

# ADO\_PEC2

Adria\_Fernandez

29/5/2020

## Contents

<b>Abstract</b>	<b>2</b>
<b>Objectius</b>	<b>2</b>
<b>Materials</b>	<b>2</b>
Software . . . . .	2
Dades . . . . .	2
Disseny experimental: . . . . .	2
<b>Mètodes</b>	<b>3</b>
Pipeline . . . . .	3
Procediment . . . . .	3
Preparació del entorn de treball i descàrrega de dades: . . . . .	3
Selecció de les dades: . . . . .	3
Preparació de les dades: . . . . .	3
Identificació de gens diferencialment expressats: . . . . .	3
Anotació dels resultats: . . . . .	3
Agrupació de les mostres: . . . . .	4
Anàlisis de significació biològica: . . . . .	4
<b>Resultats</b>	<b>5</b>
Agrupació de les mostres: . . . . .	5
Diferencies d'expressió gènica: . . . . .	6
Identificació dels gens: . . . . .	7
Anàlisi de significació biològica . . . . .	8
<b>Discusió</b>	<b>10</b>
<b>Apèndix</b>	<b>11</b>
<b>Referències</b>	<b>11</b>

## Abstract

Aquest informe està basat en les dades d'expressió (RNA-seq) d'un conjunt d'anàlisis de tiroides per a tres grups diferenciat (Not infiltrated tissues (NIT): 236 samples, small focal infiltrates (SFI): 42 samples, extensive lymphoid infiltrates (ELI): 14 samples). S'han buscat diferències significatives en l'expressió gènica dels diferents grups, tant sobre-expressió com sub-expressió utilitzant Bioconductor en R com a eina de treball.

El codi d'execució, així com al resta de fitxers descarregats i generats es poden trobar a: [[https://github.com/Adria-FerDi/ADO\\_PEC2](https://github.com/Adria-FerDi/ADO_PEC2)]

## Objectius

L'objectiu d'aquest exercici es el d'aplicar una pipeline apropiada per a l'estudi de les dades d'expressió (RNA-seq) d'un conjunt d'anàlisis de tiroides. S'efectuaran 3 comparacions d'expressió diferencial entre els tres grups disponibles: -Not infiltrated tissues (NIT): 236 samples -Small focal infiltrates (SFI): 42 samples -Extensive lymphoid infiltrates (ELI): 14 samples. Agafant, de cada grup, 10 mostres aleatories.

## Materials

### Software

Aquest informe ha utilitzat [R]<sup>1</sup> Com a programari principal de treball, mitjançant l'interfaç [RStudio]<sup>2</sup>. Algunes llibreries emprades no formen part del llistat de R, i s'ha recolzat molt en el paquet [Bioconductor]<sup>3</sup> com a eina principal per a l'anàlisis de dades òmiques, en concret el seu paquet DESeq2.

### Dades

Les dades fetes servir en aquest estudi provenen del repositori GTEx (Genotype-Tissue Expression project), el qual ens proporciona dades d'expressió gènica de 54 classes de teixit de 1000 subjectes. S'han fet servir les dades d'expressió (RNA-seq) dels teixits tiroidals, classificats en 3 grups:

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

Per tal d'homogeneitzar les mostres, s'han extret 10 mostres de cada grup al azar.

### Disseny experimental:

En aquest cas farem un experiment de comparació de classes, amb l'objectiu de visualitzar les diferències en les expressions gèniques dels tres grups. Farem les tres comparacions possibles: - NIT vs SFI: Per observar l'efecte del tractament SFI, comparant-ho amb no tractats. - NIT vs ELI: Per observar l'efecte del tractament ELI, comparant-ho amb no tractats. - SFI vs ELI: Per comparar els efectes dels dos tractaments.

<sup>1</sup><https://cran.r-project.org/index.html>

<sup>2</sup><https://www.rstudio.com/>

<sup>3</sup><https://www.bioconductor.org/>

# Mètodes

## Pipeline

- 1 Preparació del entorn de treball i descàrrega de dades.
- 2 Selecció de les dades.
- 3 Preparació de les dades.
- 4 Identificació de gens diferencialment expressats.
- 5 Anotació dels resultats
- 6 Agrupació de les mostres
- 7 Anàlisis de significància biològica.

## Procediment

### Preparació del entorn de treball i descàrrega de dades:

Primer de tot generem les carpetes necessàries per a l'efectuació del treball (dades i resultats), i direccions cap aquestes. Posteriorment descarreguem les dades proporcionades pel professor. Per a la reproduïibilitat de l'estudi s'haurien de descarregar des de github.

### Selecció de les dades:

Extraiem les dades dels archius csv, seleccionem els grups d'interès i de cadascun n'extraiem 10 noms de les mostres aleatories. Les quals es juntaran en un mateix dataframe, emprat per extreure les informacions del arxiu count. Per fer aquest procediment s'estableix una seed per a la reproduïibilitat de l'estudi.

### Preparació de les dades:

Preparam les dades per ser tractades amb el paquet DESeq2. Utilitzarem la funció DESeqDataSetFromMatrix, amb el disseny experimental descrit en funció del tractament, per generar l'objecte dds.

El següent pas es fer un prefiltratge, eliminant les observacions que es puguin considerar massa baixes, ja que podrien generar efectes de "soroll". En aquest punt caldria saber si algun gen té especialment interès per l'estudi, per tal de no eliminar-lo. En el nostre cas tractarem aquells gens amb una expressió superior a 10 entre totes les mostres. Aconseguint una disminució de 20000 gens.

A continuació caldrà normalitzar les dades, d'entre els diferents mètodes, en destaquen VST i rlog. en el nostre cas farem servir VST, ja que per la mida de les mostres és més útil que no pas rlog.

### Identificació de gens diferencialment expressats:

A partir de l'objecte dds generem un dataframe amb la funció DESeq, el qual ens retornarà aquells gens diferencialment expressats. Farem les tres comparacions experimentals ja esmentades.

Per últim farem un filtratge dels resultats obtinguts, en funció del seu valor, acceptant aquells que tinguin un p-valor inferior a 0.15.

### Anotació dels resultats:

Per a poder anotar els resultats referents a la diferent expressió gènica necessitem referenciar els gens, mitjançant el paquet annotationDbi. Com a base de dades de referència hem fet servir org.Hs.eg.db, fent la traducció dels codis d'Ensembl a símbols.

### **Agrupació de les mostres:**

Per visualitzar l'agrupació de les mostres es faran servir diferents mètodes, sempre fent servir les dades normalitzades:

- Primer de preparem les dades per fer un heatmap, que ens agrupa les mostres mitjançant unes noves variables, les components. La funció a executar es plotPCA.
- Per altra banda podem representar una matriu de distàncies, mitjançant el paquet pheatmap.
- Mitjançant genefilter podem seleccionar els 20 gens més diferenciats i observar la seva classificació.
- Per poder efectuar un plot MA convertirem les dades dels objectes ddSeq per evitar sorolls, mitjançant apeglm.
- Efectuarem també un diagrama de venn per comparar les comparacions, però per aquest no cal fer un tractament de les dades extra.

### **Anàlisis de significància biològica:**

Quan ja tenim els gens diferencialment expressats només ens resta saber si aquests es deuen a algun proces conegut o documentat. Utilitzarem el paquet clusterProfiler, cercant per patrons de repetició per les diferents categories de GO. Conjuntament amb enrichplot podrem fer gràfics per representar els resultats.

## Results

## Agrupació de les mostres:

Representem gràficament els resultats de les agrupacions de mostres, començant per les representacions de heatmaps:

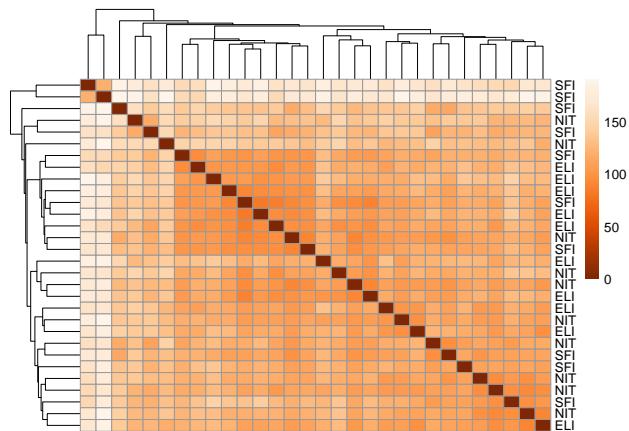


Figure 1: Heatmap de l'agrupació de les mostres

Aquest gràfic, però, no ens mostra una evidència clara entre la agrupació dels diferents tractametns.

En segon lloc, intentem visualitzar l'agrupament de les mostres mitjançant un anàlisi de components principals (o PCA):

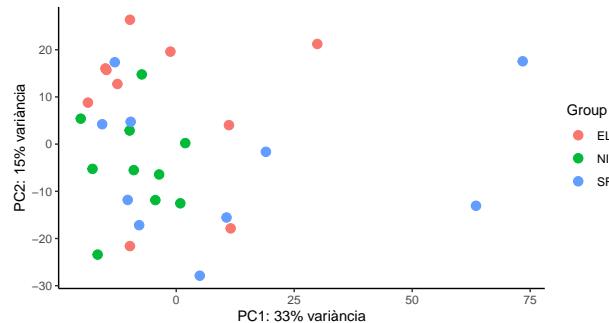


Figure 2: PCA de l'agrupació de les mostres

Seguim sense evidenciar una agrupació clara entre els diferents grups de mostreig. Per últim provem graficant

un heatmap amb els 20 gens més diferenciats:

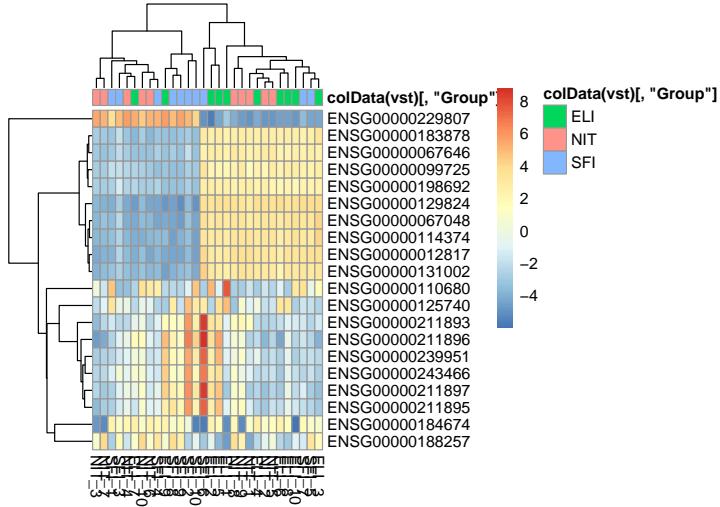


Figure 3: Heatmap amb els 20 gens més significatius

En aquest cas tampoc s'extreu una evidència sobre les diferenciacions a nivell d'expressió entre els grups.

### Diferencies d'expressió gènica:

Per altre part podem observar les diferencies obtingudes en el conteig de gens, esperant obtenir informació d'interès:

Per visualitzar les diferenciacions en les comparacions fetes (SFI-ELI, SFI-NIT, ELI-NIT) observarem els MA plots, amb l'objectiu de comparar l'expressió entre grups en els 3 casos:

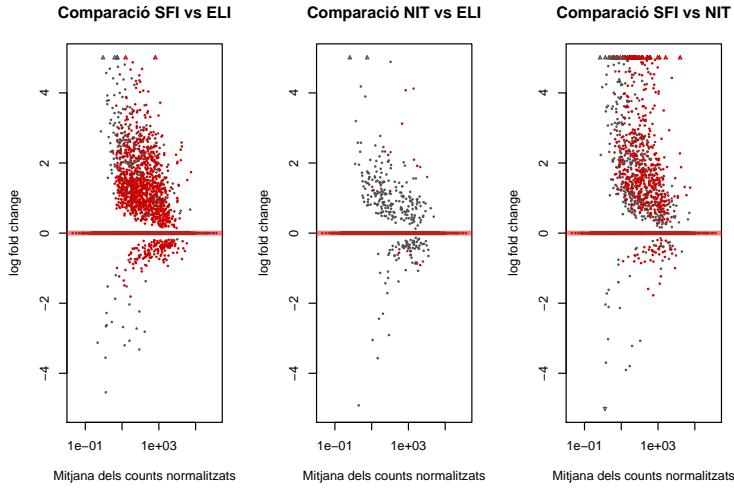


Figure 4: MA plots de les diferències d'expressió gènica

En aquests gràfics observem diferències en les comparacions, sent NIT vs ELI una comparació amb poca abundància de gens, amb una lleugera tendència cap a la sobreexpressió en el cas de NIT, les altres dues comparacions, amb més abundància de gens, mostren més clarament una sobreexpressió genica, en amds casos, per al grup SFI.

Amb el diagrama de Venn podem visualitzar les similituds:

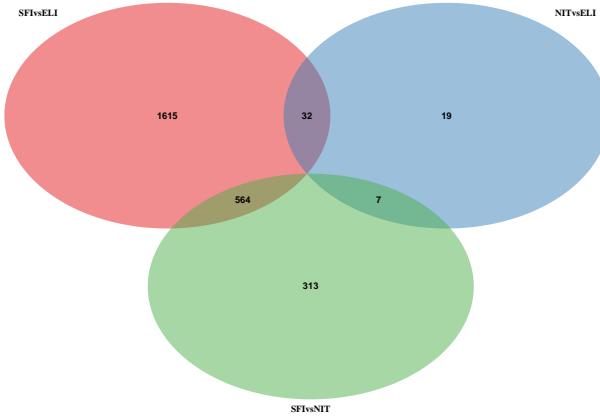


Figure 5: Diagrama de Venn de les tres comparacions

S'observen poques diferencies en la comparació NIT-ELI, fet que podrà mostrar que el tractament ELI no és gaire eficaç, mentre que el tractament amb SFI (NIT-SFI) presenta més evidències de diferenciació. Per últim, la comparació que presenta més diferències és entre els dos tractaments.

### Identificació dels gens:

Podem observar, quins són els gens diferenciats en les comparacions:

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol
ENSG00000143552	95.57215	3.718152	0.6028739	6.167379	0	0.0000099	NUP210L
ENSG00000185686	37.86416	5.161329	0.8389721	6.151967	0	0.0000099	PRAME
ENSG00000167767	210.52843	3.129576	0.5175175	6.047286	0	0.0000127	KRT80
ENSG00000173110	645.12838	3.696700	0.6235191	5.928767	0	0.0000198	HSPA6
ENSG00000143119	1407.05591	2.960416	0.5325223	5.559234	0	0.0001257	CD53

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol
ENSG00000179772	233.27562	-2.3692046	0.4912790	-4.822524	1.40e-06	0.0211266	FOXS1
ENSG00000174460	711.67467	4.5111476	0.9465337	4.765966	1.90e-06	0.0211266	ZCCHC12
ENSG00000132854	91.84208	-2.6457931	0.5860742	-4.514434	6.30e-06	0.0475744	KANK4
ENSG00000178445	800.14327	-2.6121818	0.5868287	-4.451353	8.50e-06	0.0479581	GLDC
ENSG00000196465	1269.57699	-0.5868725	0.1344938	-4.363564	1.28e-05	0.0575336	MYL6B

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol
ENSG00000143552	95.57215	3.511280	0.6018203	5.834432	0e+00	0.0001400	NUP210L
ENSG00000173110	645.12838	3.478722	0.6233591	5.580607	0e+00	0.0003108	HSPA6
ENSG00000105369	266.41256	5.318507	0.9884350	5.380735	1e-07	0.0006412	CD79A
ENSG00000128438	63.93711	6.526359	1.2497243	5.222239	2e-07	0.0011460	NA
ENSG00000172578	501.92980	2.656244	0.5171143	5.136667	3e-07	0.0013679	KLHL6

Observem que les diferències en SFI-NIT no estan anotades a la base org.Hs.eg.org, s'hauria de fer cerca externa sobre el gen si generen interès per a l'investigador.

## Anàlisi de significació biològica

Després de visualitzar els gens, intentarem encabir-los en algún procés biològic concret. Mitjançant enrichment analysis del paquet clusterProfiler. Efectuarem dotplots de les diferents comparacions per visualitzar les funcions així com la quantitat de gens implicats.

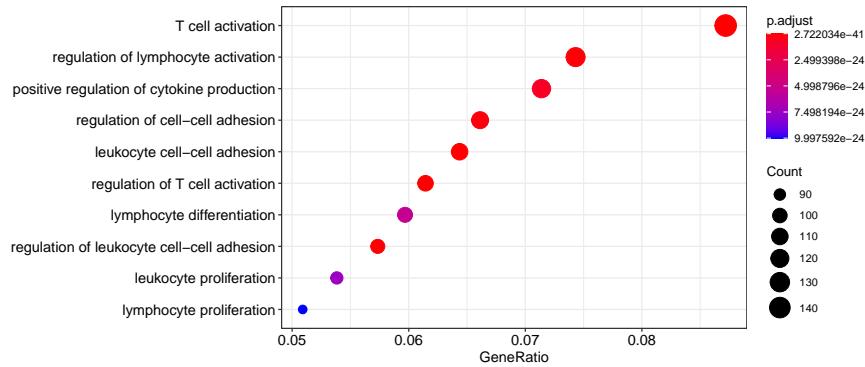


Figure 6: Dotplots SFI-ELI

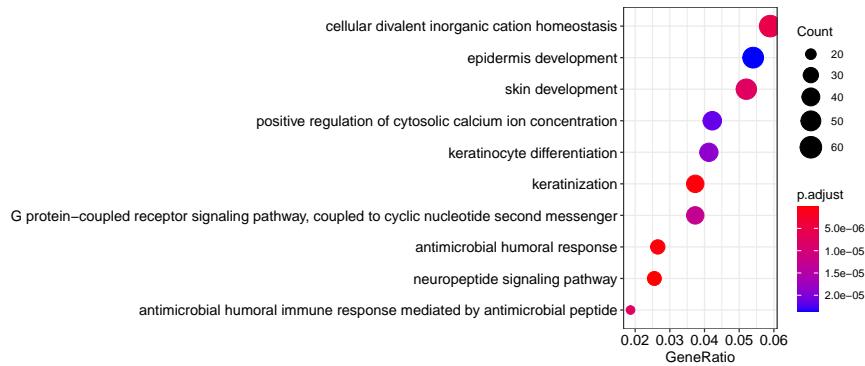


Figure 7: Dotplots NIT-ELI

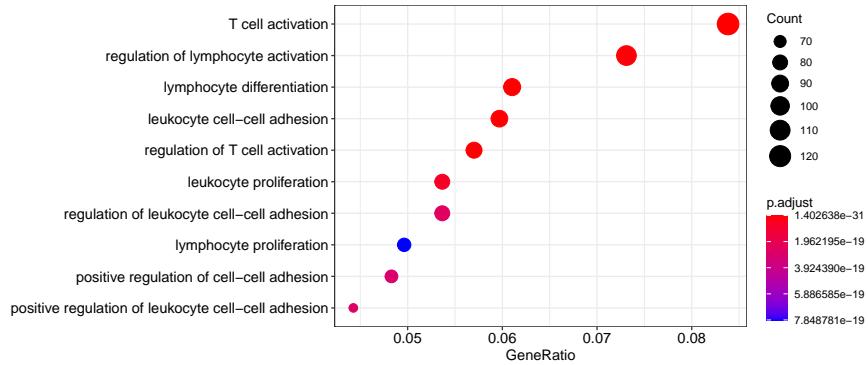


Figure 8: Dotplots SFI-NIT

La comparació SFI-ELI i la SFI-NIT presenten similituds, i es centren bastant en els limfòcits i altres parts del sistema immune. Pel que fa a la comparació NIT-ELI presenten elevada variació entre funcions, incloent algunes funcions relacionades amb la pell, la keratina, i la resposta humorala.

Visualitzarem ara una xarxa, que ens permetrà visualitzar aquells gens implicats en diferents processos. Els noms dels gens s'han omès perquè imedien la interpretació d'aquests, en cas de ser d'interés algú en concret podria extreure's fàcilment.

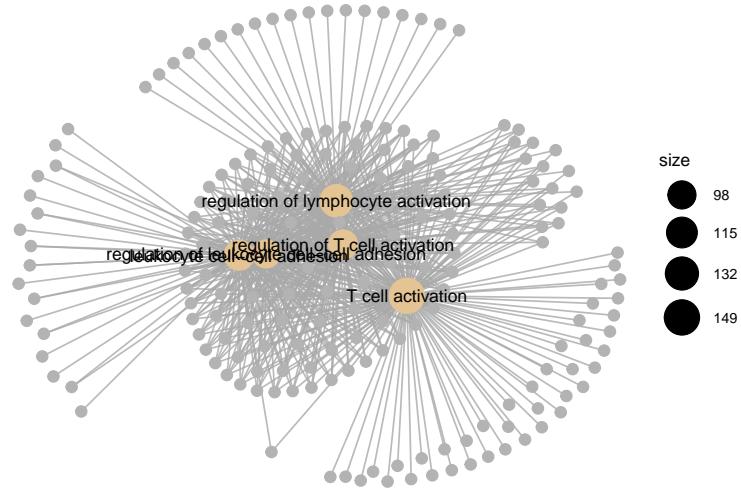


Figure 9: Xarxa de gens SFI-ELI

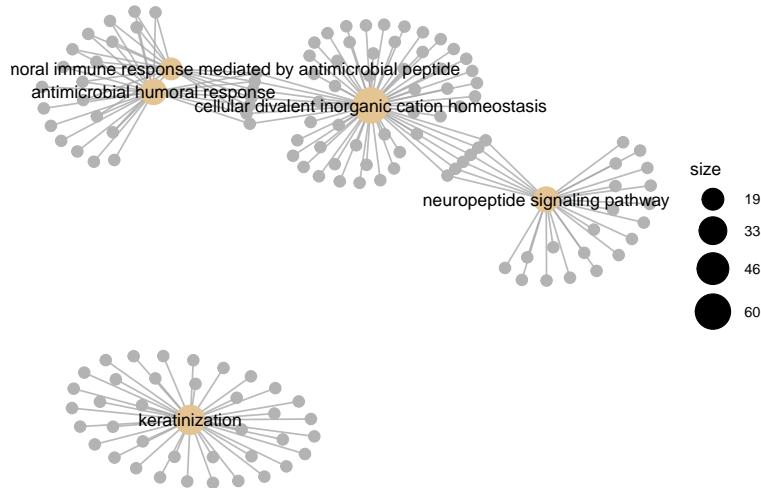


Figure 10: Xarxa de gens NIT-ELI

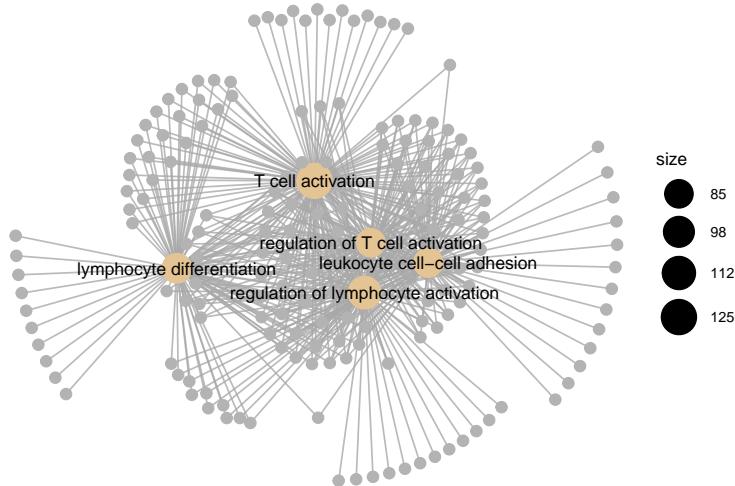


Figure 11: Xarxa de gens SFI-NIT

Seguim corroborant que SFI-ELI i SFI-NIT fan referència a limfòcits i cèl·lules T, mentre que NIT-ELI té una barreja de funcions.

Les comparacions que impliquen el tractament SFI mostren un conjunt de gens interconectats en diferents funcions, relacionades amb el cas d'estudi, que semblen mostrar que es produeix un efecte sobre el que s'espera.

## Discusió

D'aquest estudi s'han pogut extreure dades que podrien ser d'interès, sobretot del apartat de classificació de gens, però s'hauria de tenir enc compte que tansols es una mostra de 30 casos. El disseny d'experiment a l'hora de la presa de dades no sembla ser l'òptim, 10 mostres de cada grup potser no són suficients, i extreure a l'atzar 10 mostres de 236 fa que depenen de la seed es puguin obtenir resultats molt diferenciats. Idílicament un estudi d'aquest tipus obtindria resultats més fiables si s'agafessin totes les mostres possibles (tot i que

suposaria un consum computacional elevat), tot tenint en compte que aquestes haurien de ser simètriques, no és gaire útil tenir 236 mostres “control” si tansols en tenim 14 de tractament.

## Apèndix

El codi d’execució, així com al resta de fitxers descarregats i generats es poden trobar a: [[https://github.com/Adria-FerDi/ADO\\_PEC2](https://github.com/Adria-FerDi/ADO_PEC2)]

## Referències

- Zhu, A., Ibrahim, J.G., Love, M.I. (2018) Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences.
- Guangchuang Yu, Li-Gen Wang, Guang-Rong Yan, Qing-Yu He. DOSE: an R/Bioconductor package for Disease Ontology Semantic and Enrichment analysis. Bioinformatics 2015, 31(4):608-609.
- Guangchuang Yu, Li-Gen Wang, Yanyan Han, Qing-Yu He. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology. 2012, 16(5):284-287.
- Yu G, Wang L, Han Y, He Q (2012). “clusterProfiler: an R package for comparing biological themes among gene clusters.” OMICS: A Journal of Integrative Biology, 16(5), 284-287. doi: 10.1089/omi.2011.0118.
- Love MI, Anders S, Huber W (2020). “Analyzing RNA-seq data with DESeq2”.