PEC 1 MO.157- Análisis de datos Ómicos. Rubén Jové Nieto

El estudio de dónde se han obtenido las muestras es el siguiente (1): Galicia, J. C., Henson, B. R., Parker, J. S., & Khan, A. A. (2016). Gene expression profile of pulpitis. Genes and immunity, 17(4), 239–243. https://doi.org/10.1038/gene.2016.14

Enlace al repositorio de github.

https://github.com/rbenjn/PEC_01

Abstract

Las enfermedades endodónticas presentan una prevalencia, coste y dolor que requiere de una mayor comprensión de los aspectos moleculares fundamentales de su patogénesis. Las pulpas inflamadas se recogieron de pacientes diagnosticados con pulpitis irreversible (n=20) y las pulpas normales sirvieron como controles (n=20). El dolor de valoró con la EVA. El análisis de microarrays con Affymetrix GeneTitan Multichannel Instrument. Como resultado, hubo una mayor expresión de genes involucrados en la respuesta inmune en el grupo de pulpitis. En cuanto a dolor, varios genes de lo modulan junto con la inflamación, mostraron una expresión diferencial en pacientes con dolor leve vs intenso.

Objetivos

El objetivo de este estudio es idenftificar a nivel genético, los factores que contribuyen al dolor e inflamación pulpar o pulpitis. También así poder proporcionar una base molecular para el diagnóstico clínico de pulpitis. Al comprender de mejor manera la inflamación pulpar, posteriores estudios sobre el tratamiento y manejo de la pulpitis y dolor asociado, pueden tener una referencia biológica.

Materiales y métodos

El estudio fue aprobado por la Oficina Ética de la Universidad de Carolina del Norte. Los datos provienen de pacientes que se trataban en la Escuela de Odontologia. Los criterios de inclusión fueron adultos que se presentaron para tratamiento endodóntico sin evidencia de patologías periapicales y sin terapia pulpar previa. Se excluyó a aquellos que tomaban debido a otras patologias o motivos, medicamentos de acción central que interfieren en la liberación de mediadores del dolor y/o alteran la respuesta inmune. El dolor fue valorado con la escala visual análogica (EVA). Clasificando las puntuaciones en: menos de 30 leve, 31-74 moderado y 75-100 dolor intenso. Se realizaron 2 grupos, uno de casos y el otro de controles. Las pulpas inflamadas se recogieron de pacientes diagnosticados con pulpitis irreversible (n=20) y las pulpas normales sirvieron como controles (n=20).

Los datos fueron preparados para estar analizados en una microarray Affymetrix Human Gene 2.1 ST. Se obtuvieron 12 muestras (archivos .CEL), 6 para cada grupo (Normal y Pulpitis). Mediante R y BioConductor se realizó el análisis de datos. Los cambios en la expresión de genes se analizó mediante el enriquicimiento del conjunto de genes con ClusterProfiler y el Pathway de Reactome.

Se realizarán dos comparaciones, por un lado pulpas normales vs inflamadas (6vs6). Por otro lado, dentro de las inflamadas, se comparará el dolor: leve vs severo (3vs3).

"Pipeline" análisis

Los pasos o "pipeline" seguido para el análisis han sido:

1. Identificar que grupos hay y a qué grupo pertenece cada muestra.
2. Control de calidad de los datos crudos
3. Normalización
4. Control de calidad de los datos normalizados
5. Filtraje no específico
6. Identificación de genes diferencialmente expresados
7. Anotación de los resultados
8. Comparación entre distintas comparaciones
9. Análisis de significación biológica ("Gene Enrichment Analysis")

"Pipeline" ampliada

1. Identificar que grupos hay y a qué grupo pertenece cada muestra. Accedemos al Accession Display de GSE77459. En su parte inferior podemos ver los Samples (12) con sus GSM y así saber a qué grupo pertenece cada muestra: normal o inflammed, intensidad de dolor leve o severa. A partir de eso, crearemos el archivo targets.csv, dónde identificaremos cada muestra según su grupo.

Table 1: Contenido del archivo targets utilizado para el análisis

FileName	Group	Genotype	Pain	ShortName
GSM2052371_1	Normal.Ninguno	Normal	Ninguno	N.N.1
GSM2052372_2	Normal.Ninguno	Normal	Ninguno	N.N.2
GSM2052373_3	Normal.Ninguno	Normal	Ninguno	N.N.3
GSM2052374_4	Normal.Ninguno	Normal	Ninguno	N.N.4
GSM2052375_5	Normal.Ninguno	Normal	Ninguno	N.N.5
GSM2052376_6	Normal.Ninguno	Normal	Ninguno	N.N.6
GSM2052377_7	Inflammed.Leve	Inflammed	Leve	I.L.1
GSM2052378_8	Inflammed.Leve	Inflammed	Leve	I.L.2
GSM2052379_9	Inflammed.Leve	Inflammed	Leve	I.L.3
GSM2052380_10	Inflammed.Severo	Inflammed	Severo	I.S.4
GSM2052381_11	Inflammed.Severo	Inflammed	Severo	I.S.5
GSM2052382_12	Inflammed.Severo	Inflammed	Severo	I.S.6

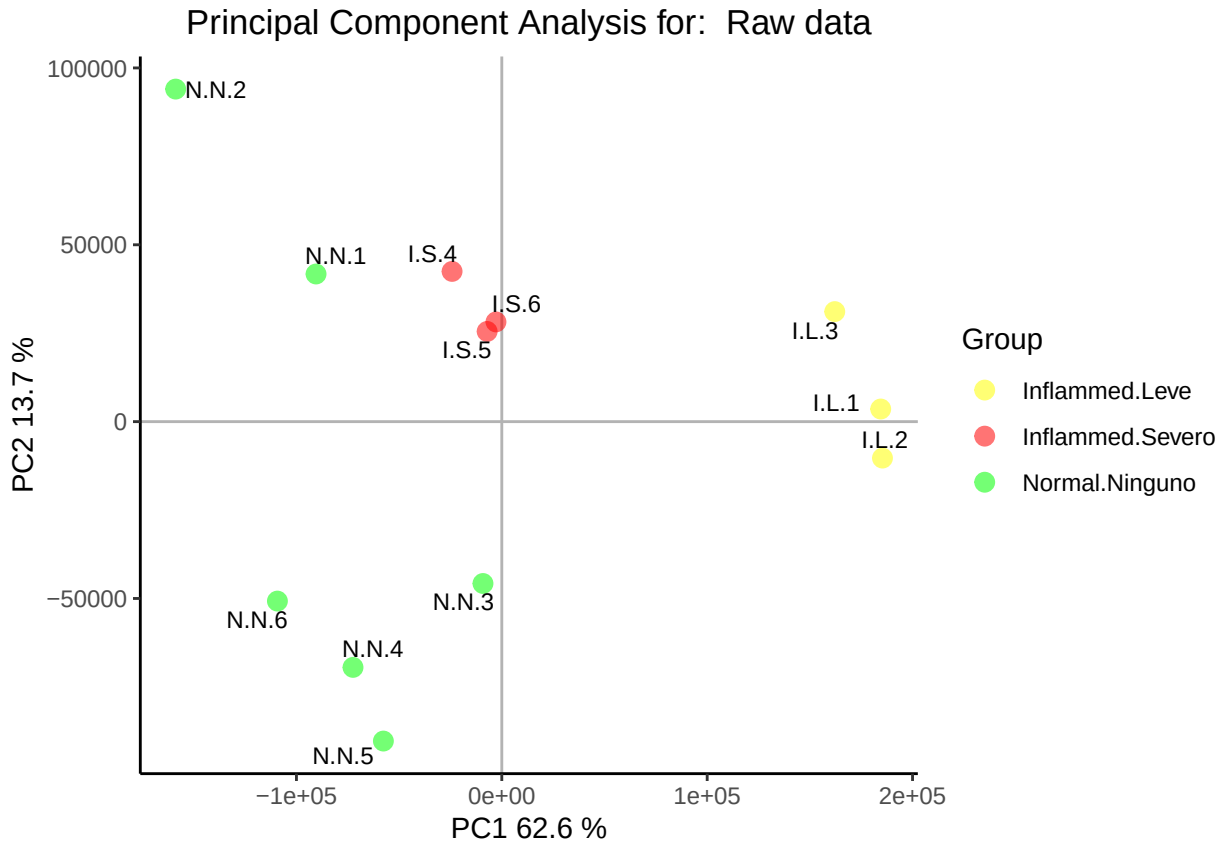
2. Control de calidad de los datos crudos Con el paquete "arrayQualityMetrics" podemos verificar si los arrays obtenidos son de buena cualidad. Según los resultados obtenidos, los mantendríamos o no en el análisis.

En nuestro caso, los resultados obtenidos como se muestra en la imagen, permiten mantener todos los arrays en el análisis. Sólo el número 2 ha sido marcado, pero sólo en una ocasión. Lo cuál indica que los problemas potenciales serán pequeños.

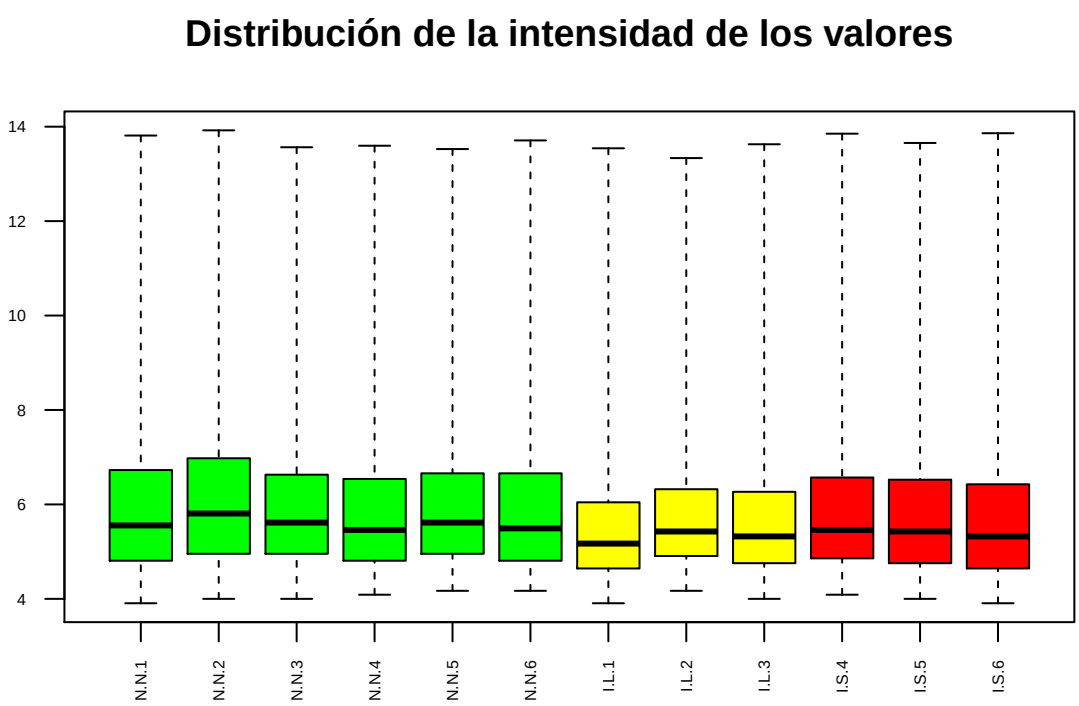
Mostramos en gráfico el análisis del componente y observamos como se distribuyen en función de su grupo, denotando unas diferencias:

	array	sampleNames	*1	*2	*3	Group	Genotype	Pain	ShortName
☐	1	N.N.1				Normal.Ninguno	Normal	Ninguno	N.N.1
☐	2	N.N.2			x	Normal.Ninguno	Normal	Ninguno	N.N.2
☐	3	N.N.3				Normal.Ninguno	Normal	Ninguno	N.N.3
☐	4	N.N.4				Normal.Ninguno	Normal	Ninguno	N.N.4
☐	5	N.N.5				Normal.Ninguno	Normal	Ninguno	N.N.5
☐	6	N.N.6				Normal.Ninguno	Normal	Ninguno	N.N.6
☐	7	I.L.1				Inflammed.Leve	Inflammed	Leve	I.L.1
☐	8	I.L.2				Inflammed.Leve	Inflammed	Leve	I.L.2
☐	9	I.L.3				Inflammed.Leve	Inflammed	Leve	I.L.3
☐	10	I.S.4				Inflammed.Severo	Inflammed	Severo	I.S.4
☐	11	I.S.5				Inflammed.Severo	Inflammed	Severo	I.S.5
☐	12	I.S.6				Inflammed.Severo	Inflammed	Severo	I.S.6

Figure 1: Tabla Control Calidad de los Datos Crudos

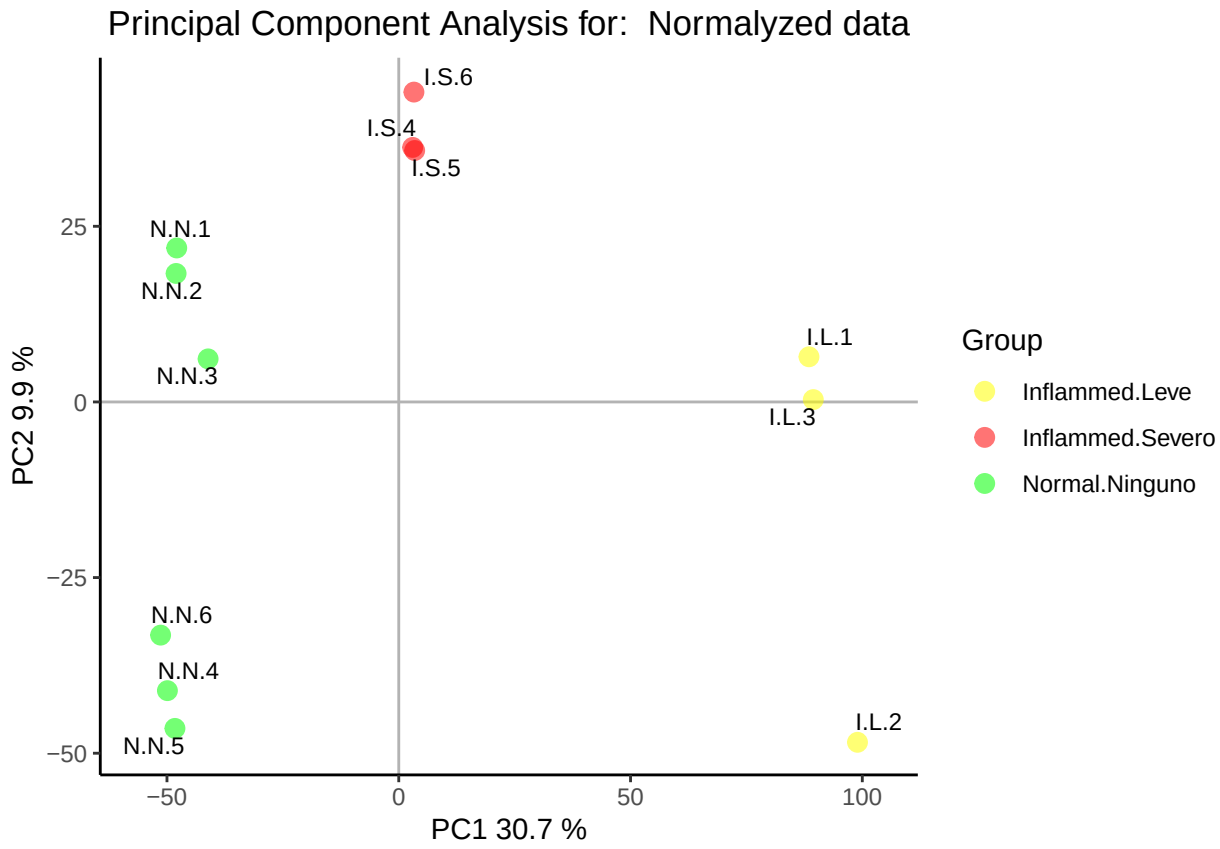


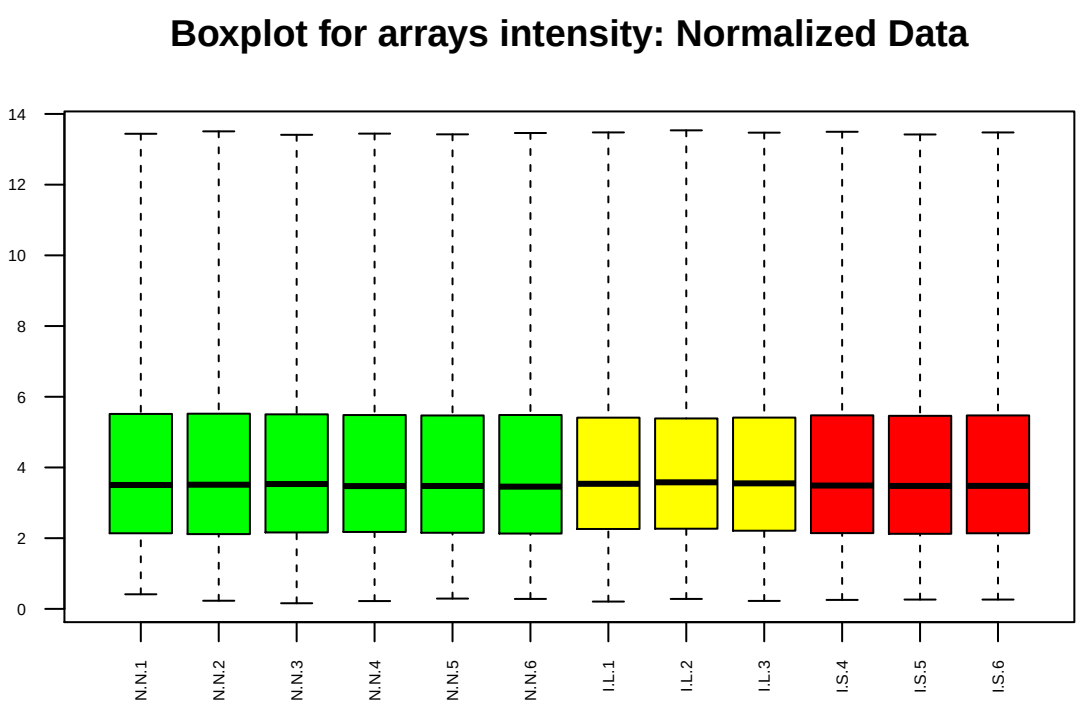
También mediante boxplot podemos visualizar la intensidad de la distribución:



3. Normalización Para normalizar los datos, usamos el método Robust Multichip Analysis.

4. Control de calidad de los datos normalizados Realizamos procedimiento que hemos llevado a cabo antes pero con los datos normalizados. Usaremos la función arrayQualityMetrics de la misma manera:


Principal Component Analysis for: Normalyzed data
I.S.6
I.S.4
I.S.5
N.N.1
N.N.2
N.N.3
I.L.1
I.L.3
N.N.6
N.N.4
N.N.5
I.L.2
25
0
−25
−50
−50
0
50
100
PC2 9.9 %
PC1 30.7 %
Group
Inflammed.Leve
Inflammed.Severo
Normal.Ninguno

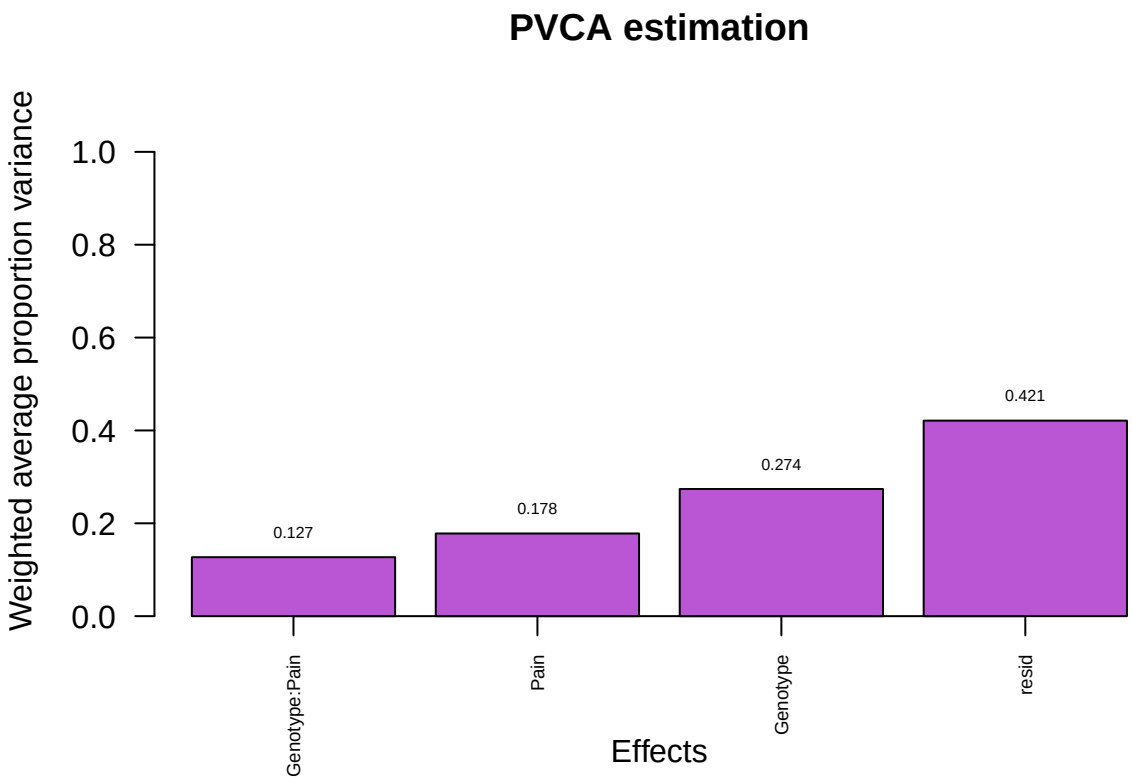


En el Boxplot observamos los datos ya normalizados, a diferencia del Boxplot del apartado 2.

5. Filtraje no específico Detectamos los pequeños cambios que se dan al procesar las muestras observando su varianza. Utilizamos el paquete "pvca".

Mostramos un gráfico que nos muestra cuáles variables muestran más varianza, en este caso la variable Genotipo:

```
## boundary (singular) fit: see ?isSingular
## boundary (singular) fit: see ?isSingular
## boundary (singular) fit: see ?isSingular
## boundary (singular) fit: see ?isSingular
```



6. Identificación de genes diferencialmente expresados Los genes con una mayor variabilidad se encuentran con una desviación estándar por encima del 90-95% de todas las desviaciones estándar.

A continuación, con la función "nsFilter" del paquete de BioConductor "genefilter" filtramos los genes con poca variabilidad y que por lo tanto no tienen una expresión diferencial.

```
## $numDupsRemoved
## [1] 2829
##
## $numLowVar
## [1] 19738
##
## $numRemoved.ENTREZID
## [1] 24470
```

Quedan 6580 genes filtrados.

7. Anotación de los resultados Creamos las matrices de diseño para realizar las posteriores comparaciones entre grupos:

```
##    I.L I.S N.N
## 1    0   0   1
## 2    0   0   1
## 3    0   0   1
```

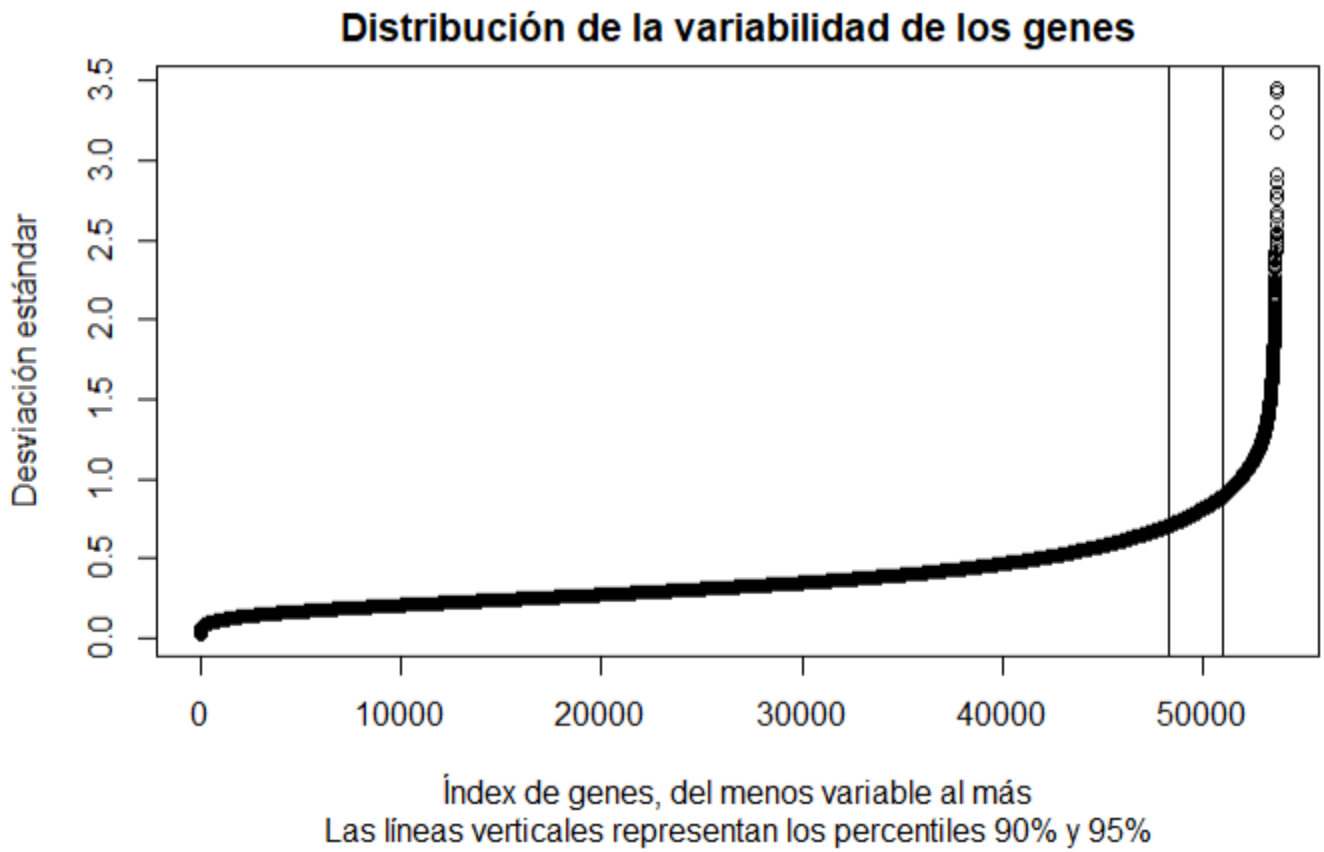


Figure 2: Distribución de la variabilidad de los genes

```
## 4    0   0   1
## 5    0   0   1
## 6    0   0   1
## 7    1   0   0
## 8    1   0   0
## 9    1   0   0
## 10   0   1   0
## 11   0   1   0
## 12   0   1   0
## attr(,"assign")
## [1] 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$Group
## [1] "contr.treatment"
```

A continuación definimos las comparaciones con las matrices de contraste:

```
##       Contrasts
## Levels InflammedvsNormal I.LvsI.S
##    I.L                1        1
##    I.S                1       -1
##    N.N               -1        0
```

Estimamos el modelo y la selección de genes con el paquete "limma":

Posteriormente, otbenemos una lista con los genes con mayor expresión diferencial.

Mostramos la cabecera para las 2 comparaciones.

En primer lugar, para la comparación Inflamado vs Normal

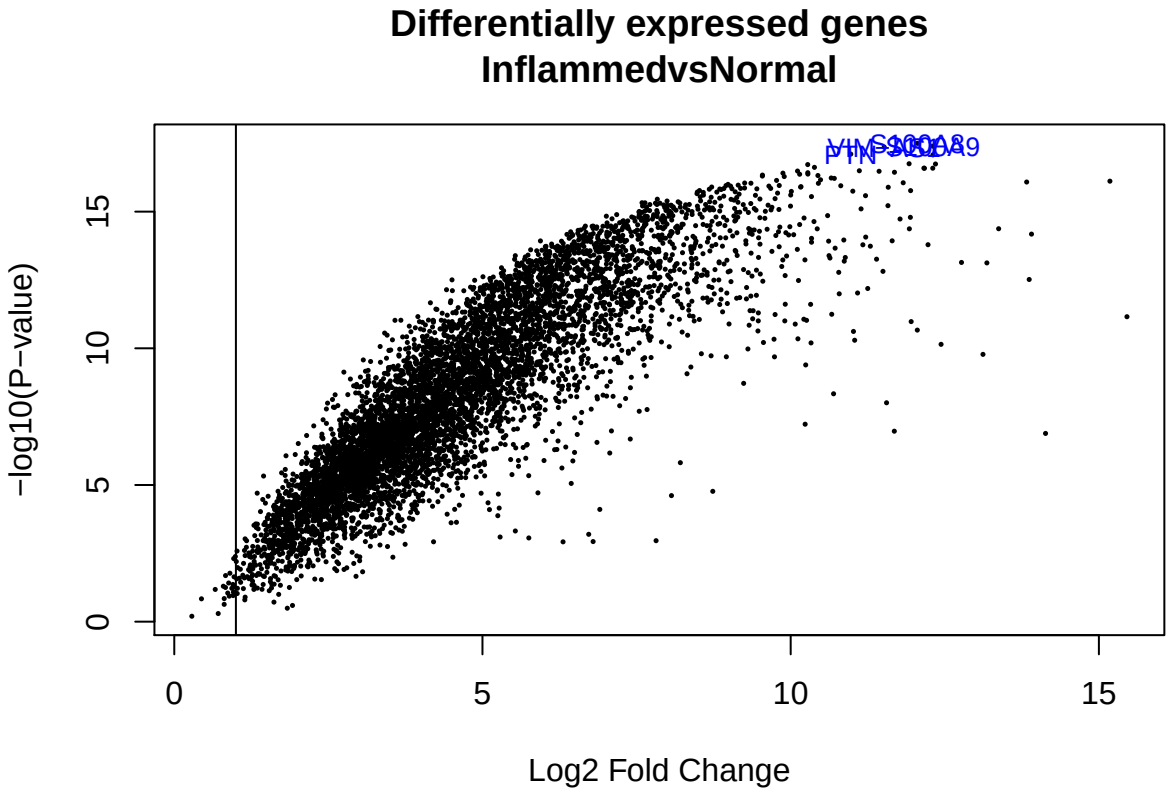
```
##             logFC   AveExpr        t      P.Value    adj.P.Val        B
## 16693414 12.05599 10.129853 84.53948 3.177221e-18 9.581665e-15 26.45255
## 16671139 12.30590  9.593264 82.93421 4.010337e-18 9.581665e-15 26.37938
## 16712272 11.51708 12.134091 82.35211 4.368540e-18 9.581665e-15 26.35199
## 17063254 10.97504 11.766863 78.46616 7.857653e-18 1.104801e-14 26.15667
## 16859763 12.30778  8.686143 78.03981 8.395146e-18 1.104801e-14 26.13383
## 17126266 11.92108 11.013317 73.39414 1.768824e-17 1.409511e-14 25.86473
```

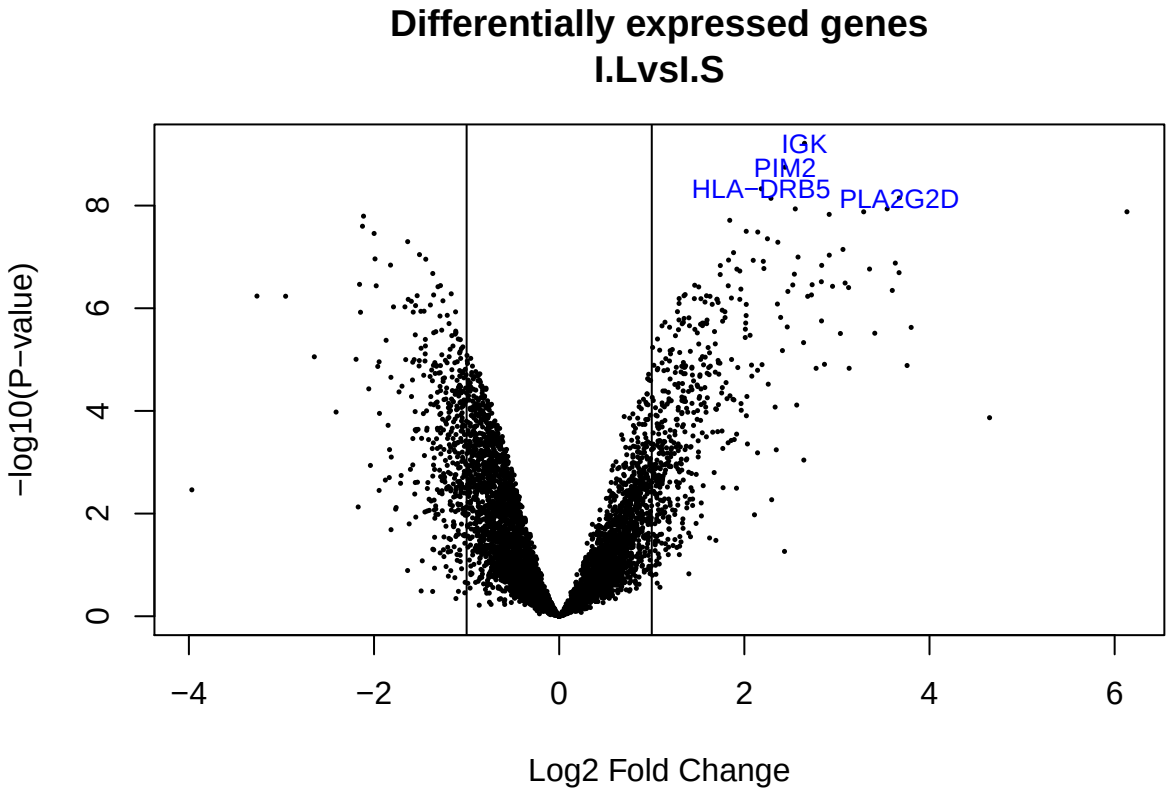
En este caso, para la comparación de Dolor entre muestras del grupo Inflamado:

```
##             logFC  AveExpr        t      P.Value    adj.P.Val        B
## 16882555 2.649697 8.121571 17.31203 6.133384e-10 4.035767e-06 13.06798
## 17110670 2.439695 4.651897 15.79809 1.786757e-09 5.878431e-06 12.12189
## 17017885 2.182432 6.943299 14.53149 4.709305e-09 9.515255e-06 11.24050
## 16682785 3.674498 3.372725 14.02292 7.101155e-09 9.515255e-06 10.86076
## 16689400 2.287339 5.267487 14.00095 7.230437e-09 9.515255e-06 10.84400
## 16976827 3.542751 5.027086 13.43985 1.156302e-08 9.613389e-06 10.40555
```

Visualizamos los genes expresado de manera distinta mediante un volcano plot:

```
## 'select()' returned 1:1 mapping between keys and columns
```

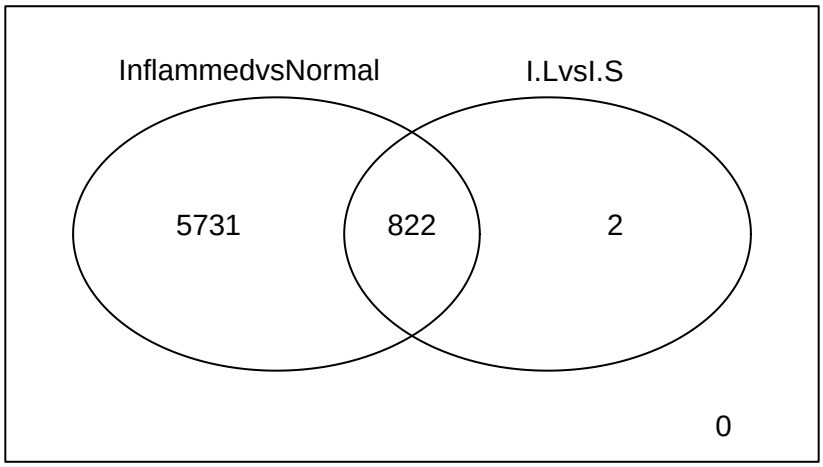

Differentially expressed genes
InflammedvsNormal
−log10(P−value)
0
5
10
15
0
5
10
15
Log2 Fold Change



En el volcano plot mostramos los nombres de los 4 genes expresados de de mayor manera diferencial. Observamos las diferencias entre grupos según el gráfico obtenido. Distinto entre las 2 comparaciones, debido a sus diferencias de grupo.

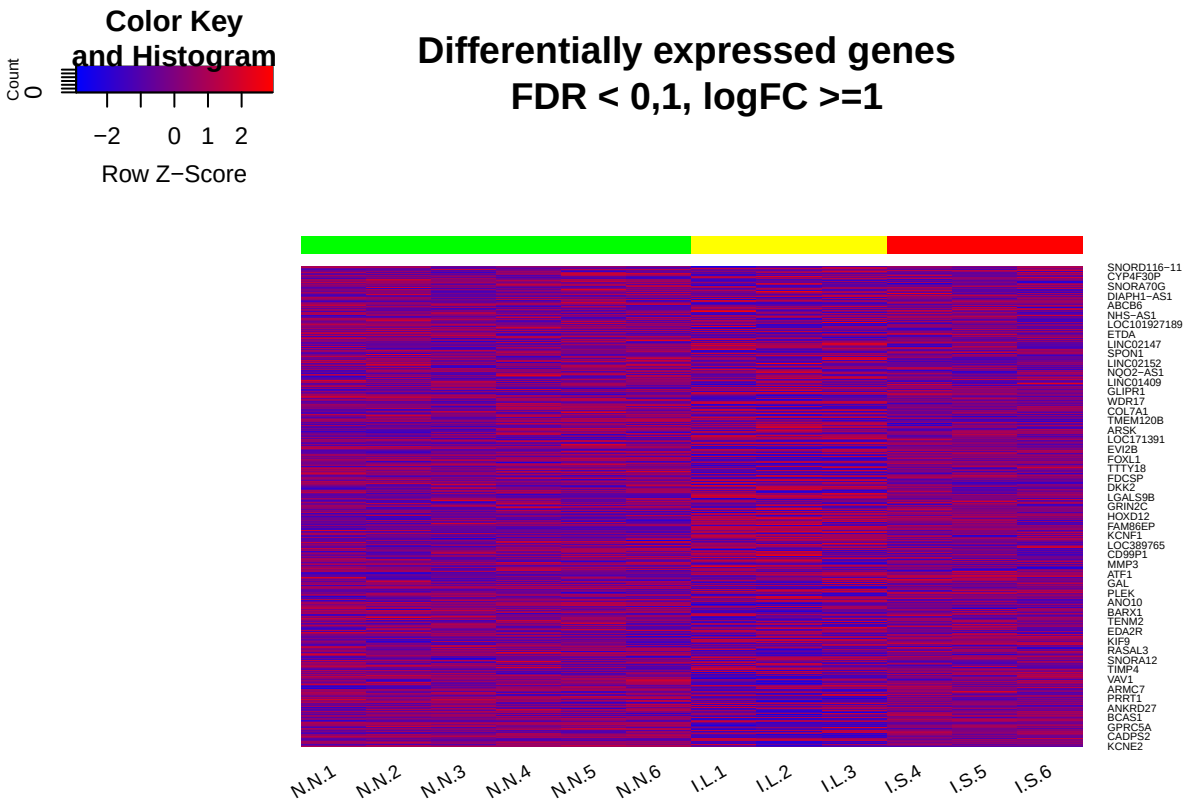
8. Comparación entre distintas comparaciones Realizamos las múltiples comparaciones:

```
##        InflammedvsNormal I.LvsI.S
## Down                   0      366
## NotSig                27     5756
## Up                  6553      458
```



El Diagrama de Venn nos muestra los genes comunes entre las 2 comparaciones. Un total de 822
Los Mapas de Calor:

```
## 'select()' returned 1:1 mapping between keys and columns
```



9. Análisis de significación biológica ("Gene Enrichment Analysis") Preparamos lista de genes analizados

```
## 'select()' returned 1:1 mapping between keys and columns
## 'select()' returned 1:1 mapping between keys and columns
```

```
## InflammedvsNormal          I.LvsI.S
##              6551              2020
```

Los genes a tener en cuenta, tienen al menos una anotación en Gene Ontology.

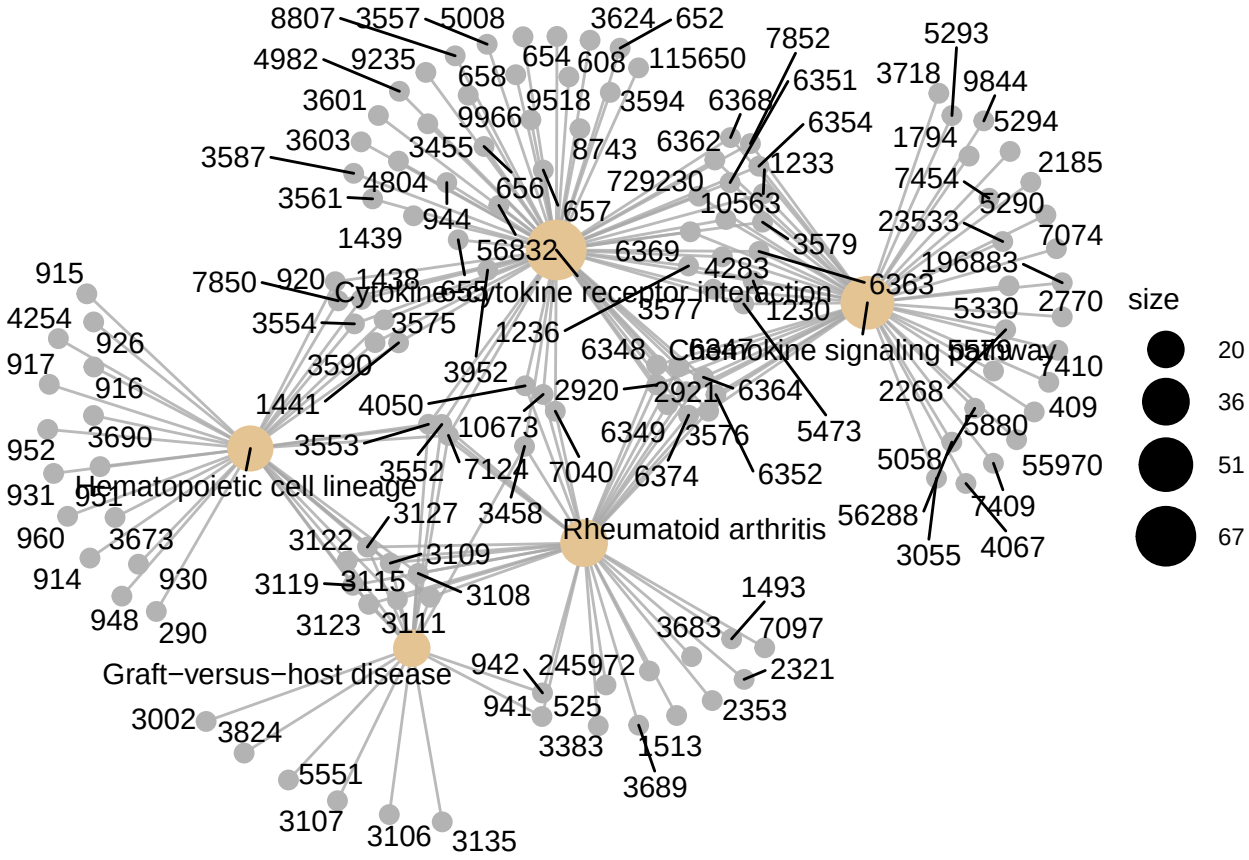
Con el paquete "clusterProfiler" y siguiendo el Pathway de ReactomePA, realizamos el análisis de la significación biológica.

Mostramos para cada comparación, los genes más expresados diferencialmente y con la descripción según su función biológica ("Gene Ontology"):

ID	Description	GeneRatio	BgRatio	pvalue
hsa04640	Hematopoietic cell lineage	61/1841	98/7991	7.75703789476954e-17
hsa04060	Cytokine-cytokine receptor interaction	126/1841	294/7991	1.27065310331772e-14
hsa05323	Rheumatoid arthritis	54/1841	92/7991	1.57873150510695e-13
hsa04061	Viral protein interaction with cytokine and cytokine receptor	55/1841	100/7991	3.68780810401393e-12
hsa05332	Graft-versus-host disease	30/1841	41/7991	1.27972801230647e-11
hsa05150	Staphylococcus aureus infection	51/1841	93/7991	2.61812481736328e-11

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust
hsa05323	Rheumatoid arthritis	36/717	92/7991	3.20304563264605e-15	1.00575632865086e-12
hsa04060	Cytokine-cytokine receptor interaction	67/717	294/7991	2.80614605621923e-13	4.4056493082642e-11
hsa04062	Chemokine signaling pathway	49/717	189/7991	3.64046978404517e-12	3.81035837396728e-10
hsa04640	Hematopoietic cell lineage	33/717	98/7991	6.42509396696605e-12	5.04369876406835e-10
hsa05332	Graft-versus-host disease	20/717	41/7991	3.89329285797259e-11	2.44498791480679e-09
hsa04380	Osteoclast differentiation	35/717	127/7991	8.03170461635628e-10	4.20325874922645e-08

Los datos al completo los encotramos almacenados en los respectivos archivos .csv para cada comparación.



Resultados

Obtenemos resultados significativos de ambas comparaciones.

Muestra Normal vs Muestra Pulpitis

El Gene Set Enrichment Analysis mostró genes que se expresan de manera diferencial. Con una mayor expresión en la muestra de Pulpitis de genes asociados con la activación de la respuesta inmunitaria.

Muestra Dolor Leve vs Dolor Severo Entre las personas que referían dolor (3 severo y 3 leve), encontramos genes expresados de forma diferencial entre los 2 grupos. Con especial atención a aquellos genes relacionados con el sistema inmune adaptativo y la interacción citoquina-citoquina.

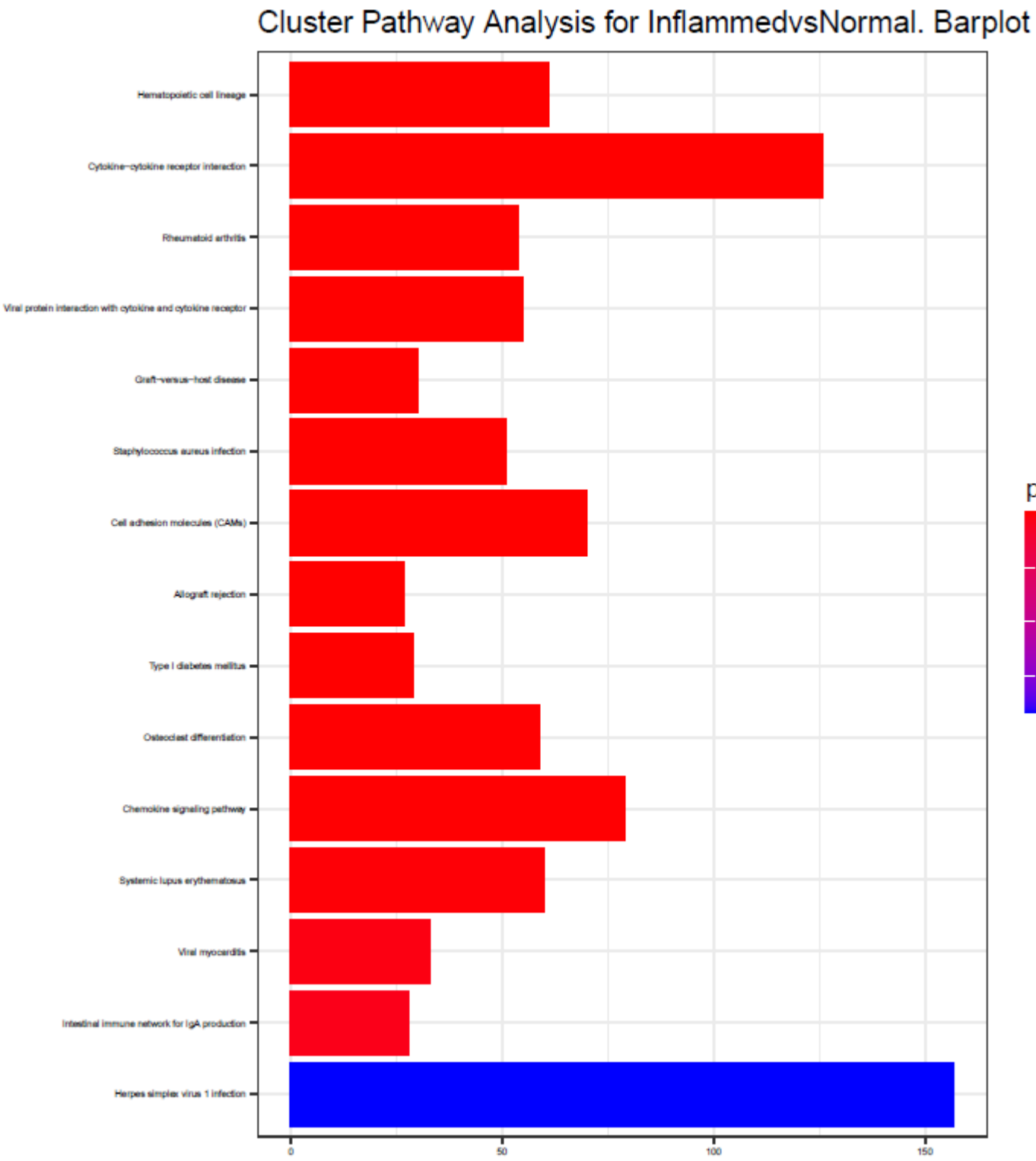


Figure 3: Barplot Genes expresados diferencialmente Normal vs Pulpitis

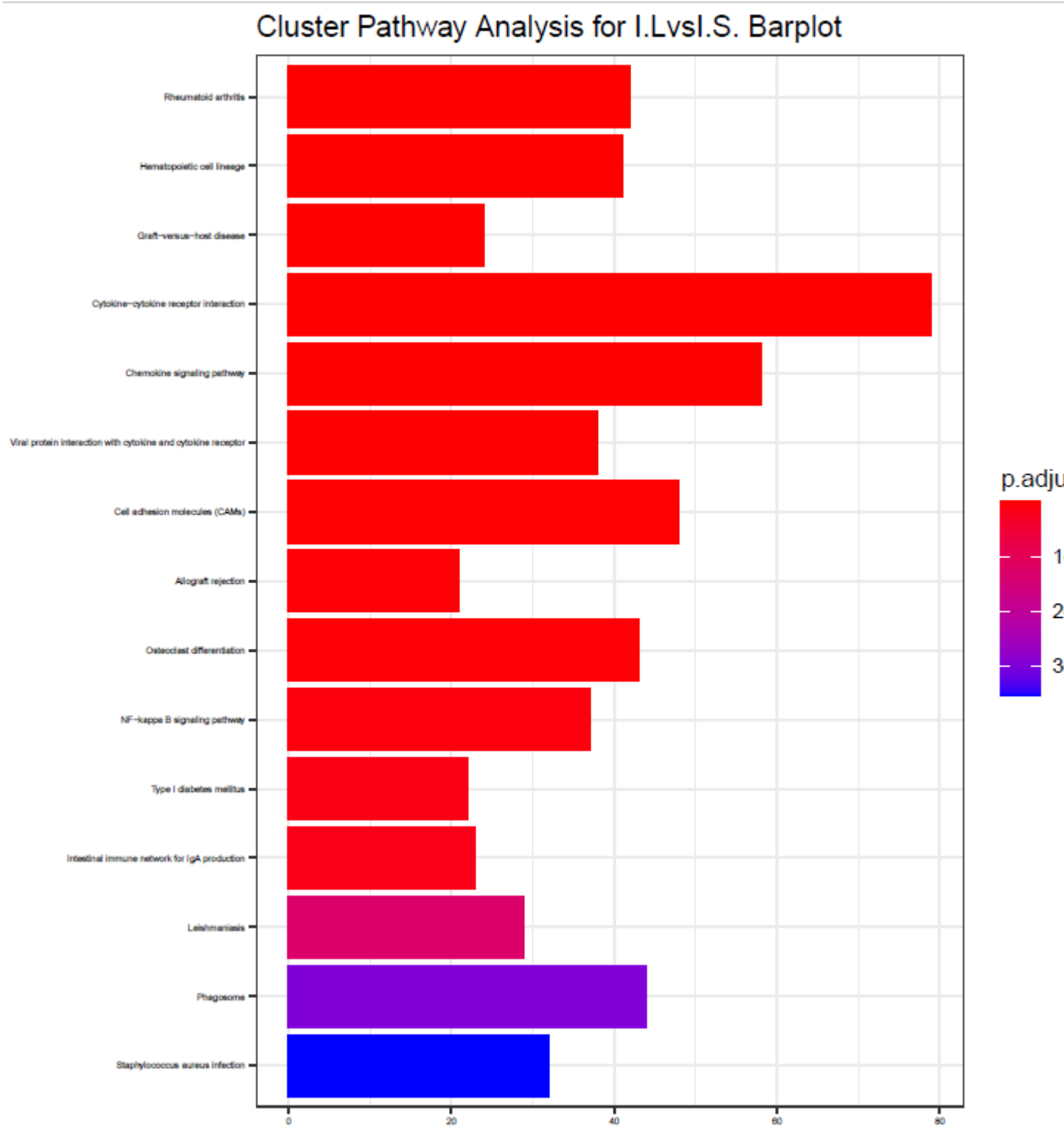


Figure 4: Barplot Genes expresados diferencialmente Dolor Leve vs Dolor Severo

Discusión

Aunque el uso del análisis de microarrays es una herramienta poderosa para estudiar la expresión simultánea de varios genes, hay algunas limitaciones.

Por ejemplo, factores técnicos como el rango limitado y la hibridización cruzada. Además el perfil celular de las pulpas inflamadas difiere de las normales. En las inflamadas se caracterizan por tener una afluencia ya por encima de lo normal de células inmunes.

Es decir, que los resultados pueden ser en parte debido a diferencias en la composición celular.

Apéndice

Enlace al repositorio de github, dónde se halla el código R utilizado entre los otros archivos del estudio.

https://github.com/rbenjn/PEC_01

Bibliografía

(1) Galicia, J. C., Henson, B. R., Parker, J. S., & Khan, A. A. (2016). Gene expression profile of pulpitis. Genes and immunity, 17(4), 239–243. https://doi.org/10.1038/gene.2016.14

(2) Yu, Guangchuang, and Qing-Yu He. 2016. "ReactomePA: An R/Bioconductor Package for Reactome Pathway Analysis and Visualization." Molecular BioSystems 12 (2): 477–79. https://doi.org/10.1039/C5MB00663E.

(3) Yu, Guangchuang, Li-Gen Wang, Guang-Rong Yan, and Qing-Yu He. 2015. "DOSE: An R/Bioconductor Package for Disease Ontology Semantic and Enrichment Analysis." Bioinformatics 31 (4): 608–9. https://doi.org/10.1093/bioinformatics/btu684.