

PEC2 - análisis de datos de ultrasecuenciación

Sonia Doblado Martín

11 de junio, 2020

Índice

1. Abstract	1
2. Introducción	1
3. Métodos	2
3.1. Tratamiento de los datos	2
3.2. Sample distances	3
3.3. PCA	4
3.4. Expresión diferencial	6
4. Plot de conteo	7
4.1. Anotación de los resultados	7
5. Enrichment Analysis	8
6. Resultados	8
7. Sumario de resultados	14
8. Comentarios	15
Bibliografía	15

1. Abstract

El presente informe contiene el estudio de los resultados del análisis RNA-Seq a muestras de tejido tiroideo. Se comparan tres tipos de infiltración (NIT, SFI y ELI) y se muestran resultados de expresión diferencial en dichas comparaciones. Con los genes diferencialmente expresados, se realiza un análisis de significación biológica.

Para la reproducción del análisis aquí realizado, el código R empleado se encuentra en el siguiente repositorio **Github**

2. Introducción

El archivo targets.csv contiene la información de las muestras de un estudio obtenido del repositorio (GTEx1). Este repositorio contiene datos de múltiples tipos en un total de 54 tejidos. Nosotros nos centraremos en los datos de expresión (RNA-seq) pertenecientes a un análisis del tiroides en donde se compara tres tipos de infiltración medido en un total de 292muestras pertenecientes a tres grupos:

- Not infiltrated tissues (NIT)

- Small focal infiltrates (SFI)
- Extensive lymphoid infiltrates (ELI)

También se proporciona el archivo `counts.csv`, que contiene el conteo de 56202 genes para las 292 muestras.

Al desconocer los detalles del estudio de referencia, a la hora de elegir umbrales se han seleccionado los comúnmente aceptados o dejado los valores por defecto de las funciones empleadas. Se asume también que los individuos muestreados son humanos. En este informe se incluyen sólo las informaciones que se consideran relevantes para evaluar las comparaciones solicitadas. Más información, incluyendo figuras adicionales, puede encontrarse en el repositorio Github mencionado anteriormente.

3. Métodos

Para el informe se han utilizado los archivos `targets.csv` y `counts.csv`. De entre las 292 muestras, se han seleccionado las 30 muestras solicitadas por el enunciado, 10 de cada tipo. Se realizarán tres comparaciones, ELI vs NIT (EvsN), ELI vs SFI (EvsS) y SFI vs NIT (SvsN).

3.1. Tratamiento de los datos

El primer paso será unir esas 10 muestras de cada grupo en una única tabla que contenga los conteos del conjunto de 30 muestras. Para ello, se unifican los nombres de muestras de los archivos `targets.csv` y `counts.csv`. Después, se elabora una tabla que contenga las columnas de `counts.csv` cuyo encabezamiento coincida con el nombre de las muestras escogidas de la tabla de `targets.csv`.

Para añadir la información de la tabla `targets.csv` en el nuevo elemento creado (llamado `conteo`), se selecciona dicha información sólo de las muestras seleccionadas y se crea una nueva tabla (llamada `infoTargets`). Esta tabla será la utilizada como `colData` en la función `DESeqDataSetFromMatrix` del paquete `DESeq2` para crear el objeto `DESeqDataSet`, necesario para continuar el análisis.

Para la filtración de los datos, sólo se eliminarán aquellas filas del `DESeqDataSet` que no contengan ningún conteo o tengan uno sólo a lo largo de todas las muestras. No se filtrará a mayores al no tener conocimiento de los detalles del estudio de referencia, por lo que se opta por la opción con la que se pierde el mínimo de información. Tras realizar esta operación, quedan 43186 genes.

Los valores de la matriz de conteo no se han normalizado ya que el modelo del paquete `DeSeq2` supone que los datos introducidos son los datos obtenidos en el experimento, en forma de una matriz numérica.

3.1.1. Transformación de los datos

Para una mejor visualización de los resultados se realiza una transformación de los datos. Se ha empleado la función `vst`, que devuelve un objeto `DESeqTransform`. Los datos transformados ya no son un conteo, se almacenan en el slot `assay`, y se sigue pudiendo acceder a la información de `colData`. A continuación se muestran dos gráficos con diferentes transformaciones, `log2` y `VST`. Se ha escogido `VST` porque, como puede observarse, en este caso se comprimen las diferencias con genes con un conteo bajo, para los cuales los datos no ofrecen mucha información sobre su expresión diferencial.

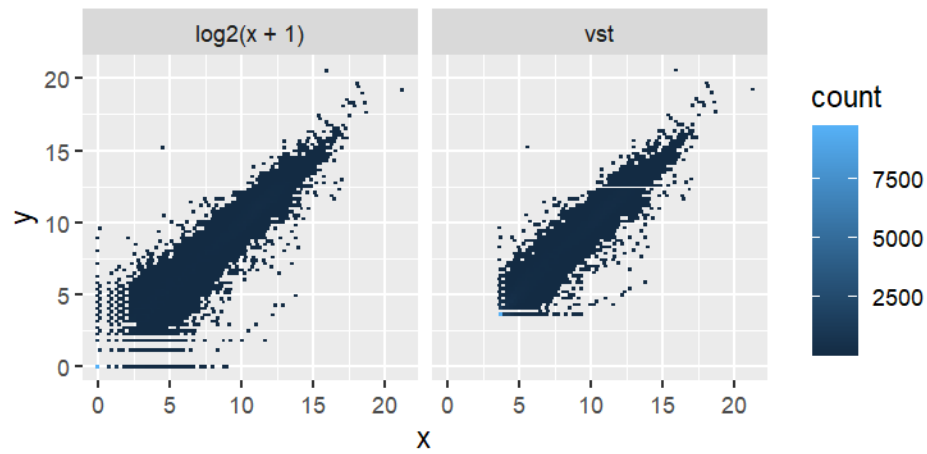


Figura 1: Transformaciones de los datos según los métodos log2 (izq) y VST (dcha)

3.2. Sample distances

En caso de que fuese práctico conocer las diferencias de unas muestras respecto a otras, se incluye un mapa de calor con las distancias intermuestrales.

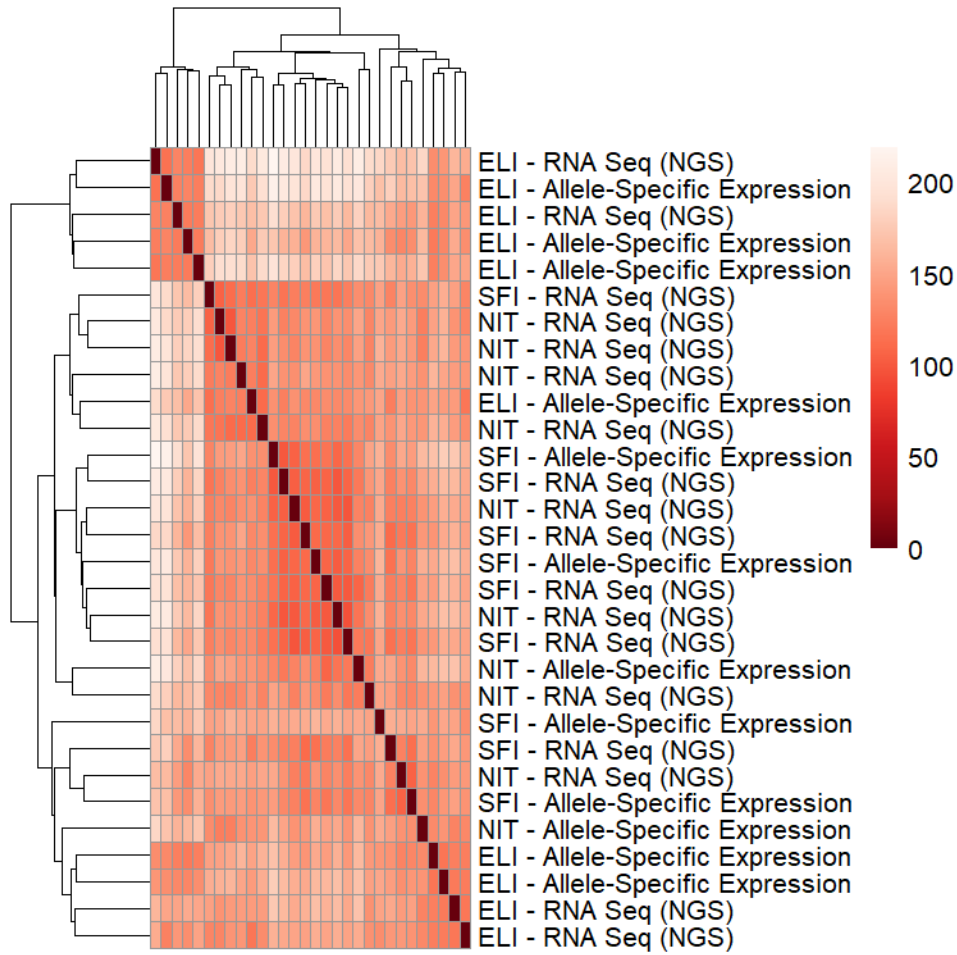


Figura 2: Heatmap de las distancias entre las muestras, basados en los datos transformados por el método VST

3.3. PCA

Se realizó un análisis de componentes principales, cuyo resultado puede verse en la Figura 3 y se añade un gráfico MDS que ayuda a la interpretación de los datos de distancias entre muestras (Figura 4).

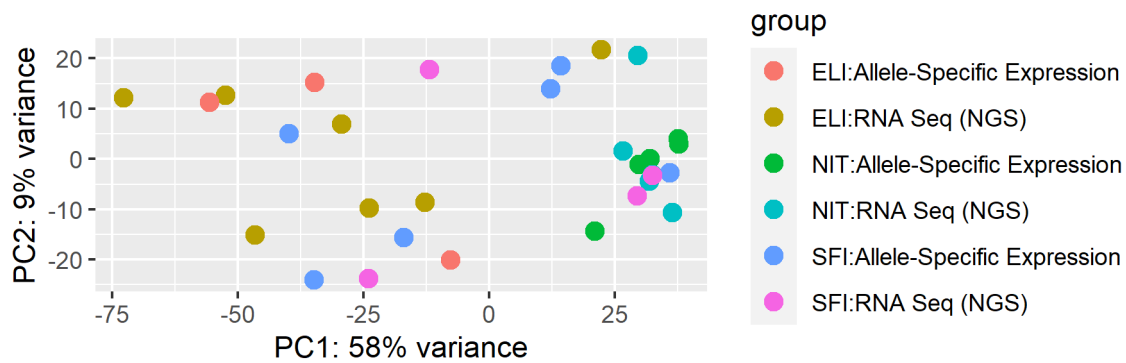


Figura 3: PCA plot usando los datos VST. Cada combinación de tratamiento y sexo tiene un color único

En el análisis de los componentes principales (PCA) se observa que el primer componente del PCA es responsable del 58 % de la variabilidad de las muestras. Como se observa en el gráfico de la figura 3, esta variabilidad podría atribuirse ya a los diferentes grupos ELI, NIT y SFI. Esta diferencia es más clara entre las muestras ELI y el resto, sobre todo entre ELI y NIT. Se muestra también en los gráficos la diferencia entre las muestras *NGS* y *allele-specific expression* para comprobar visualmente que dicha diferencia es despreciable.

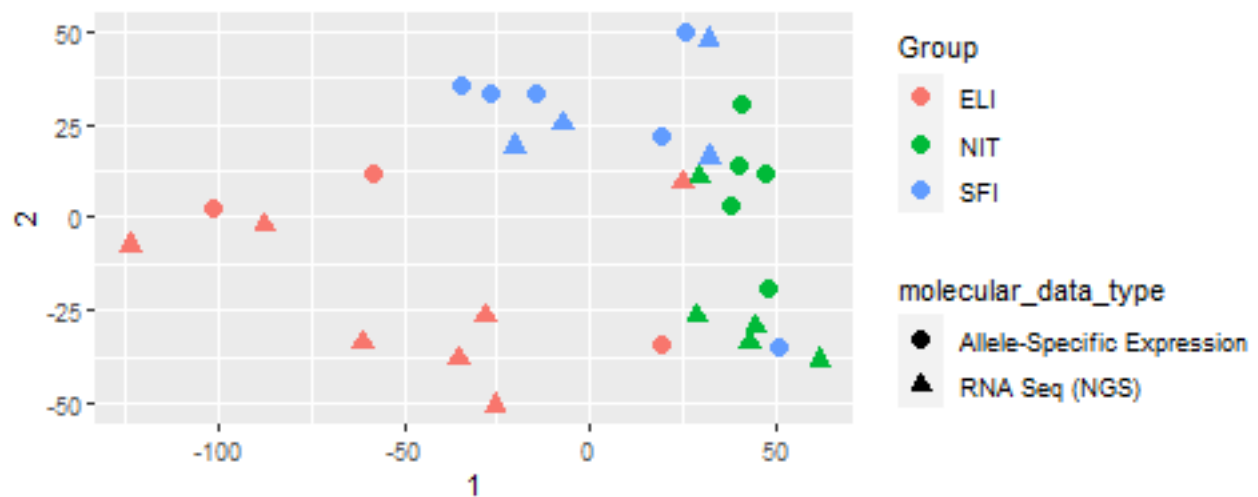


Figura 4: MDS plot usando los datos VST. Cada tipo de infiltración tiene un color único y cada tipo de datos tiene una forma única

3.4. Expresión diferencial

Para el análisis de la expresión diferencial, se emplea la función `DESeq` sobre los datos **sin transformar**.

Del objeto `DESeqDataSet` obtenido, se extraen los resultados de interés. En este caso, serán contrastes entre las muestras de diferentes tipos de infiltración dos a dos, como solicita el enunciado (en el orden EvsN, EvsS y SvsN).

4. Plot de conteo

En el gráfico 5 se observan los conteos del objeto `DESeqDataSet` para cada uno de los tipos de infiltración. Se incluyen los genes de la comparación `EvsN`:

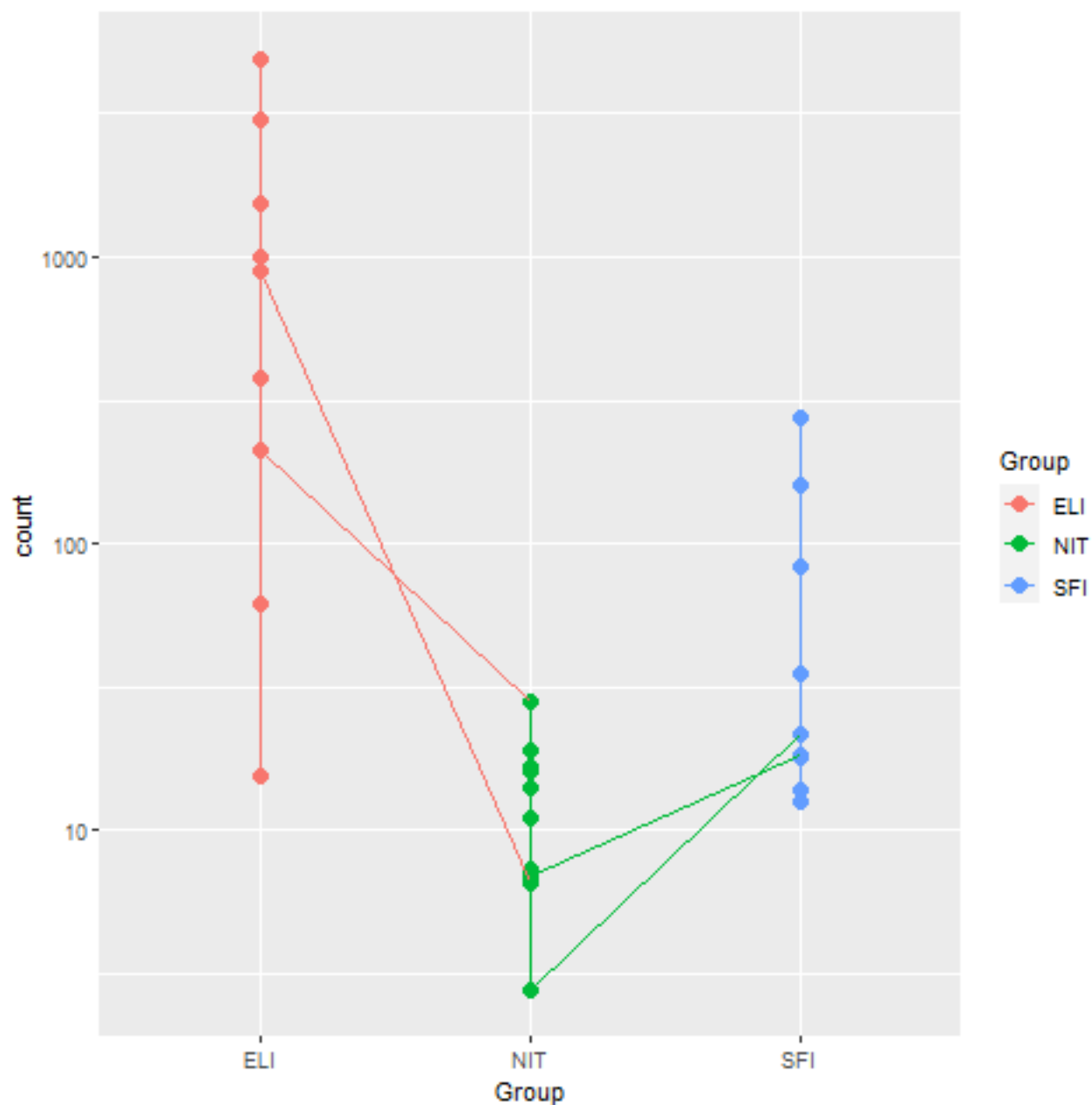


Figura 5: Count plot obtenido para cada uno de los distintos tipos de infiltración utilizados en el experimento, para los genes incluidos en la comparación `EvsN`. La gráfica no varía sustancialmente si se incluyen los genes de `EvsS` y `SvsN`.

4.1. Anotación de los resultados

Se añaden las anotaciones de ENTREZID, GO y SYMBOL correspondientes a las anotaciones ENSEMBL ya presentes en los datos. Para ello se utiliza la función `mapIDs` del paquete `AnnotationDbi`, y el paquete de

anotación org.Hs.eg.db.

Para buscar correspondencias, se han eliminado las versiones de las anotaciones de ENSEMBL (es decir, toda aquella información del nombre ENSEMBL posterior al punto). A continuación se muestra una parte de los resultados obtenidos para la comparación ELI vs NIT:

log2 fold change (MLE): Group ELI vs NIT

Wald test p-value: Group ELI vs NIT

DataFrame with 6 rows and 9 columns

	baseMean	log2FoldChange	lfcSE	stat	pvalue
	<numeric>	<numeric>	<numeric>	<numeric>	<numeric>
ENSG00000117322	478.742	6.78408	0.649129	10.45105	1.44917e-25
ENSG00000167483	566.440	6.28964	0.614395	10.23714	1.35178e-24
ENSG00000177455	711.711	7.36347	0.749873	9.81962	9.26879e-23
ENSG00000160856	485.075	5.63517	0.580375	9.70954	2.74576e-22
ENSG00000083454	879.935	6.19594	0.639632	9.68672	3.43368e-22
ENSG00000269404	240.654	5.59562	0.597603	9.36344	7.71842e-21
	padj	symbol	entrez	GO	
	<numeric>	<character>	<character>	<character>	
ENSG00000117322	4.43793e-21	CR2	1380	GO:0001618	
ENSG00000167483	2.06985e-20	NIBAN3	199786	NA	
ENSG00000177455	9.46158e-19	CD19	930	GO:0001923	
ENSG00000160856	2.10216e-18	FCRL3	115352	GO:0002638	
ENSG00000083454	2.10306e-18	P2RX5	5026	GO:0001614	
ENSG00000269404	3.85728e-17	SPIB	6689	GO:0000981	

Los archivos guardados incluyen los genes ordenados según su p-valor. Se incluyen también estadísticos como `log2FoldChange`, `lfcSE` y `padj`, que pueden ser útiles para otras interpretaciones aparte de las aquí presentadas.

5. Enrichment Analysis

Una vez tenemos los genes anotados, procedemos a la visualización de los resultados de la expresión diferencial obtenidos. En este informe se muestran varios ejemplos de visualización. Para mostrar una visión general de la expresión diferencial se realiza un **volcano plot** para cada una de las comparaciones.

Por último, para ayudar a interpretar la importancia biológica de los resultados del experimento, se realizó un **test enrichgo** (Gene Enrichment Analysis) mediante el paquete **clusterProfiler** (Guangchuang 2020). Para representar los resultados de este test, se han utilizado diagramas de barras mediante la función **barplot**, un diagrama en forma de red especificando qué genes son los que actúan mediante la función **cnetplot** y un gráfico **emapplot**. Se han calculado los resultados para las ontologías **BP**(Biological Process), **CC**(Cellular Component) y **MF** (Molecular Function), para tener una visión más general de la importancia biológica de los resultados. El criterio de selección de genes diferencialmente expresados ha sido un p-valor ajustado menor de 0.05. Se seleccionan todos los genes que tienen al menos una anotación en la base de datos GO.

Las figuras resultantes de los análisis han sido guardadas en archivos .png y .pdf por si esta información fuese útil para este u otros estudios.

6. Resultados

Los resultados para cada comparación tras utilizar la función **DESeq** son:

```
out of 43182 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)      : 3112, 7.2%
```



```
LFC < 0 (down)      : 1640, 3.8%
outliers [1]        : 0, 0%
low counts [2]       : 14237, 33%
(mean count < 2)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

NULL

```
out of 43182 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)        : 3573, 8.3%
LFC < 0 (down)      : 1486, 3.4%
outliers [1]        : 0, 0%
low counts [2]       : 12562, 29%
(mean count < 1)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

NULL

```
out of 43182 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)        : 982, 2.3%
LFC < 0 (down)      : 676, 1.6%
outliers [1]        : 0, 0%
low counts [2]       : 15911, 37%
(mean count < 2)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

NULL

En los siguientes plots se observan resaltados en rojo los genes con un p-valor ajustado menor de 0.1 para cada una de las comparaciones.

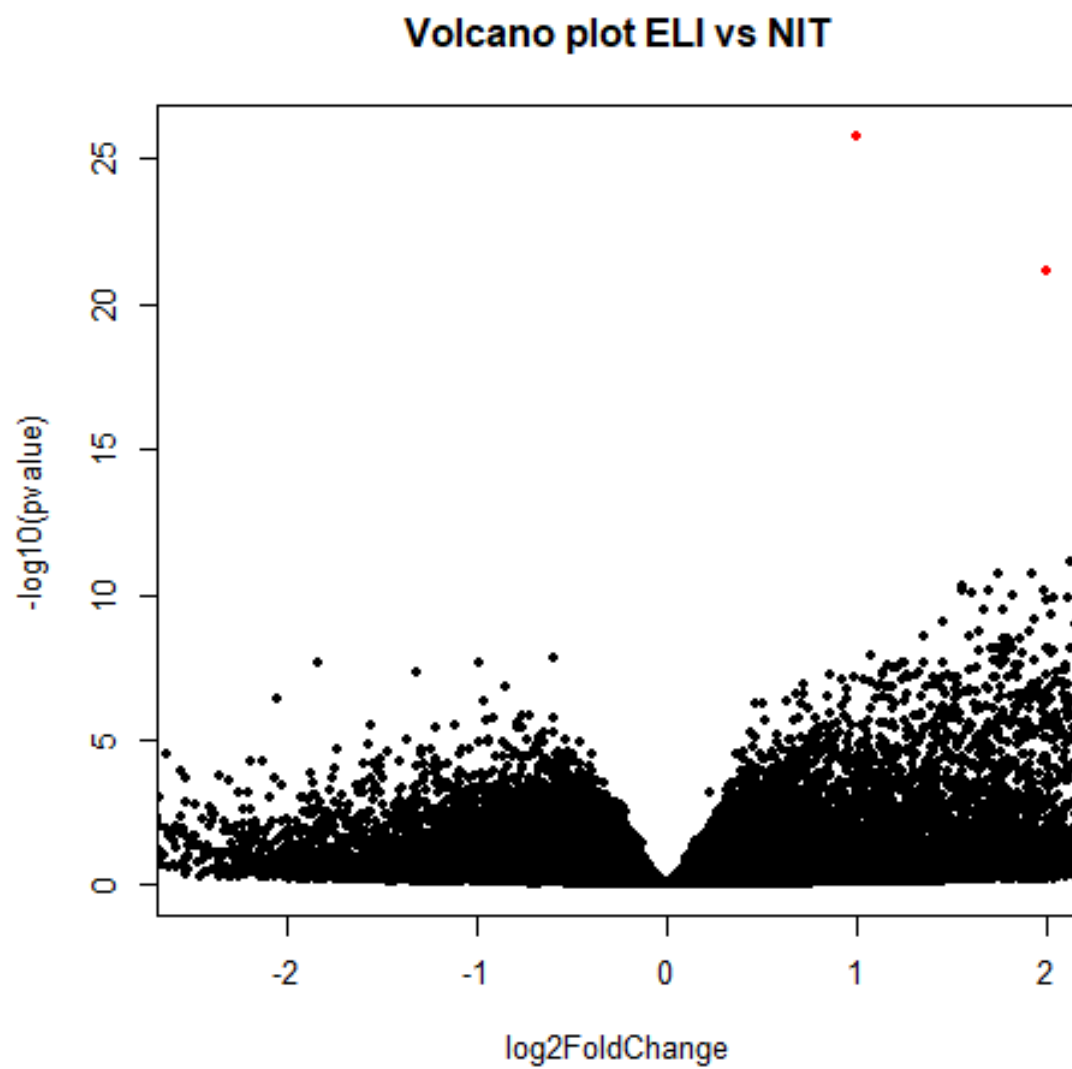


Figura 6: Volcanoplots para ELI vs NIT, ELI vs SFI y SFI vs NIT. Resaltados aparecen los 4 primeros genes

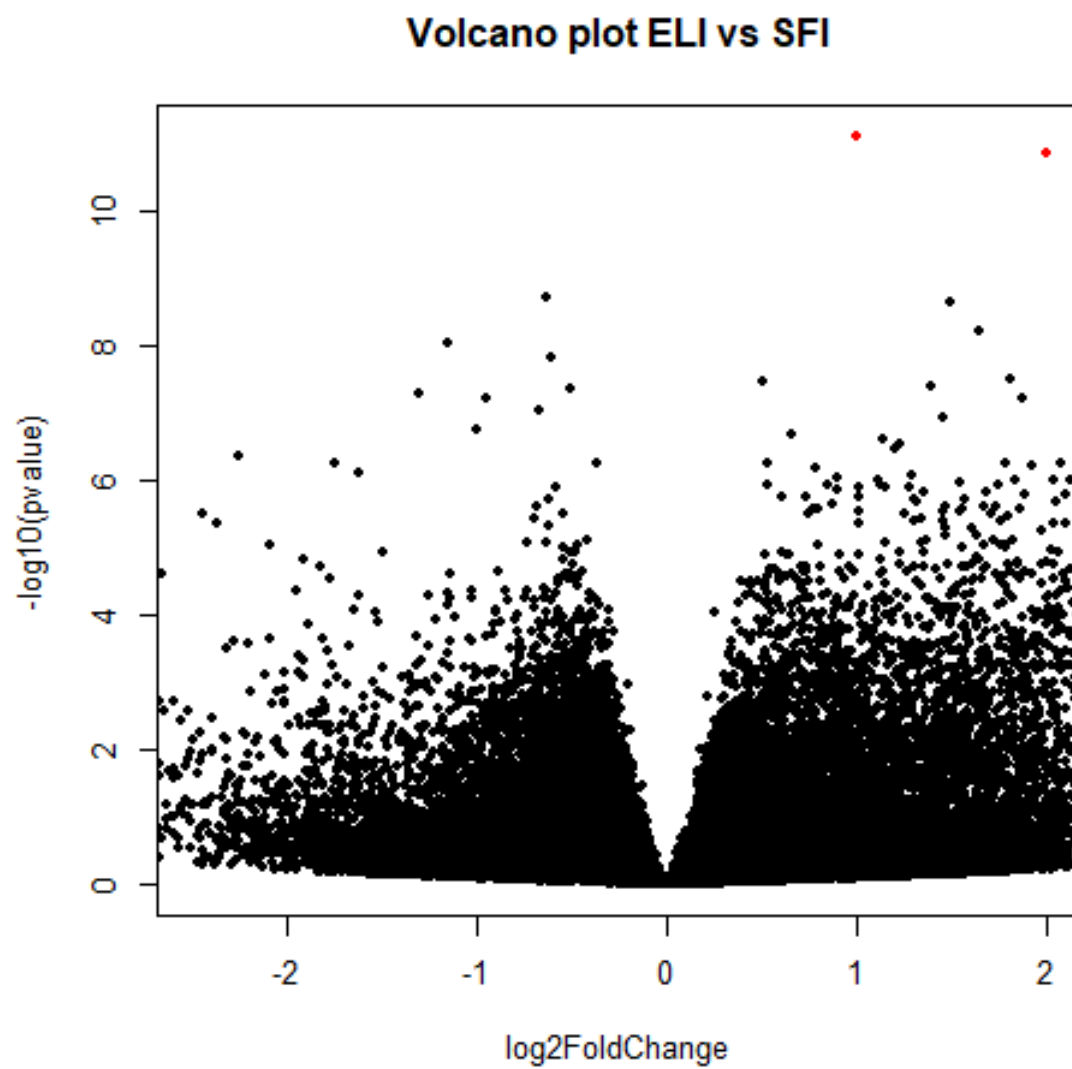


Figura 7: Volcanoplots para ELI vs NIT, ELI vs SFI y SFI vs NIT. Resaltados aparecen los 4 primeros genes

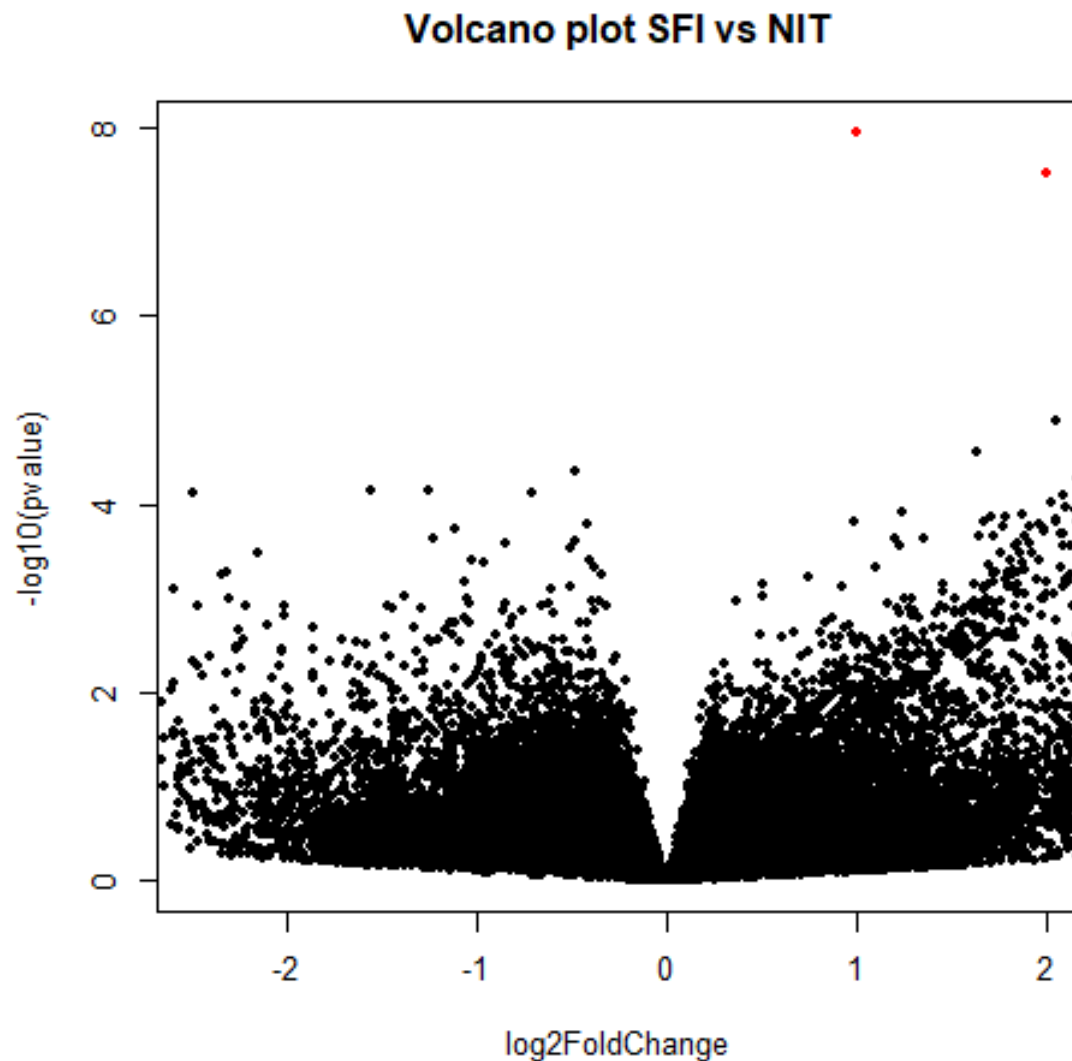


Figura 8: Volcanoplots para ELI vs NIT, ELI vs SFI y SFI vs NIT. Resaltados aparecen los 4 primeros genes

En la comparación SvsN intervienen una cantidad ligeramente mayor de genes “*down-regulated*”, mientras que en las comparaciones EvsN y EvsS ocurre al contrario.

6.0.1. Resultados del “Gene Enrichment Analysis”

Los genes seleccionados para este análisis en cada una de las comparaciones son:

EvsN	EvsS	SvsN
16224	17766	16772

Como la pregunta del estudio no está clara, sólo se muestran aquí los gráficos que a criterio técnico son más fácilmente interpretables. resultados del análisis realizado con enrichGO para procesos biológicos (Figuras 9 y 10):

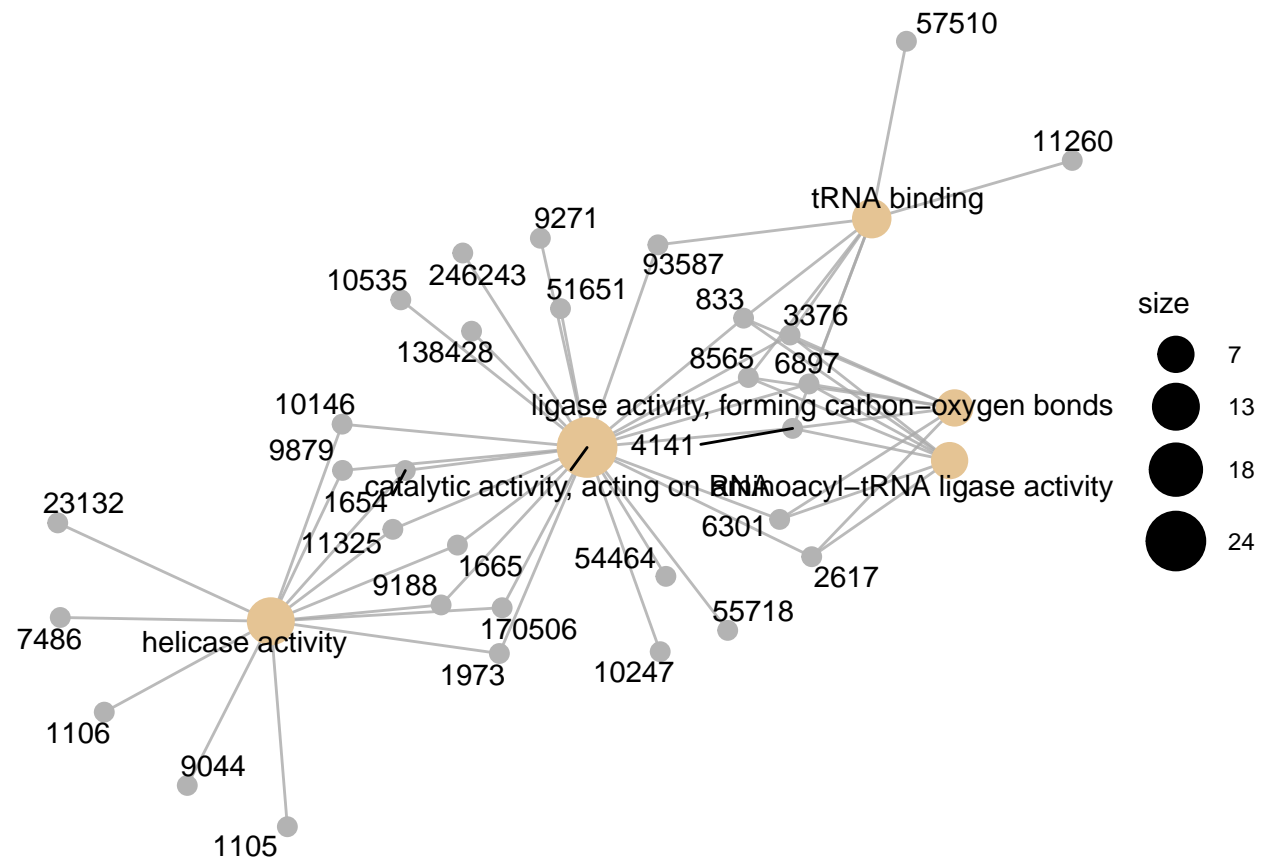


Figura 9: Network obtenida mediante el análisis enrichGO en la lista obtenida de la comparación entre UvsP1 y UvsP4

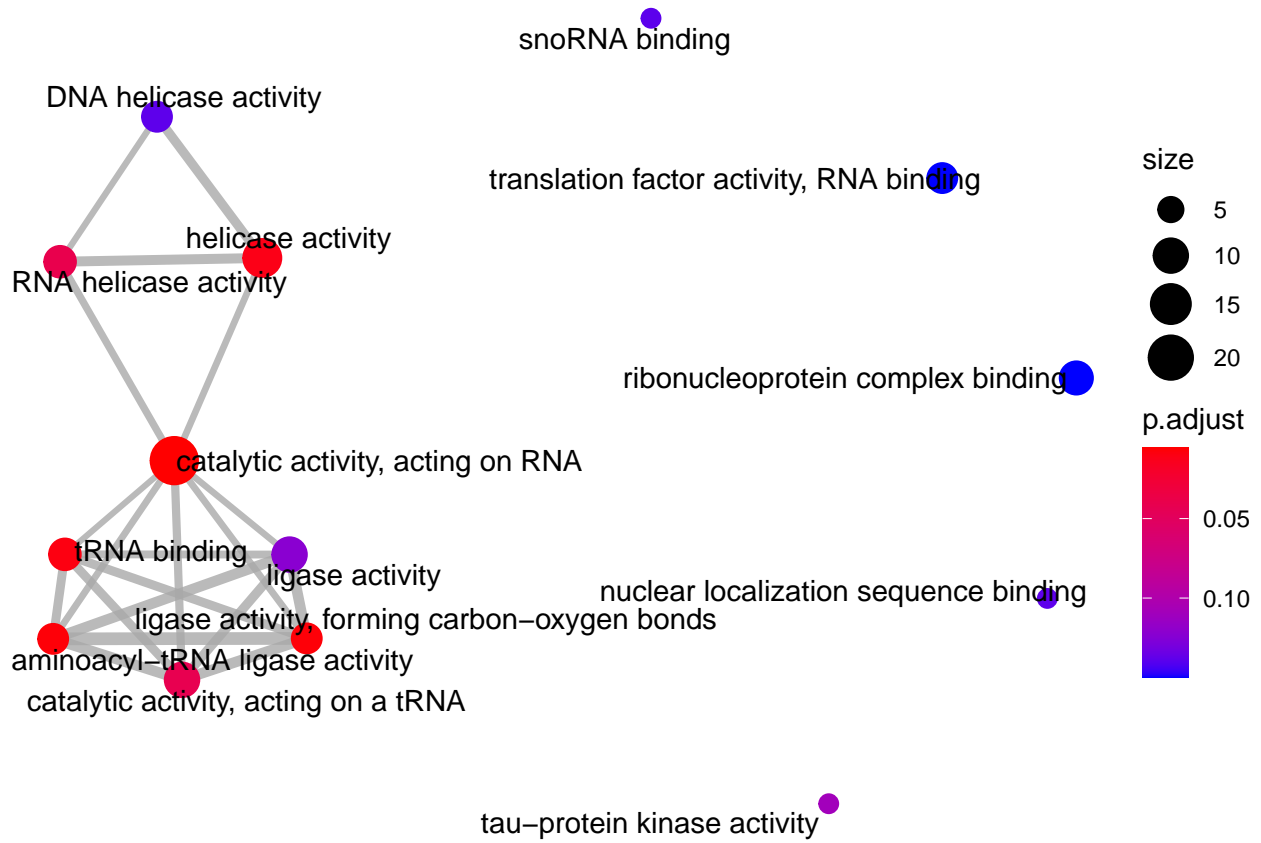


Figura 10: 'Enrichment Map' mediante el análisis enrichGO en la lista obtenida de la comparación entre UvsP1 y UvsP4

7. Sumario de resultados

[1] 938

[1] 873

[1] 7

Para la comparación EvsN, se han encontrado 938 procesos biológicos mediante el análisis de “enrichGO”, mientras que para EvsS se han encontrado 873 y para SvsN 7. Es interesante observar que, aunque la diferencia en el número de procesos biológicos señalados por el análisis es notable entre las comparaciones, en las primeras posiciones comparten la activación de las células T y la regulación de la activación de los linfocitos. En cambio, sólo la comparación SvsN presenta en las primeras posiciones la activación y proliferación de células B (Cuadros 1, 2 y 3).

package 'kableExtra' successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\Sonia D\AppData\Local\Temp\RtmpA1TM5M\downloaded_packages

Cuadro 1: Primeras filas y columnas de resultados de enrichGO para la comparación EvsN

	Description	GeneRatio	BgRatio	pvalue
GO:0042110	T cell activation	179/2192	464/18670	1.84599190457565e-51
GO:0030098	lymphocyte differentiation	142/2192	353/18670	2.15697820070576e-43
GO:0050863	regulation of T cell activation	128/2192	314/18670	6.54540611563134e-40
GO:0051249	regulation of lymphocyte activation	166/2192	485/18670	8.49192736454003e-40

Cuadro 2: Primeras filas y columnas de resultados de enrichGO para la comparación EvsS

	Description	GeneRatio	BgRatio	pvalue
GO:0042110	T cell activation	167/2315	464/18670	3.70248486971099e-40
GO:0030098	lymphocyte differentiation	130/2315	353/18670	1.40979681509864e-32
GO:0050863	regulation of T cell activation	119/2315	314/18670	2.96714105990843e-31
GO:1903039	positive regulation of leukocyte cell-cell adhesion	91/2315	218/18670	8.23597714700428e-28

Cuadro 3: Primeras filas y columnas de resultados de enrichGO para la comparación Svsn

	Description	GeneRatio	BgRatio	pvalue
GO:0042100	B cell proliferation	11/405	95/18670	6.66345227124652e-06
GO:0006520	cellular amino acid metabolic process	21/405	348/18670	2.61065803749542e-05
GO:0006418	tRNA aminoacylation for protein translation	7/405	45/18670	4.78165845618903e-05
GO:0042113	B cell activation	19/405	310/18670	5.1091771969831e-05

8. Comentarios

En el caso de que las infiltraciones NIT, SFI Y ELI fuesen pasos sucesivos dentro de un mismo proceso, los resultados parecen indicar que entre NIT y SFI se produce la proliferación de las células B, mientras que una vez llegados a un estado ELI hay activación de células T y diferenciación de linfocitos.

Se aconseja realizar una pregunta más concreta a la hora de solicitar el análisis de los datos. Ésto puede ayudar a elaborar un informe más preciso y relevante. Para una solicitud de hallar diferencias en comparaciones, sin más detalle, se hace difícil afinar sobre todo a la hora de definir umbrales de selección, y de decidir qué es importante y qué no para el estudio de referencia.

Bibliografía

Guangchuang, Yu. 2020. *Package clusterProfiler*. <http://bioconductor.riken.jp/packages/release/bioc/manuals/clusterProfiler/man/clusterProfiler.pdf>.

Love, Michael, Simon Anders, y Wolfgang Huber. 2020. *Analyzing RNA-seq data with DESeq2*. <http://bioconductor.riken.jp/packages/release/bioc/manuals/clusterProfiler/man/clusterProfiler.pdf>.

Love, Michael, Charlotte Soneson, Simon Anders, Vladislav Kim, y Wolfgang Huber. 2017. *RNA-seq workflow: gene-level exploratory analysis and differential expression*. https://www.bioconductor.org/help/course-materials/2017/CSAMA/labs/2-tuesday/lab-03-rnaseq/rnaseqGene_CSAMA2017.html#experimental-data.

Turner, Stephen. 2014. *Using Volcano Plots in R to Visualize Microarray and RNA-seq Results*. <https://www.r-bloggers.com/using-volcano-plots-in-r-to-visualize-microarray-and-rna-seq-results/>.