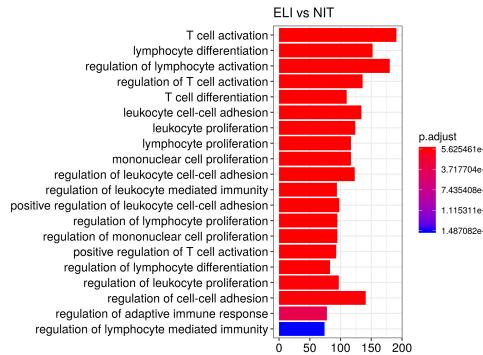


Figure 12: Gráfico de barras para la comparación ELI vs NIT.

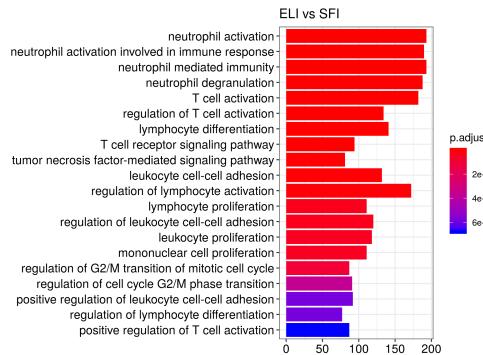


Figure 13: Gráfico de barras para la comparación ELI vs SFI.

Vemos que la mayoría de términos expuestos en las gráficas están relacionados con la inflamación y la respuesta inmunológica. En la comparación ELI-SFI llama la atención la aparición de rutas relacionadas con

los neutrófilos y la regulación del ciclo celular.

7 Discusión

Una de las características en los datos que parece haber introducido dificultad a la hora de analizarlos, es que las características de los tejidos en los tres grupos experimentales parece seguir un continuo en lugar de ser tres grupos rígidamente separados. Esto por lo menos es lo que parecen sugerir tanto el análisis de componentes principales como el mapa de calor de distancia entre muestras. Las muestras se representan bastante entremezcladas, siendo los grupos más distantes NIT y ELI con el grupo SFI haciendo de puente. También parece que las características de SFI y NIT son más parecidas entre ambos grupos que comparados con ELI.

Un aspecto que me habría gustado explorar más, pero no he podido por falta de tiempo, es el efecto del sexo. En el análisis de componentes principales se revela como importante a la hora de explicar la varianza de los datos muestrales, y en el mapa de calor de genes se revela que los patrones de expresión de los genes con mayor variabilidad entre muestras están muy influenciados por el sexo.

Me gustaría incidir en el cuidado a la hora de comunicar con qué repositorio de genes se está trabajando, ya que este mismo informe hemos visto las grandes diferencias entre ENSEMBL y Entrez, y que en muchas ocasiones no hay traducción directa entre los genes de uno y otro repositorio. Es importante entender que en cada repositorio la definición de lo que se considera como un gen, el método con el que entra al repositorio, y cómo se mantiene el propio repositorio son diferentes. Hay que conocer cuáles son las diferencias y qué significa trabajar con un repositorio u otro.

También quisiera comentar otra parte del análisis en la que hubiera querido profundizar si hubiese tenido más tiempo. Al hacer las comparaciones múltiples, originalmente tenía intención de generar una lista con los genes con expresión diferencial de las comparaciones SFI-NIT y ELI-NIT, menos los genes de la comparación ELI-SFI. Y otra lista con los genes exclusivos de la comparativa ELI-SFI. Estas listas las habría usado en el análisis de significatividad biológica para examinar los cambios de expresión génica exclusivos de patológico frente a sano por un lado, y los exclusivos de patológico leve frente a patológico severo.

Por último, una corta reflexión acerca de la enorme cantidad de datos producida en estos análisis. Aunque el resultados de los mismos es como una destilación y resumen de los datos brutos de los que partimos, las listas de genes y términos de interés son tan abultadas que no se pueden incluir en el propio informe, sino que hay que enlazarlas como ficheros adicionales.

8 Apéndice A: Código

El documento original en formato .Rmd, que incluye el código completo en lenguaje R usado para generar este informe, se puede consultar y descargar en el siguiente repositorio de Github: [jorgevallejo/RNAseq_analysis_PEC2](https://github.com/jorgevallejo/RNAseq_analysis_PEC2)

9 Apéndice B: Reproducibilidad

```
sessionInfo() # For better reproducibility

## R version 3.6.3 (2020-02-29)
## Platform: i686-pc-linux-gnu (32-bit)
## Running under: Ubuntu 16.04.7 LTS
##
## Matrix products: default
```

```

## BLAS:    /usr/lib/libblas/libblas.so.3.6.0
## LAPACK:  /usr/lib/lapack/liblapack.so.3.6.0
##
## locale:
## [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C
## [3] LC_TIME=en_GB.UTF-8       LC_COLLATE=en_US.UTF-8
## [5] LC_MONETARY=en_GB.UTF-8   LC_MESSAGES=en_US.UTF-8
## [7] LC_PAPER=en_GB.UTF-8     LC_NAME=C
## [9] LC_ADDRESS=C              LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] grid      parallel  stats4   stats    graphics grDevices utils
## [8] datasets  methods   base
##
## other attached packages:
## [1] enrichplot_1.6.1           clusterProfiler_3.14.3
## [3] VennDiagram_1.6.17         futile.logger_1.4.3
## [5] pheatmap_1.0.12            genefilter_1.68.0
## [7] org.Hs.eg.db_3.10.0        AnnotationDbi_1.48.0
## [9] dplyr_1.0.0                DESeq2_1.26.0
## [11] SummarizedExperiment_1.16.1 DelayedArray_0.12.3
## [13] BiocParallel_1.20.1        matrixStats_0.56.0
## [15] Biobase_2.46.0             GenomicRanges_1.38.0
## [17] GenomeInfoDb_1.22.1        IRanges_2.20.2
## [19] S4Vectors_0.24.4           BiocGenerics_0.32.0
## [21] gplots_3.0.4               RColorBrewer_1.1-2
## [23] reshape_0.8.8              gridExtra_2.3
## [25] ggplot2_3.2.1              knitr_1.25
##
## loaded via a namespace (and not attached):
## [1] stringr_1.4.0          httr_1.4.1          tidyrr_1.1.0
## [4] purrr_0.3.4            jsonlite_1.6        ggraph_2.0.3
## [7] gdata_2.18.0            gtools_3.8.1        Formula_1.2-3
## [10] rlang_0.4.6             RCurl_1.98-1.1     htmlTable_1.13.2
## [13] triebeard_0.3.0         generics_0.0.2      blob_1.2.1
## [16] bitops_1.0-6            base64enc_0.1-3    RSQLite_2.2.0
## [19] pillar_1.4.2            reshape2_1.4.3      R6_2.4.0
## [22] XVector_0.26.0          polyclip_1.10-0    bit_1.1-15.2
## [25] plyr_1.8.4              Hmisc_4.2-0         stringi_1.4.3
## [28] lifecycle_0.2.0          viridis_0.5.1       tidygraph_1.2.0
## [31] munsell_0.5.0            europePMC_0.4       tweenr_1.0.1
## [34] gridGraphics_0.5-0       rvcheck_0.1.8       urltools_1.7.3
## [37] highr_0.8                rstudioapi_0.10    htmlwidgets_1.5.1
## [40] DBI_1.1.0                evaluate_0.14       memoise_1.1.0
## [43] DOSE_3.12.0              foreign_0.8-76     pkgconfig_2.0.3
## [46] tools_3.6.3              acepack_1.4.1       vctrs_0.3.1
## [49] xtable_1.8-4              locfit_1.5-9.4     cluster_2.1.0
## [52] ggplotify_0.0.5           lambda.r_1.1.9      compiler_3.6.3
## [55] DESeq_1.38.0              caTools_1.17.1.2    DO.db_2.9
## [58] ggridges_0.5.2            annotate_1.64.0     igraph_1.2.5
## [61] GenomeInfoDbData_1.2.2    labeling_0.3         gtable_0.3.0
## [64] glue_1.4.1                survival_3.1-12     digest_0.6.21
## [67] Matrix_1.2-18             htmtools_0.4.0      cowplot_1.0.0

```

```

## [70] XML_3.99-0.3           KernSmooth_2.23-17      data.table_1.12.4
## [73] splines_3.6.3            ggforce_0.3.2          farver_2.0.3
## [76] progress_1.2.2           ggrepel_0.8.2          withr_2.1.2
## [79] backports_1.1.5          lattice_0.20-41        lazyeval_0.2.2
## [82] crayon_1.3.4             magrittr_1.5           checkmate_1.9.4
## [85] qvalue_2.18.0            colorspace_1.4-1       futile.options_1.0.0
## [88] xml2_1.3.2               Rcpp_1.0.2              prettyunits_1.1.1
## [91] geneplotter_1.64.0        zlibbioc_1.32.0        GO.db_3.10.0
## [94] viridisLite_0.3.0         bit64_0.9-7            scales_1.0.0
## [97] BiocManager_1.30.10       tibble_2.1.3            latticeExtra_0.6-28
## [100] rmarkdown_1.17            yaml_2.2.0              nnet_7.3-14
## [103] graphlayouts_0.7.0        MASS_7.3-52             fgsea_1.12.0
## [106] fastmatch_1.1-0          GOSemSim_2.12.1        rpart_4.1-15
## [109] hms_0.5.3                tidyselect_1.1.0        xfun_0.10

```

Referencias

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for Rna-Seq Data with Deseq2.” *Genome Biology* 15 (12). Springer: 550.

Naqvi, Sahin, Alexander K Godfrey, Jennifer F Hughes, Mary L Goodheart, Richard N Mitchell, and David C Page. 2019. “Conservation, Acquisition, and Functional Impact of Sex-Biased Gene Expression in Mammals.” *Science* 365 (6450). American Association for the Advancement of Science: eaaw7317.