Análisis de datos ómicos

PEC 2

Análisis de datos de ultrasecuenciación

José Barreiro González
UOC | https://github.com/jbarreiro20uoc/ado_pec2_2020.git

Índice

Contenido

1. ABSTRACT

Análisis de muestras de tejidos del tiroides con diferentes niveles de infiltración, mediante estudio de datos de expresión (RNA-seq) utilizando RStudio/Bioconductor. Se han realizado tres comparaciones de grupos dos a dos, partiendo de una tabla de conteo y un archivo de metadatos.

2. OBJETIVOS

El objetivo del estudio consiste en un análisis de expresión diferencial de genes, buscando patrones de expresión y genes diferencialmente expresados en las muestras de tejidos. Se buscarán entre los resultados de las tres comparaciones patrones utilizando diferentes gráficos, y se crearán tres tablas de resultados de genes diferencialmente expresados.

3. MATERIALES Y MÉTODOS

Datos

Con el estudio se realiza un análisis de varias muestras de tejidos del tiroides, donde se compararán tres tipos de infiltración de un total de 292 muestras, pertenecientes a tres grupos:

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

Se extraerán 10 muestras aleatorias de cada grupo.

Métodos

Se compararán tres grupos, SFI-NIT, ELI-NIT, ELI-SFI, con diferentes niveles de infiltración. El análisis se centra en los datos de expresión (RNA-seq)

1) El procedimiento de análisis consiste en un "pipeline" que consta de los siguientes pasos:

    a. Definición de los datos.

    b. Preprocesado de los datos: filtraje y normalización.

    c. Identificación de genes diferencialmente expresados.

    d. Anotación de los resultados.

    e. Busca de patrones de expresión y agrupación de las muestras (comparación entre las distintas comparaciones)

    f. Análisis de significación biológica ("Gene Enrichment Analysis")

2) Para el estudio se ha utilizado el software RStudio, programa open source, en su versión 1.3.959. Se ha optado por usar el programa R/Bioconductor, con el paquete DESeq2.

3) El análisis RNA-seq consiste en los siguientes pasos:

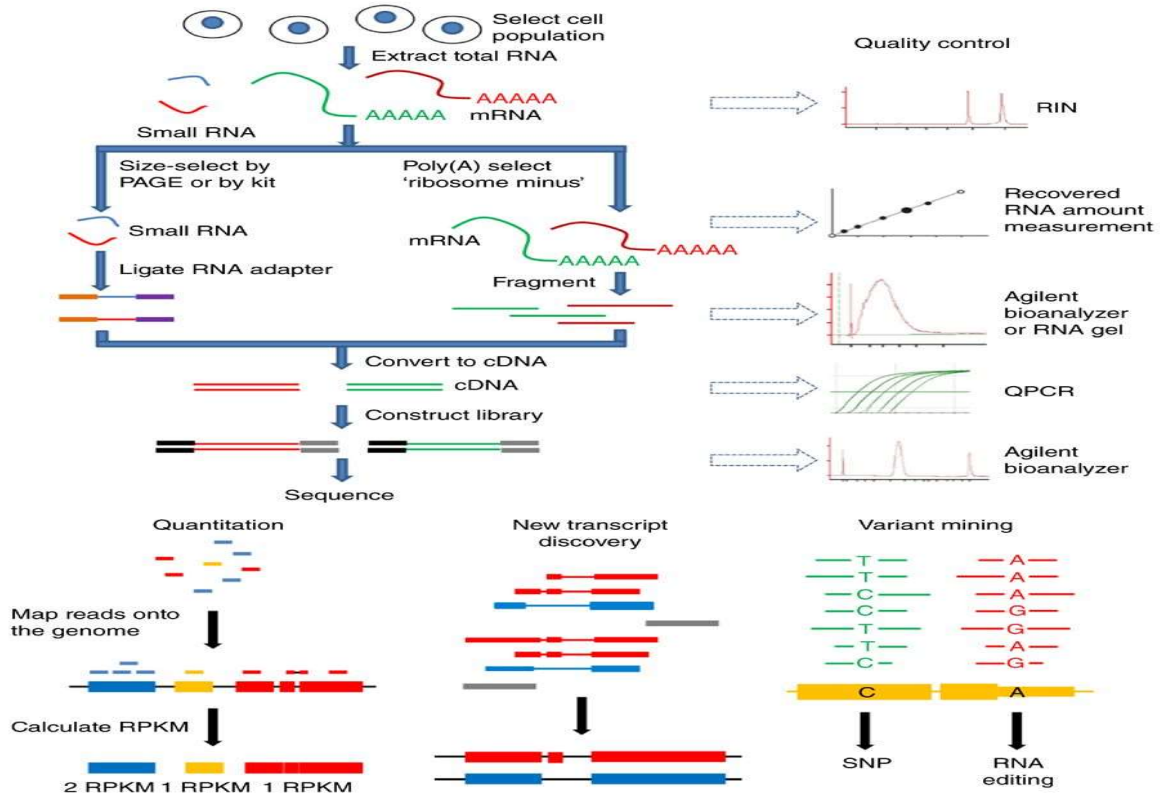
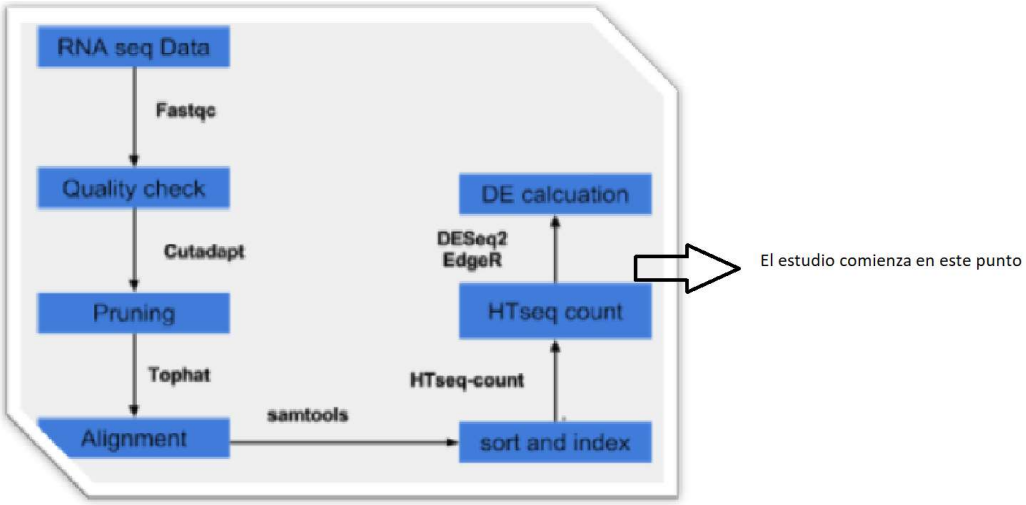


Ilustración del proceso

El análisis que se realiza a continuación se centra en la parte bioinformática, ya que partimos de tablas de contajes.

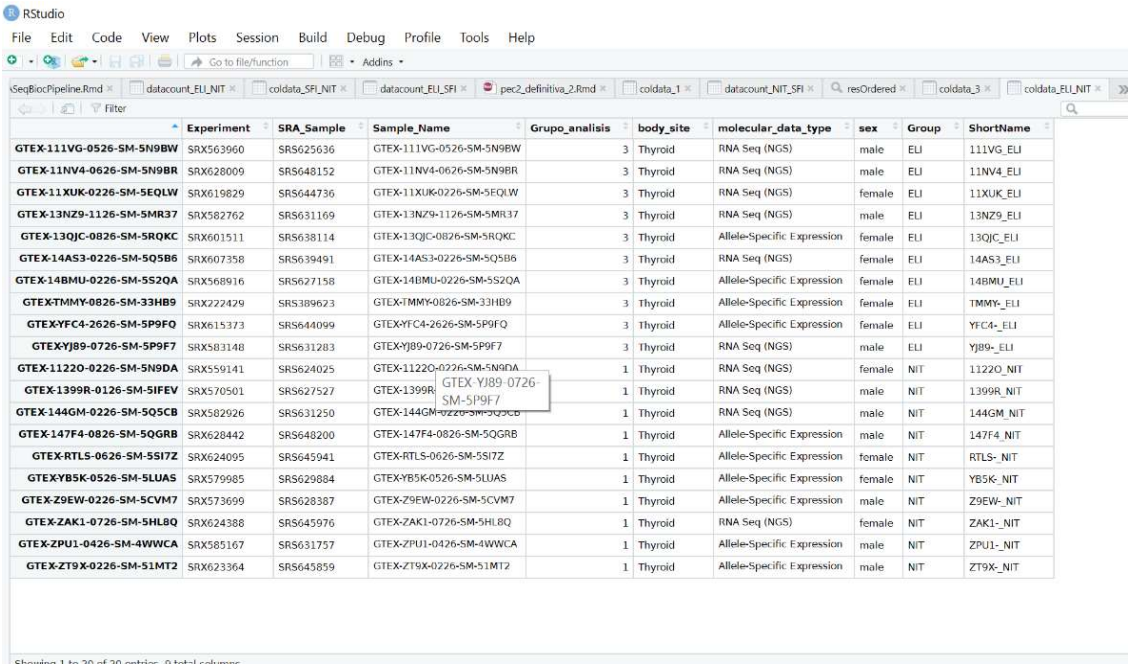


Proceso

Se muestra en detalle la primera comparación, de las otras dos se mostrarán los resultados para no ser repetitivos y no alargar el documento innecesariamente, ya que el proceso es el mismo para las tres.

En primer lugar, debemos modificar y adaptar los datos a lo pedido en el ejercicio.

Cargamos los archivos y los modificamos, filtrándolos y seleccionando 10 muestras aleatorias de cada grupo.


RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

	Experiment	SRA_Sample	Sample_Name	Grupo_analisis	body_site	molecular_data_type	sex	Group	ShortName
GTEX-111VG-0526-SM-5N9BW	SRX563960	SRS625636	GTEX-111VG-0526-SM-5N9BW	3	Thyroid	RNA Seq (NGS)	male	ELI	111VG_ELI
GTEX-11NV4-0626-SM-5N9BR	SRX628009	SRS648152	GTEX-11NV4-0626-SM-5N9BR	3	Thyroid	RNA Seq (NGS)	male	ELI	11NV4_ELI
GTEX-11XUK-0226-SM-5EQLW	SRX619829	SRS644736	GTEX-11XUK-0226-SM-5EQLW	3	Thyroid	RNA Seq (NGS)	female	ELI	11XUK_ELI
GTEX-13NZ9-1126-SM-5MR37	SRX582762	SRS631169	GTEX-13NZ9-1126-SM-5MR37	3	Thyroid	RNA Seq (NGS)	male	ELI	13NZ9_ELI
GTEX-13QJC-0826-SM-5RQKC	SRX601511	SRS638114	GTEX-13QJC-0826-SM-5RQKC	3	Thyroid	Allele-Specific Expression	female	ELI	13QJC_ELI
GTEX-14AS3-0226-SM-5Q5B6	SRX607358	SRS639491	GTEX-14AS3-0226-SM-5Q5B6	3	Thyroid	RNA Seq (NGS)	female	ELI	14AS3_ELI
GTEX-14BMU-0226-SM-5S2QA	SRX568916	SRS627158	GTEX-14BMU-0226-SM-5S2QA	3	Thyroid	Allele-Specific Expression	female	ELI	14BMU_ELI
GTEX-TMMY-0826-SM-33HB9	SRX222429	SRS389623	GTEX-TMMY-0826-SM-33HB9	3	Thyroid	Allele-Specific Expression	female	ELI	TMMY-_ELI
GTEX-YFC4-2626-SM-5P9FQ	SRX615373	SRS644099	GTEX-YFC4-2626-SM-5P9FQ	3	Thyroid	Allele-Specific Expression	female	ELI	YFC4-_ELI
GTEX-YJ89-0726-SM-5P9F7	SRX583148	SRS631283	GTEX-YJ89-0726-SM-5P9F7	3	Thyroid	RNA Seq (NGS)	male	ELI	YJ89-_ELI
GTEX-1122O-0226-SM-5N9DA	SRX559141	SRS624025	GTEX-1122O-0226-SM-5N9DA	1	Thyroid	RNA Seq (NGS)	female	NIT	1122O_NIT
GTEX-1399R-0126-SM-5IFEV	SRX570501	SRS627527	GTEX-1399R	1	Thyroid	RNA Seq (NGS)	male	NIT	1399R_NIT
GTEX-144GM-0226-SM-5Q5CB	SRX582926	SRS631250	GTEX-144GM-0226-SM-5Q5CB	1	Thyroid	RNA Seq (NGS)	male	NIT	144GM_NIT
GTEX-147F4-0826-SM-5QGRB	SRX628442	SRS648200	GTEX-147F4-0826-SM-5QGRB	1	Thyroid	Allele-Specific Expression	male	NIT	147F4_NIT
GTEX-RTLS-0626-SM-5SI7Z	SRX624095	SRS645941	GTEX-RTLS-0626-SM-5SI7Z	1	Thyroid	Allele-Specific Expression	female	NIT	RTLS-_NIT
GTEX-YB5K-0526-SM-5LUAS	SRX579985	SRS629884	GTEX-YB5K-0526-SM-5LUAS	1	Thyroid	Allele-Specific Expression	female	NIT	YB5K-_NIT
GTEX-Z9EW-0226-SM-5CVM7	SRX573699	SRS628387	GTEX-Z9EW-0226-SM-5CVM7	1	Thyroid	Allele-Specific Expression	male	NIT	Z9EW-_NIT
GTEX-ZAK1-0726-SM-5HL8Q	SRX624388	SRS645976	GTEX-ZAK1-0726-SM-5HL8Q	1	Thyroid	RNA Seq (NGS)	female	NIT	ZAK1-_NIT
GTEX-ZPU1-0426-SM-4WWCA	SRX585167	SRS631757	GTEX-ZPU1-0426-SM-4WWCA	1	Thyroid	Allele-Specific Expression	male	NIT	ZPU1-_NIT
GTEX-ZT9X-0226-SM-51MT2	SRX623364	SRS645859	GTEX-ZT9X-0226-SM-51MT2	1	Thyroid	Allele-Specific Expression	male	NIT	ZT9X-_NIT

Se muestra uno de los archivos countdata obtenidos para la comparación ELI-NIT

Una vez obtenidos los archivos se procede al análisis de los datos.

Como se explicó más arriba se ha usado el paquete de R/Bioconductor “DESeq2”.

Visualización

El primer paso es crear un dataset con los datos para comenzar el análisis estadístico, para ello se ha usado la función “DESeqDataSetFromMatrix “. Se obtiene:

```
class: DESeqDataSet
dim: 6 20
metadata(1): version
assays(6): counts mu ... replaceCounts replaceCooks
rownames(6): ENSG00000223972 ENSG00000227232 ... ENSG00000268020
  ENSG00000240361
rowData names(23): baseMean baseVar ... maxCooks replace
colnames(20): GTEX-11O72-2326-SM-5BC7H GTEX-11TUW-0226-SM-5LU8X ...
  GTEX-ZPU1-0426-SM-4WWCA GTEX-ZT9X-0226-SM-51MT2
colData names(11): Experiment SRA_Sample ... sizeFactor replaceable
```

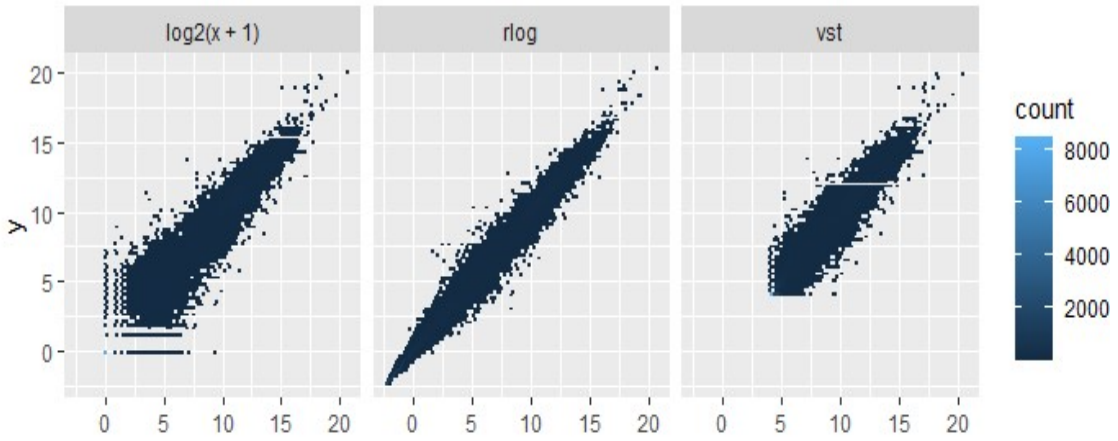
A continuación, prefiltraremos y normalizaremos los datos, eliminando los valores 0 y 1, para reducir el tamaño del archivo y agilizar los cálculos.

De los transcritos iniciales nos quedamos con:

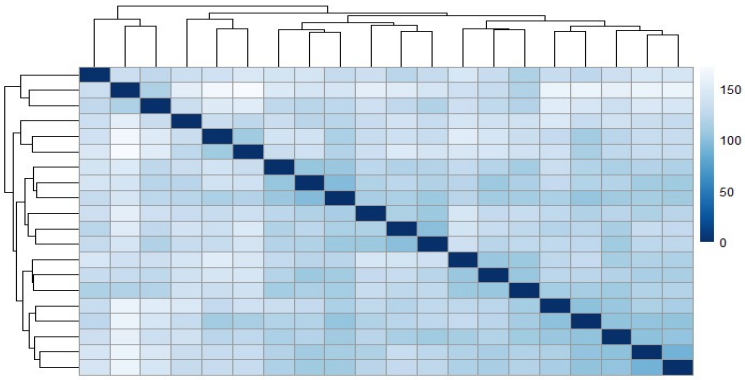
[1] 56202

[1] 41436

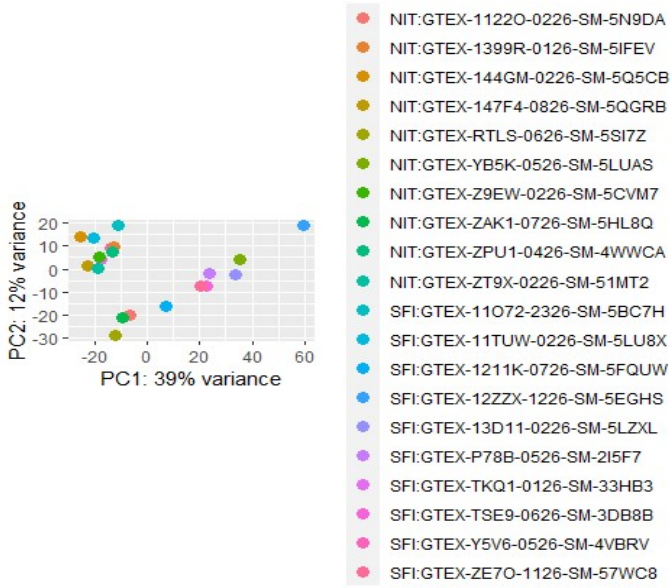
El siguiente paso será, en primer lugar, realizar varios gráficos para una primera visualización de los datos; y, en segundo lugar, realizar el análisis de expresión diferencial mediante paquetes estadísticos.

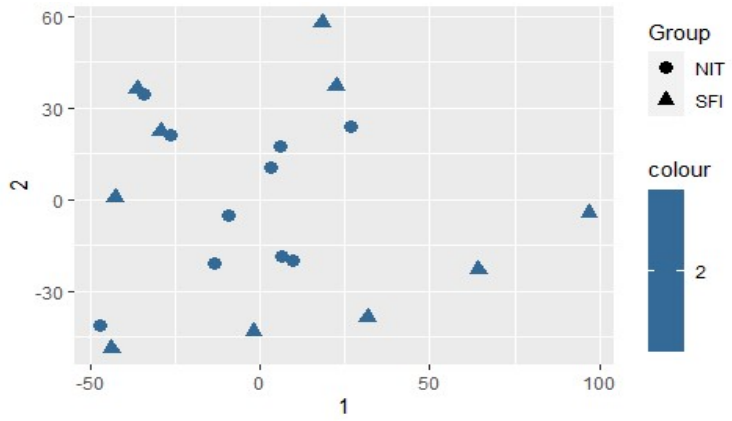


Heatmap



Plot-PCA





Expresión diferencial de genes

Se ha utilizado la función DESeq para este paso. Se obtiene:

	row <chr>	baseMean <dbl>	log2FoldChange <dbl>	lfcSE <dbl>	stat <dbl>	pvalue <dbl>	padj <dbl>
1	ENSG00000223972	2.7623352	0.2580906	0.6140587	0.42030280	0.6742643	0.999794
2	ENSG00000227232	774.2449210	-0.1945517	0.2087884	-0.93181272	0.3514333	0.999794
3	ENSG00000243485	2.0531693	0.8789200	0.7065835	1.24390111	0.2135360	0.999794
4	ENSG00000237613	1.7962436	0.6875734	0.7256347	0.94754758	0.3433598	0.999794
5	ENSG00000268020	0.5826471	-0.1015649	1.1217965	-0.09053773	0.9278599	*NA*
6	ENSG00000240361	1.1449497	-0.0874194	0.8352989	-0.10465643	0.9166484	0.999794

6 rows

Ordenamos en función del p-valor.

Podemos seguir dos criterios diferentes más restrictivos a la hora de considerar genes significativamente diferenciados.

1) Bajando el umbral de la tasa de detección falsa:

```
FALSE  TRUE
28500    82
```

2) Si queremos elevar el umbral del log2:

```
FALSE  TRUE
21336    16
```

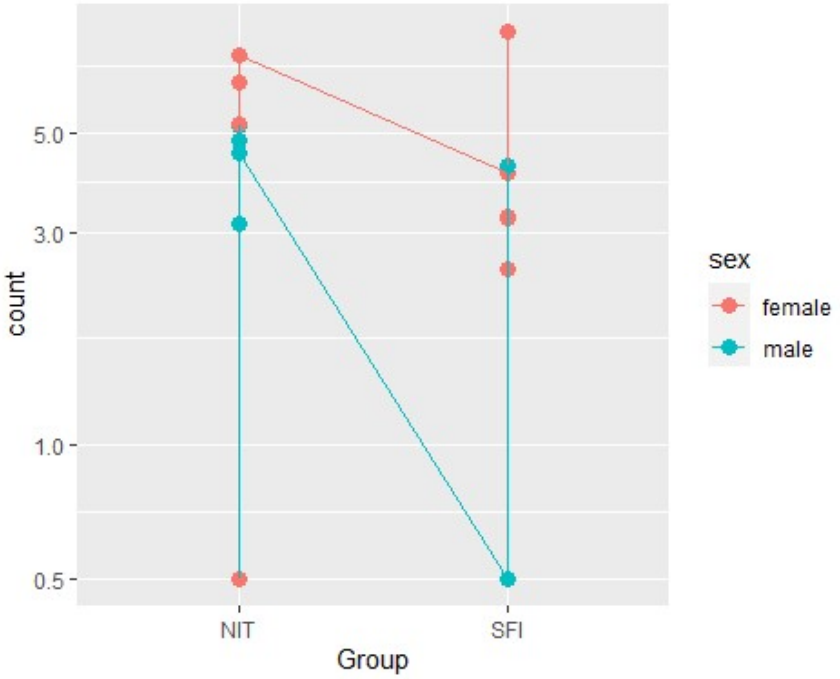
Por lo tanto, si consideramos que una fracción de 10% de falsos positivos es aceptable, podemos considerar que todos los genes con un valor p ajustado por debajo del 10% a 0,1 son significativos. ¿Cuántos genes de este tipo hay?

```
[1] 107
```

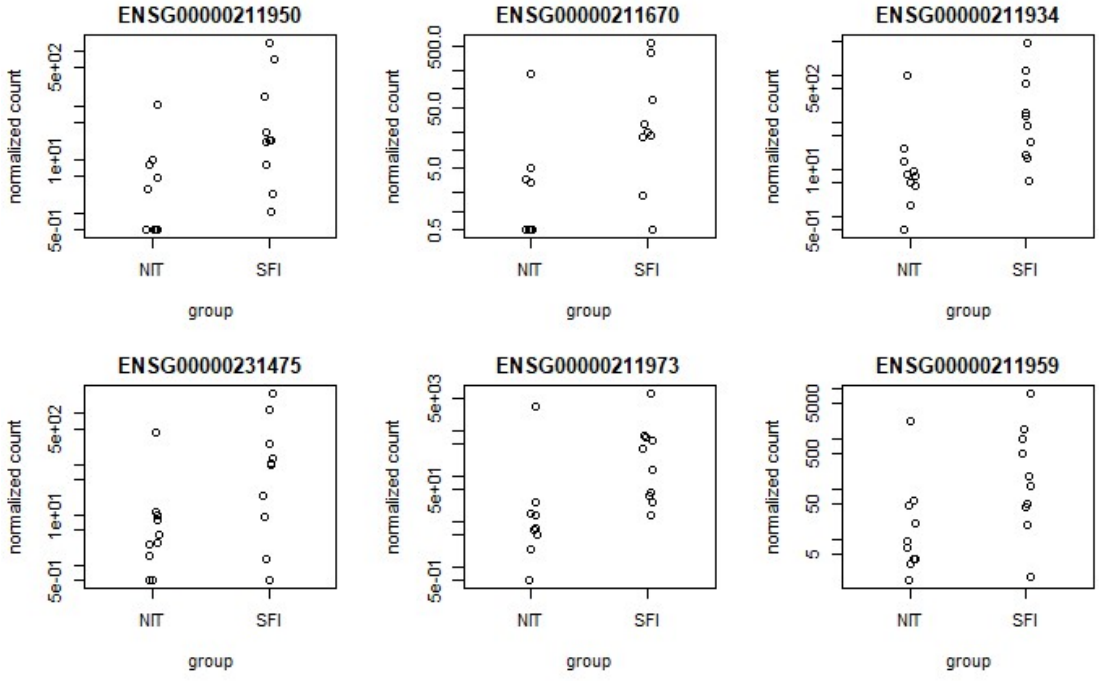
Colocamos la tabla de resultados en estos genes y luego la ordenamos por la estimación de cambio de pliegue log2 para obtener los genes significativos con la regulación descendente más fuerte:

```
log2 fold change (MLE): Group SFI vs NIT
Wald test p-value: Group SFI vs NIT
DataFrame with 6 rows and 6 columns
                 baseMean log2FoldChange     lfcSE      stat      pvalue        padj
                <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric>
ENSG00000211950  125.6517        6.15338  1.216330   5.05897 4.21528e-07 6.05957e-04
ENSG00000211670   55.5411        5.97609  1.358290   4.39971 1.08393e-05 6.67792e-03
ENSG00000211934  366.1685        5.96189  0.976632   6.10454 1.03097e-09 3.11228e-05
ENSG00000231475  220.4259        5.94893  1.125562   5.28529 1.25503e-07 2.58803e-04
ENSG00000211973  489.0421        5.67600  1.040012   5.45763 4.82536e-08 1.61853e-04
ENSG00000211959  578.0581        5.57921  1.011176   5.51754 3.43772e-08 1.48254e-04
```

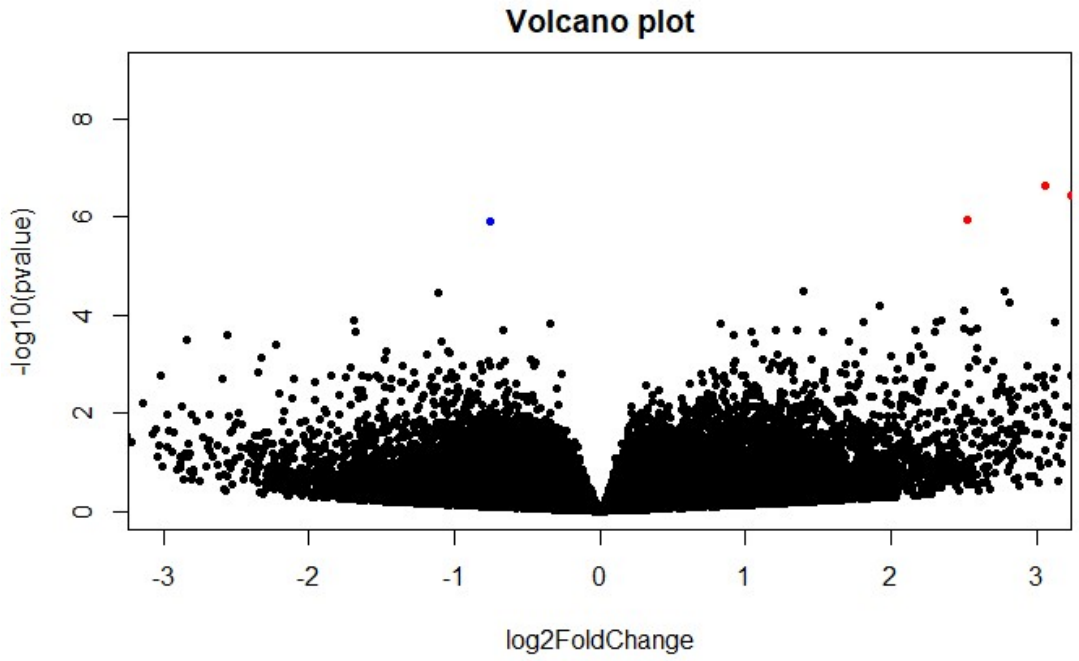
Visualizamos los resultados, añadiendo la variable sexo, para identificar posibles relaciones.



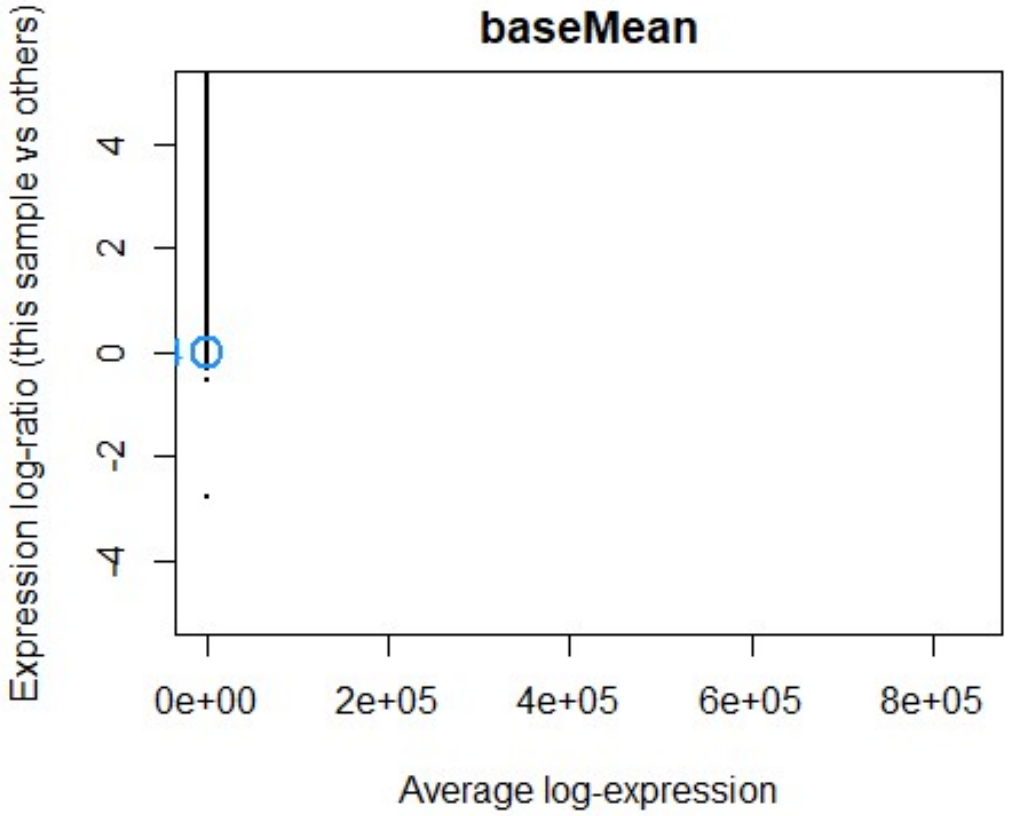
Vemos los “plot-counts”



Volcano plot

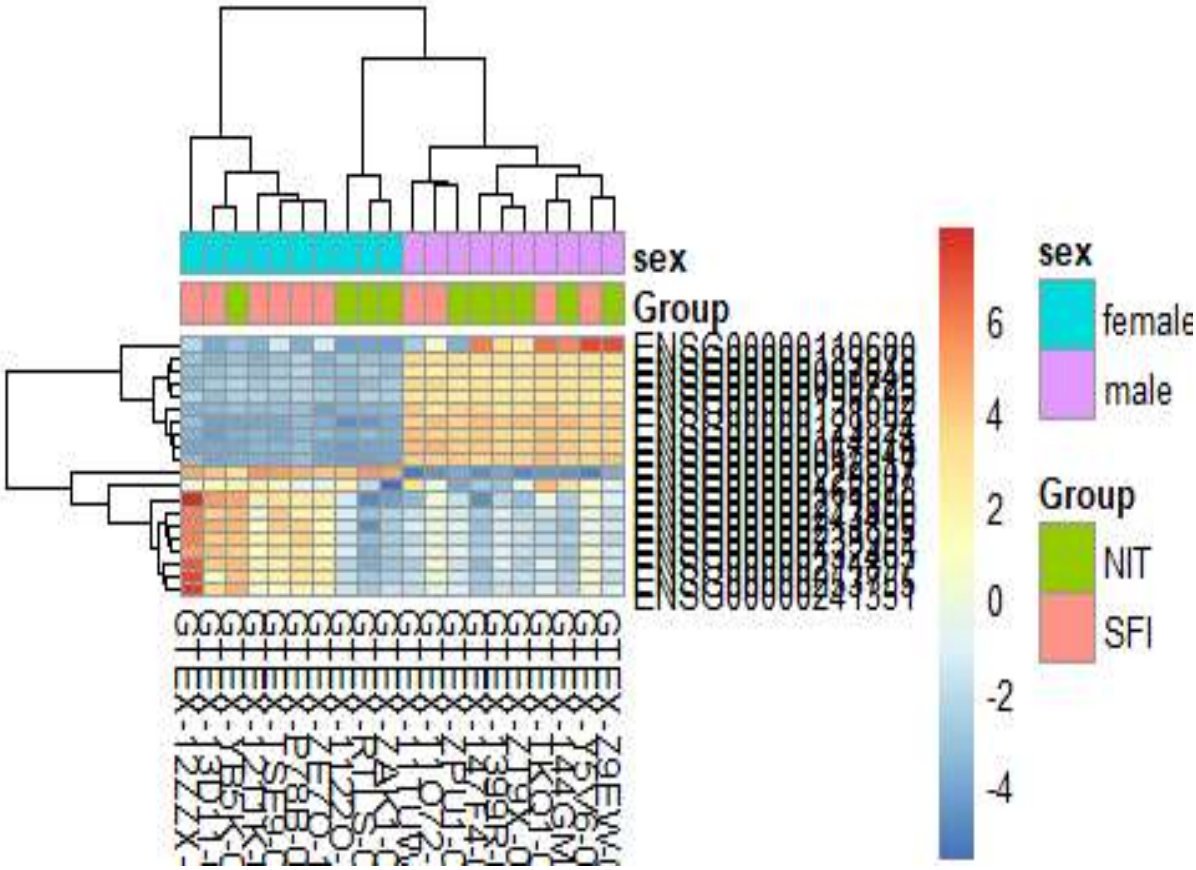


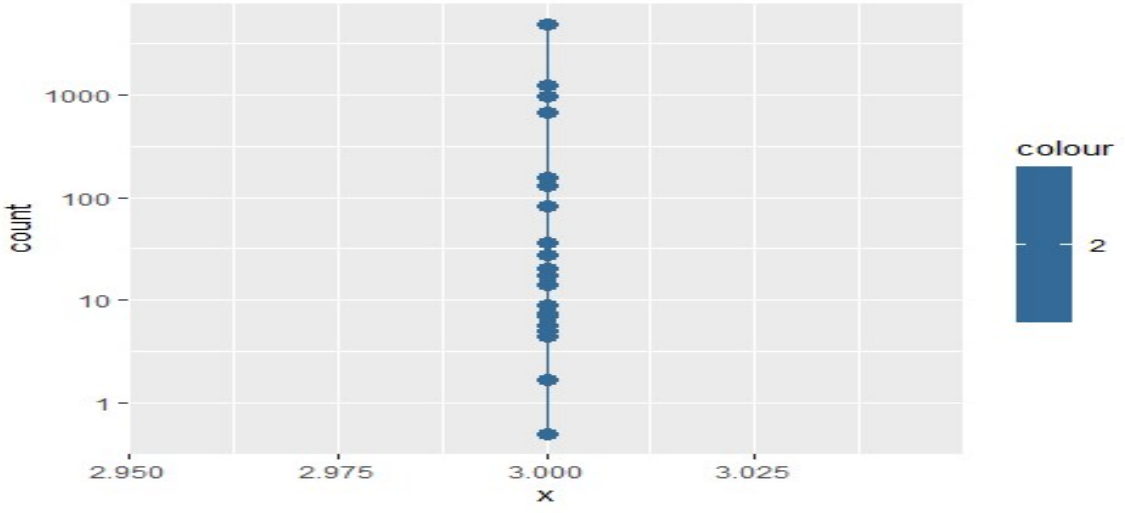
MA-plot



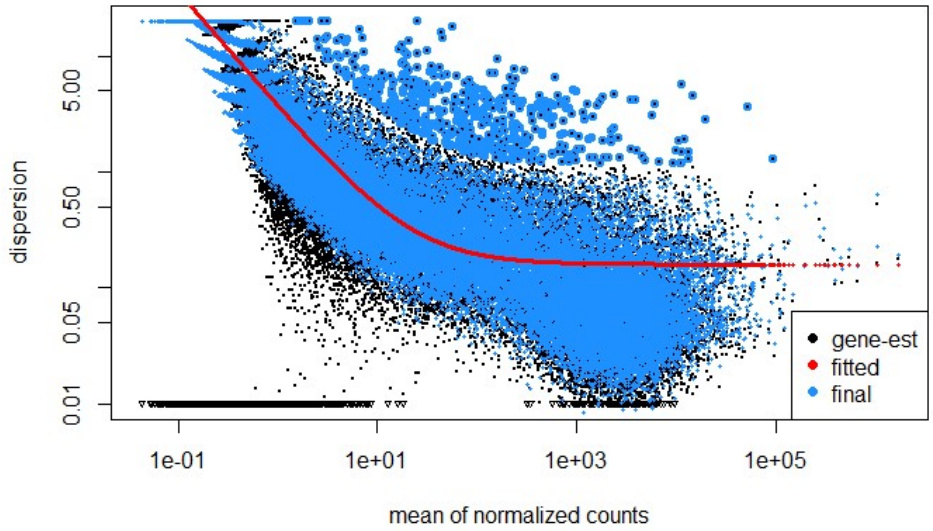
Gene clustering

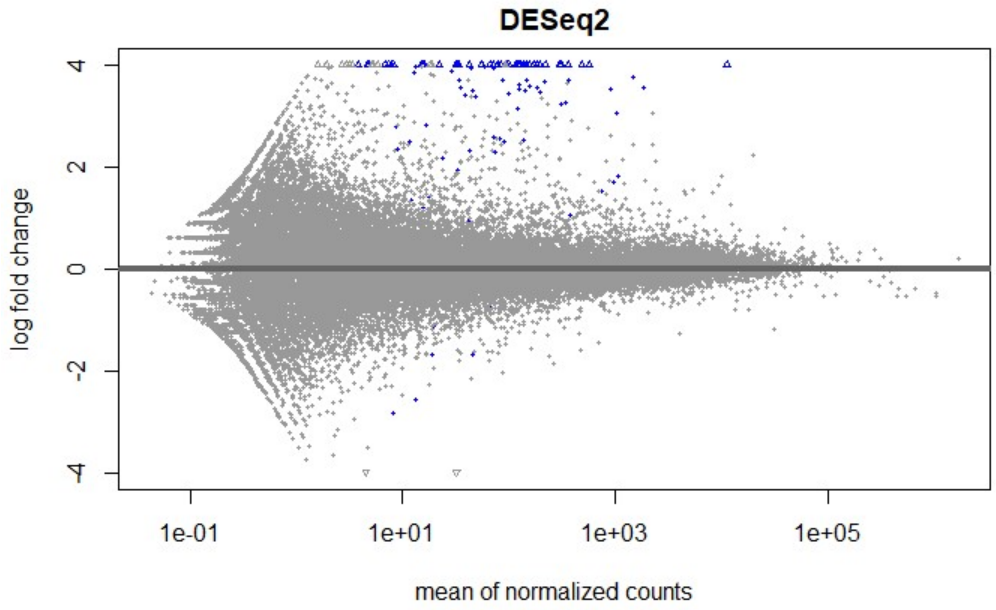
Visualizamos un clúster de los genes de más variación. Usaremos la transformación de estabilización de la varianza(vst), que integra el paquete DESeq2 y permite mejor tratamiento para los datos que generalmente tienen el mismo rango de varianza en diferentes rangos de los valores medios. Podemos añadir la variable "sex" para ver la relación.

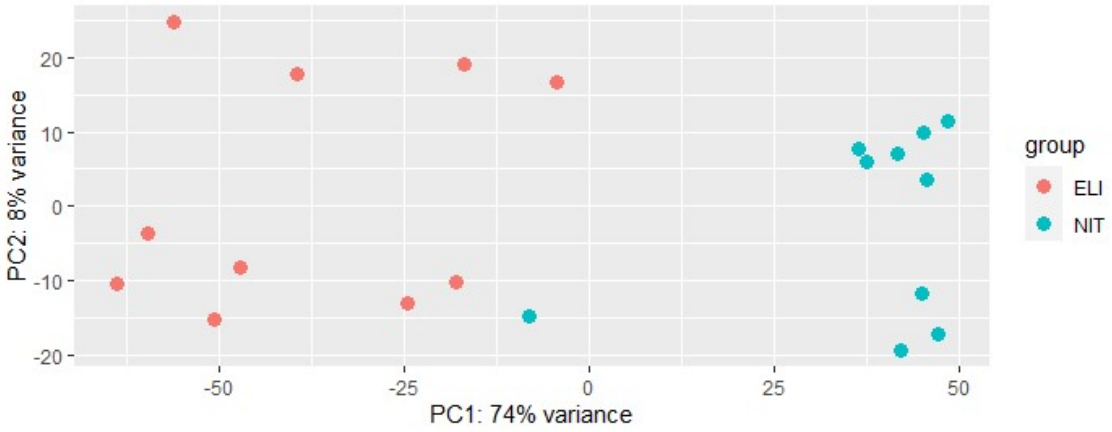




Estimamos los parámetros de dispersión:







Anotaciones

Se preparan los paquetes necesarios y se instalan. En este caso:

library("AnnotationDbi")

library(org.Hs.eg.db)

```
log2 fold change (MAP): Group SFI vs NIT
Wald test p-value: Group SFI vs NIT
DataFrame with 6 rows and 7 columns
                  baseMean log2FoldChange      lfcSE      pvalue        padj      symbol
entrez
                 <numeric>      <numeric>  <numeric>   <numeric>   <numeric> <character>
<character>
ENSG00000211934   366.169    2.41682e-06 0.00144270 1.03097e-09 3.11228e-05          NA
NA
ENSG00000240041   136.387    3.08080e-06 0.00144270 5.21696e-09 7.87448e-05          NA
NA
ENSG00000211663   313.320    3.17248e-06 0.00144270 9.61120e-09 9.67143e-05          NA
NA
ENSG00000211966   303.101    1.76164e-06 0.00144269 1.84026e-08 1.24882e-04          NA
NA
ENSG00000241755   144.267    3.16994e-08 0.00144269 2.06840e-08 1.24882e-04          NA
NA
ENSG00000211900   100.618    2.23248e-06 0.00144269 3.02414e-08 1.48254e-04          NA
NA
```

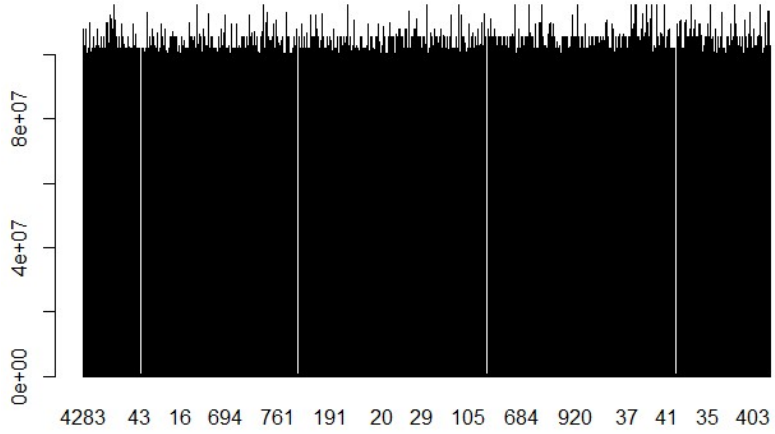
En este paso se han descargado ficheros de anotaciones del genoma humano para asociar los resultados de transcritos diferencialmente expresados con el gen asociado correspondiente. Se añaden dos columnas, “symbol” y “entrez”.

Observamos que en algunos casos obtenemos valores NA, no disponibles.

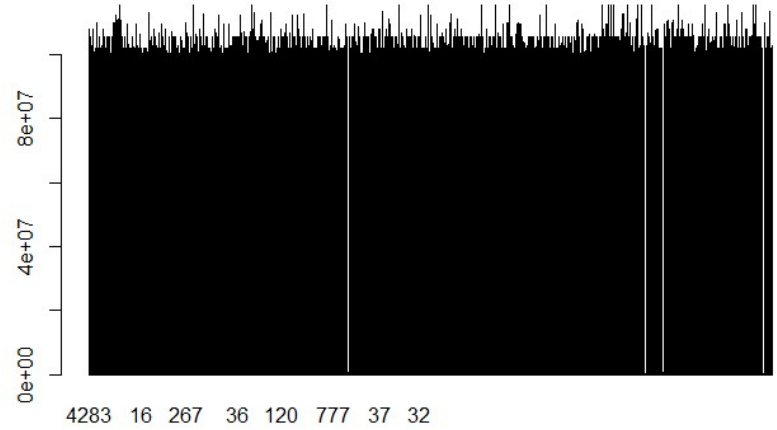
Se exportan tres ficheros txt con los resultados de las anotaciones asociados a los transcritos diferencialmente expresados.

Análisis de significación biológica.

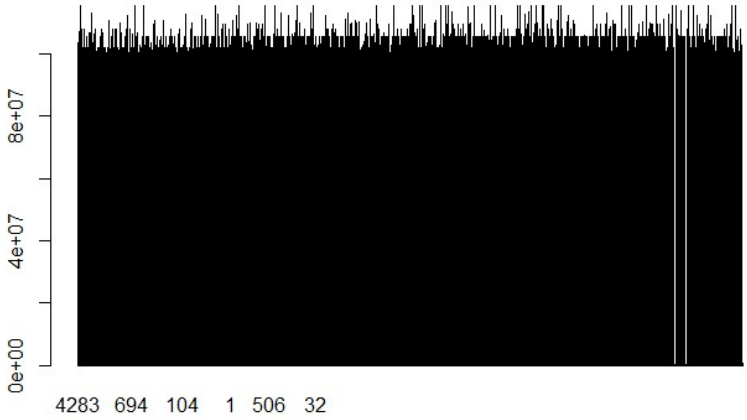
Gene Enrichment Analysis.



1.SFI-NIT



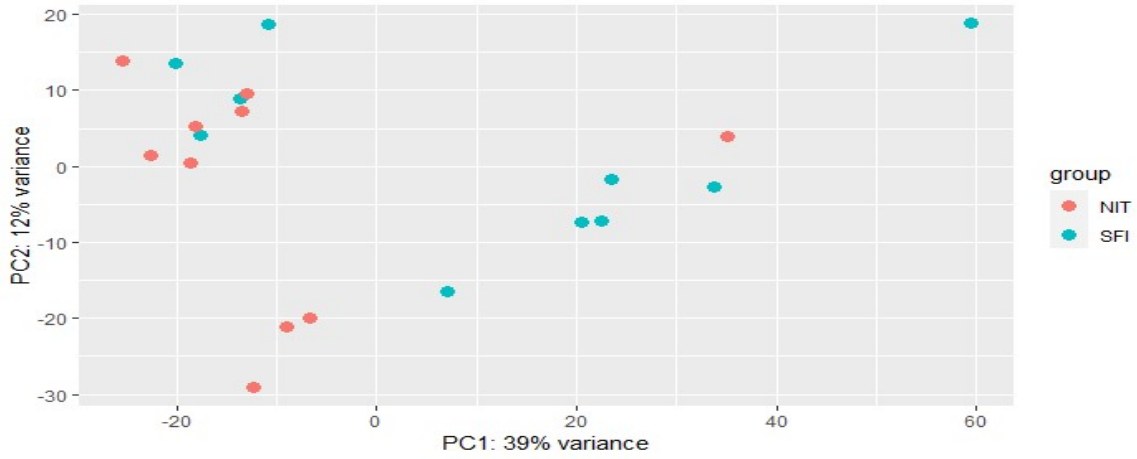
2.ELI-NIT

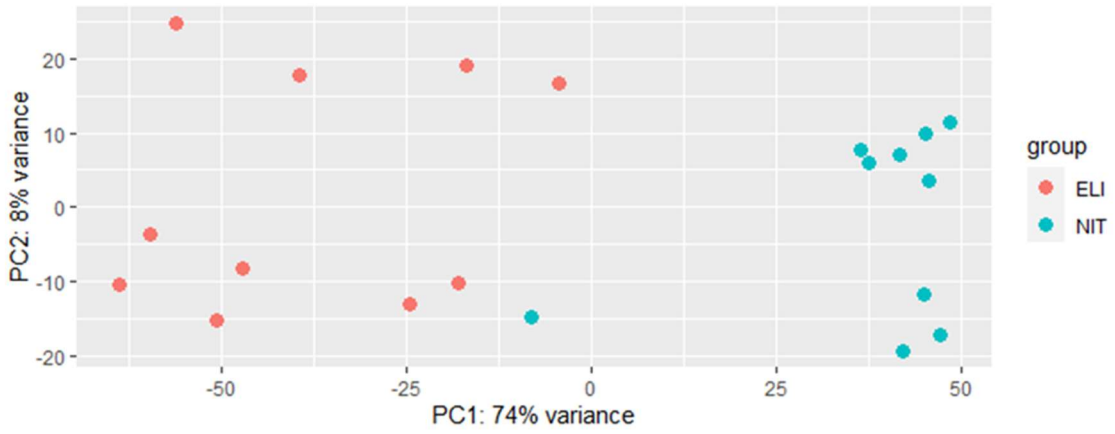


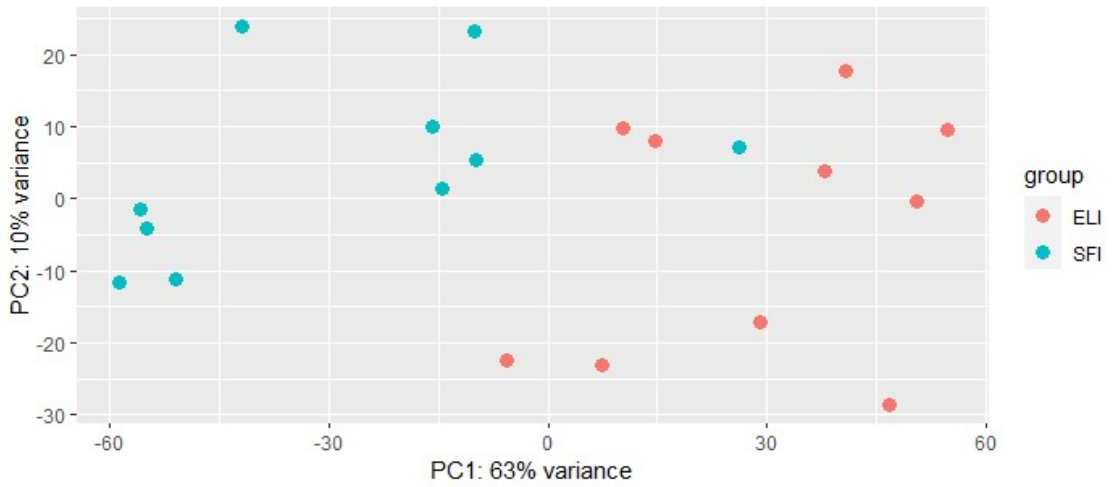
3.ELI-SFI

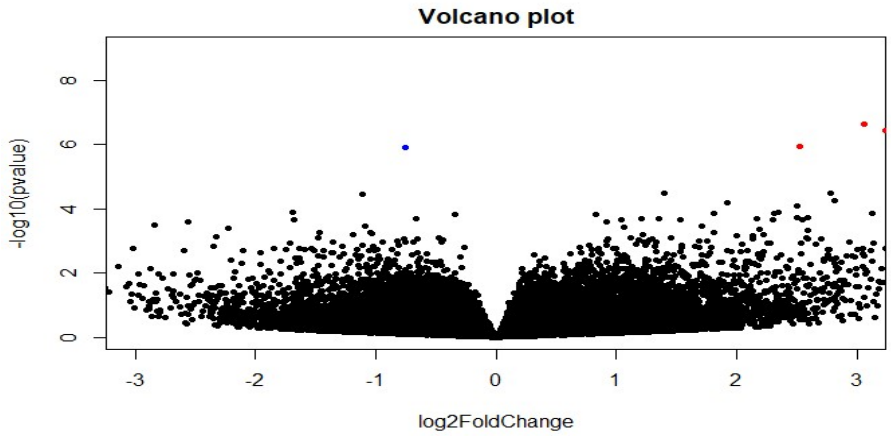
4. RESULTADOS

Se comparan los resultados gráficamente para ver las relaciones:

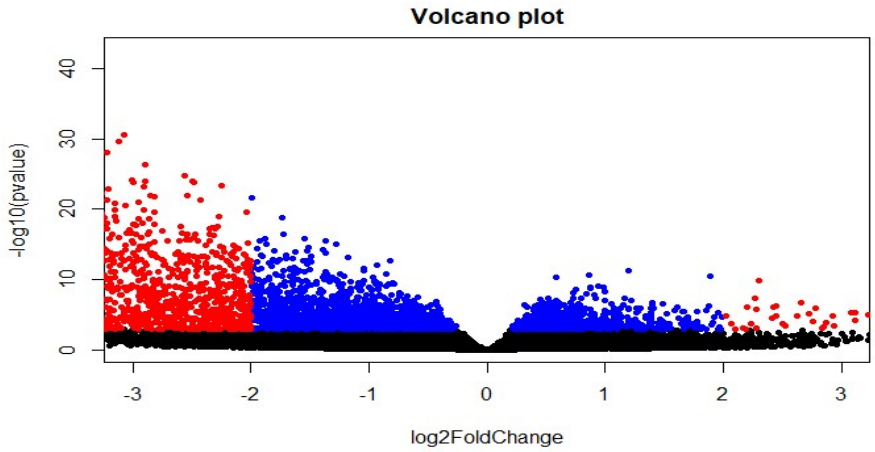




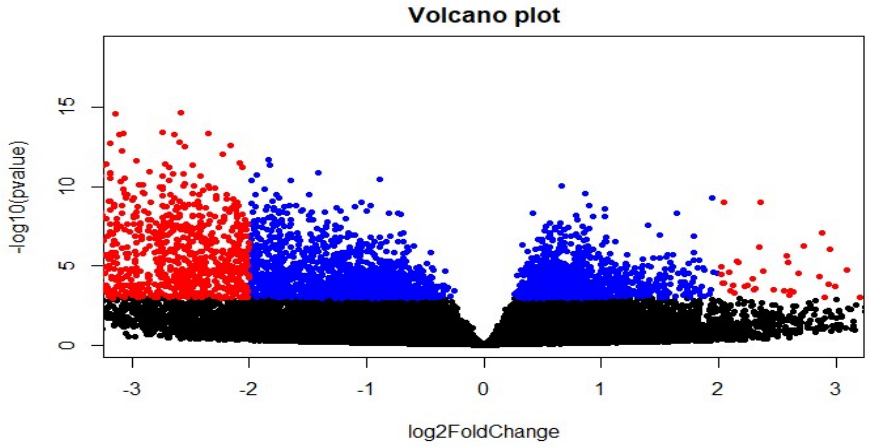




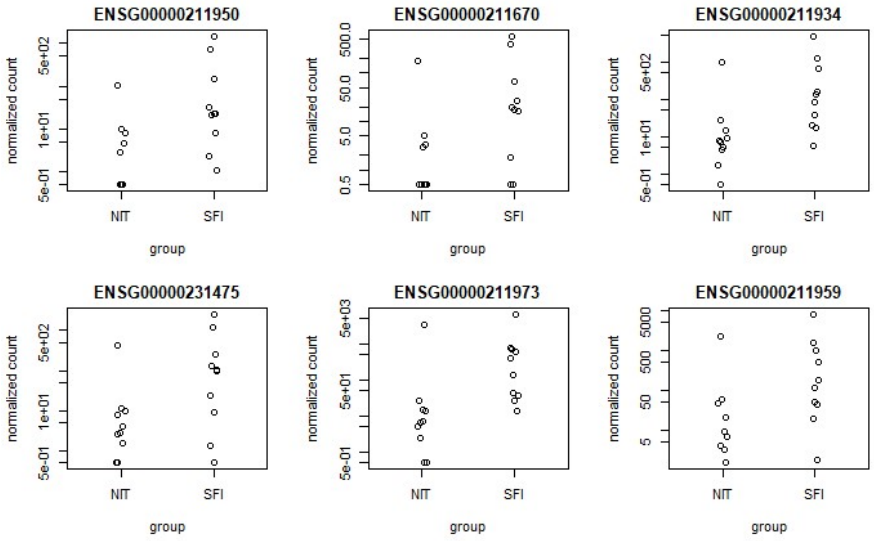
SFI-NIT



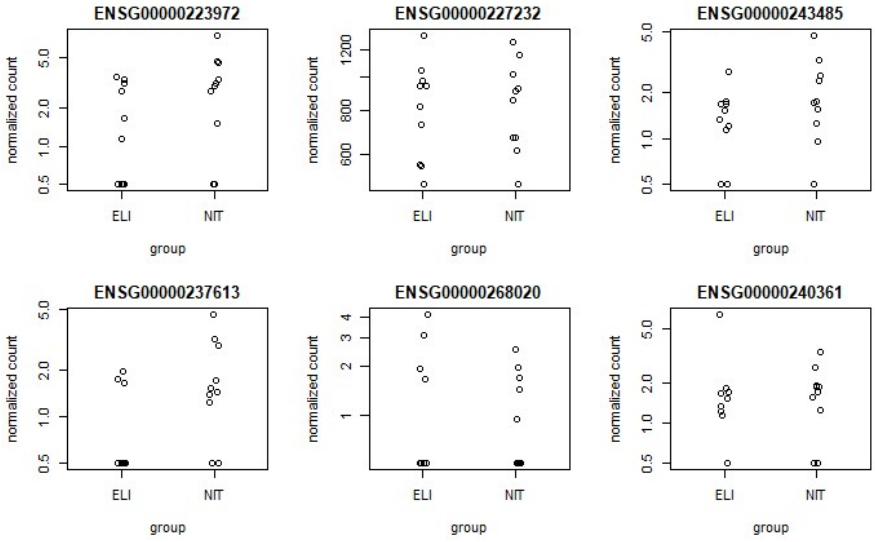
ELI-NIT



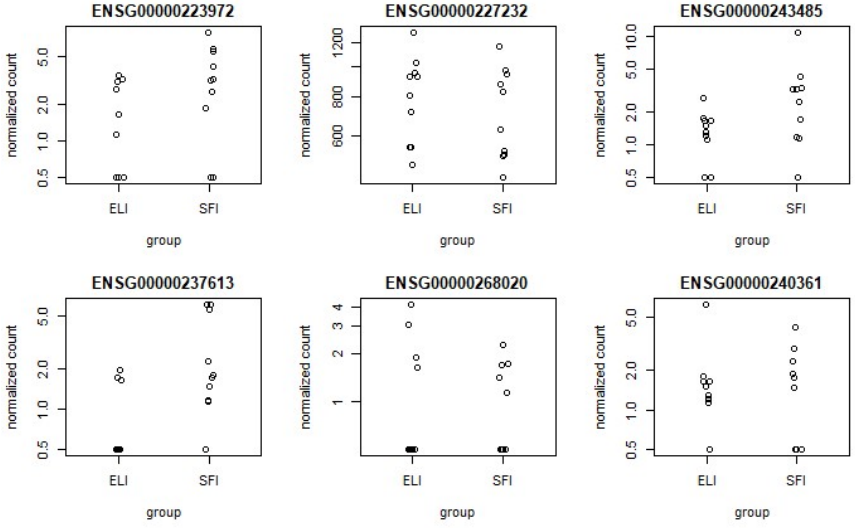
ELI-SFI



1ª Comparación

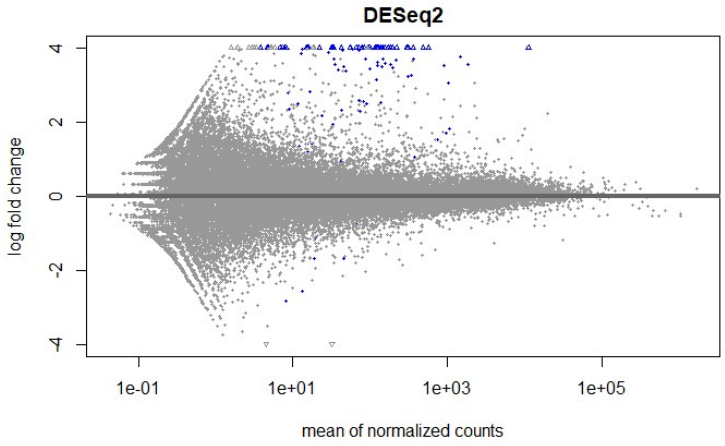


2ª Comparación

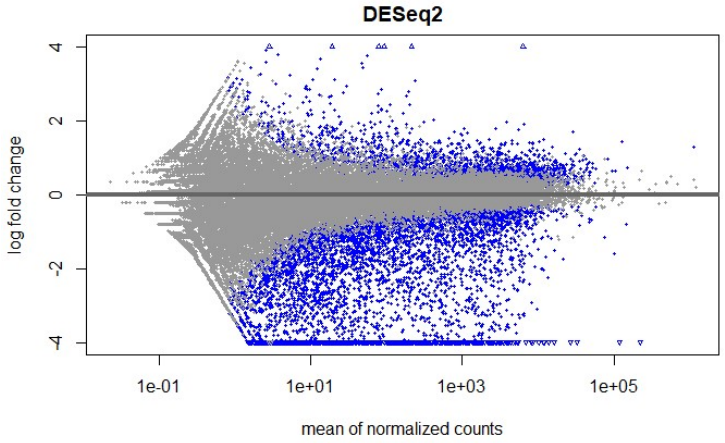


3ª Comparación

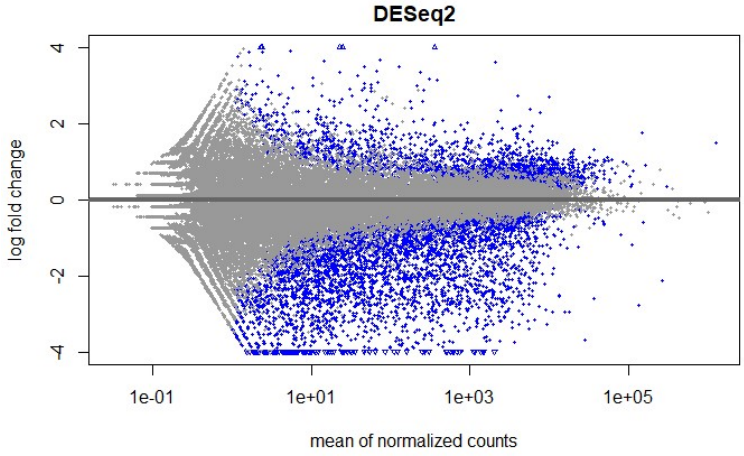
SFI-NIT



ELI-NIT



ELI-SFI



5. DISCUSIÓN

La parte del análisis de significación biológica no es muy completa por la dificultad de trabajar y modificar archivos con R. Si bien tiene ventajas a la hora de manejar grandes cantidades de datos para tareas de filtrado, cálculos y selección, puede ser frustrante al principio al realizar la preparación de los datos.