

Prova d’Avaluació Continuada 2. Anàlisi de Dades Òmiques

Joan Serrano Quílez

06 June, 2020

Índice

1. Resum	2
2. Objectius	2
3. Materials i mètodes	2
3.1. Dades i disseny experimental	2
3.2. Mètodes i eines emprades	2
3.2.1. Procediment general d’anàlisi	2
3.2.2. Software emprat	3
3.3. Descripció pas a pas	3
3.3.1. Obtenció pseudoaleatòria de les dades	3
3.3.2. Preprocesament de les dades	3
3.3.3. Identificació de gens diferencialment expressats	3
3.3.4. Anotació dels resultats	4
3.3.5. Agrupació de les mostres	4
3.3.6. Anàlisi de significació biològica	4
4. Resultats	4
4.1. Normalització i filtratge	4
4.2. Agrupació de les mostres	5
4.3. Descobriment de gens diferencials	6
4.4. Anàlisi de significació biològica	8
5. Discusió	11
6. Conclusió	11
7. Apèndix	11
8. Referències	11

L’arxiu executable d’aquest treball és pot trobar a github a l’adreça web: https://github.com/joaseki/PEC2_ADO

1. Resum

En el present estudi d'expressió gènica mitjançant RNA-Seq s'estudia el comportament de dos tractaments diferents, SFI i ELI front a l'absència de tractament (NIT); amb l'objectiu de comprovar si hi ha diferències entre els diferents grups. Després de triar 10 mostres aleòries de cada grup i dur a terme els anàlisis necessaris es comprova com no hi ha pràcticament separació entre les mostres SFI i NIT, mentre que ELI és diferent a aquestes dues, activant rutes de senyalització limfoide.

2. Objectius

L'objectiu d'aquest estudi utilitzant dades de RNA-seq és el d'analitzar i identificar possibles efectes d'infiltració en tiroides, bé amb infiltrats locals o amb extensius en limfòcits. Així podem dividir l'objectiu com a doble:

- D'una banda saber si cadascun dels tractaments resulta eficaç pels pacients.
- Saber si existeixen diferències entre els dos tractaments esmentats.

3. Materials i mètodes

3.1. Dades i disseny experimental

Les dades emprades han estat proporcionades pel professor i provenen de dades d'expressió obtingudes per **RNA-Seq**. Es troben en un parell d'arxius, **targets.csv** que conté la informació de cada mostra, sobre el seu grup i d'altres característiques. D'altra banda, **counts.csv** que conté els contejos per a cada geni mostra. Els arxius originals contenen un total de 292 mostres, però de manera pseudoaleatòria s'extreuen 10 mostres de cada grup, com s'explica a l'apartat següent, per tant se'ns queden un total de 30 mostres distribuïdes de la següent manera:

- 10 mostres de teixits no infiltrats (a partir d'ara, **NIT**)
- 10 mostres d'infiltrats locals petits (a partir d'ara, **SFI**)
- 10 mostres d'infiltrats extensius limfoides (a partir d'ara, **ELI**)

Per tant, l'experiment que tenim davant és del tipus de **comparació de classes**, ja que volem veure com canvia l'expressió dels gens entre els diferents grups, tal com s'ha dit als Objectius. Pel que fa al disseny experimental, tindria només un factor en aquest cas (A) que seria el tipus de tractament, d'*efectes fixes** (ja que només s'estudien aquests), amb 3 nivells (NIT, SFI i ELI). Podríem fàcilment representar aquest disseny experimental amb l'equació:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

On μ seria la mitjana general per a un gen determinat, α_i és l'efecte de cadascun dels tres tipus de tractament i , ϵ_{ij} els errors aleatoris per a cada tractament i i rèplica j . Per tant, y_{ij} és l'expressió per a un gen determinat en una observació i grup determinat, que per tant, anirà determinat pel grup a què pertany.

Una vegada hem establert quin hauria de ser el disseny experimental, procedim a assenyalar quin haurien de ser els contrastos d'interès:

- α_{NIT} vs α_{SFI} : per a veure si el tractament amb SFI té un efecte sobre no tractats;
- α_{NIT} vs α_{ELI} : per a veure si el tractament amb ELI té un efecte sobre els no tractats i, per últim;
- α_{SFI} vs α_{ELI} : per a comprovar si ambdós tractaments mostren diferències entre sí.

3.2. Mètodes i eines emprades

3.2.1. Procediment general d'anàlisi

El *work-flow* que s'ha fet servir és molt semblant al trobat a l'enunci de la PAC:

- 1. Obtenció pseudoaleatòria de les dades
- 2. Preprocesament de les dades: filtratge i normalització

- 3. Identificació de gens diferencialment expressats
- 4. Anotació de resultats
- 5. Agrupació entre mostres
- 6. Anàlisi de significació biològica

3.2.2. Software emprat

Durant aquest estudi, el *software* que s'ha decidit fer servir és **R**, en concret s'ha emprat el paquet **DESeq2**, que ens permet fer un anàlisi complet de les dades, obtenint la seva significació biològica.

3.3. Descripció pas a pas

3.3.1. Obtenció pseudoaleatòria de les dades

En primer lloc, s'ha procedit a l'extracció aleatòria de les dades. Tanmateix, no s'ha fet de manera totalmente aleatòria, sinó que amb la fi de poder traballar sempre amb les mateixes dades i que aquestes no canviessin massa, s'ha fet un *set.seed()* (fincant com a número el meu NIF), de manera que les dades triades serien, en tot cas, *pseudoaleatòries*. Primer he creat un nou *dataframe* per a cada grup, i dintre d'ell, amb *sample* triem 10 d'elles segons el *seed*, després unim els *dataframe* en un de nou, que anomeno *sampled_targets*. Amb aquest, faig un subset del document de *counts*. El codi emprat es pot veure a l'arxiu .Rmd.

3.3.2. Preprocesament de les dades

En aquest cas, com s'ha dit amb anterioritat, el paquet que s'emprarà durant el present estudi és el **DESeq2** per la seva facilitat d'utilització i la seva versatilitat. En primer cas, el que s'ha de fer és saber de quin tipus de dades partim, i les nostres són **dataMatrix** o matriu de dades, per la qual cosa, haurem d'utilitzar la funció *DESeqDataSetFromMatrix*, emprant un disseny experimental que depengui del grup de tractament (*Group*), el qual serà el *dds*.

Un cop tenim aquest objecte ben creat, hem de procedir al següent pas, que és el **prefiltratge**, en què eliminarem aquelles observacions que continguin un nombre de counts massa baix, perquè la variabilitat no tindrà perquè deure's a variabilitat biològica. Un bon punt de tall serien aquelles observacions amb menys de 10 *counts*.

El pas següent de cara a fer les comprovacions inicials és fer una normalització que després emprarem. Les dues més emprades soLEN ser la *VST* i la *rlog* les quals es pot veure a la Figura 1.

Cadascun té els seus avantatges i inconvenients, com per exemple que la transformació per *rlog* no va massa bé per grups grans de mostres (en tenim 30).

Ens quedarem amb la *VST* principalment per conveniència amb el nombre de mostres i el fet de ser una transformació molt més ràpida.

3.3.3. Identificació de gens diferencialment expressats

Seguint amb l'objecte *dds* que hem obtingut del pas anterior, ja filtrat. Amb aquest el que fem és un **DESeq**, amb què obtindrem un Data Set amb el format *DESeq*. Obtenim els resultats i els fiquem en objectes diferents per a cada comparació.

També podem guardar els resultats, filtrant segons els paràmetres que ens convinguin, com per exemple, tots els que tinguin un *p*-*valor* inferior a 0.1.

Un cop tinguts aquests resultats ja es podrien exportar, tanmateix, preferim esperar fins que els gens estiguin anotats, és a dir, que podem saber el nom dels gens a què fa referència.

3.3.4. Anotació dels resultats

Amb la fi d'anotar els resultats i poder identificar quins són els gens diferencialment expressats que s'han obtingut (en els objectes *res*), emprarem el paquet **AnnotationDbi**. Com a base de dades de referència emprarem **org.Hs.eg.db**, que ens permet convertir els IDs dels gens procedents d'*Ensembl* a el que vulguem. Per conveniència hem triat fer el canvi als *símbols*, ja que és el nom pel que solen ser més reconeguts pels investigadors al seu camp.

3.3.5. Agrupació de les mostres

Després de tenir les dades de l'anotació, podem dur a terme l'agrupació de les mostres, mitjançant diverses tècniques. Com pot ser creant una matriu de distàncies i veure com agrupen els diferents grups, com es pot comprovar a la secció de resultats, seguint sempre la normalització per VST, amb el paquet *pheatmap*.

Una altra manera de veure les diferències seria la de fer un *heatmap* en què ens agrupa les mostres creant unes noves variables, les components, a partir de tots els gens; el qual podem fer amb la funció *plotPCA*.

Amb **genefilter**, a partir dels resultats de la normalització VST, també podem seleccionar, per exemple, els 20 més diferencialment expressats entre els tres grups, i veure com aquests es classifiquen.

D'altra banda, per a cada comparació feta també podem fer *plots MA*, però abans convertirem les dades originals dels *dds* per a poder visualitzar-los millor amb el mètode *apeglm* [ref] per a poder eliminar possibles sorolls de fons que puguin interferir al gràfic.

D'altra banda, també farem un gràfic de Venn en què compari quins són significatius que canvien. Alternativament, exportarem els 100 més significatius de cada comparació com a *html* separat per a cada comparació amb el paquet *ReportingTools*.

3.3.6. Anàlisi de significació biològica

Un cop ja tenim els gens diferencialment expressats per a cada comparació, podem passar a fer un enriquiment dels processos més repetits. Aquest procediment el durem a terme amb el paquet **clusterProfiler** degut a la seva versatilitat, permetent-nos buscar patrons repetits segons diferents categories de la GO. Amb l'ajuda d'*enrichplot* farem els gràfics necessaris per a cada comparació.

4. Resultats

4.1. Normalització i filtratge

Primer de tot, passem a veure el resultat de les diferents transformacions possibles:

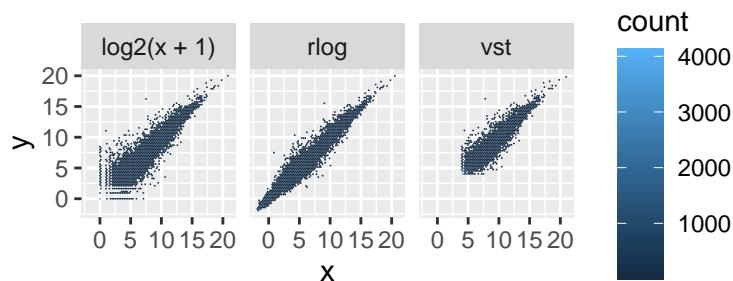


Figura 1: Diferents transformacions efectuades a les mostres

Podem comprovar com les dues (*rlog* i *vst*) són capaces d'arreglar la gran variabilitat que hi ha a les mostres amb contejos baixos. Però els contejos són reduïts clarament a la transformació VST, que és la que, com hem dit anteriorment, agafarem per a fer l'estudi.

4.2. Agrupació de les mostres

D'altra banda, un cop trobats els gens diferencials, podem veure com s'agrupen les mostres fent un anàlisi de components principals (o PCA):

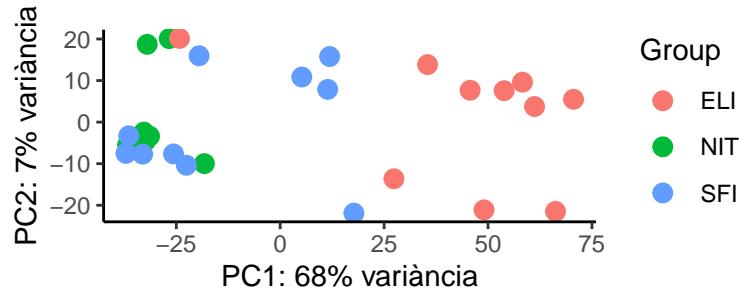


Figura 2: Anàlisi de components principals per als tres grups d'estudi

És aquí on podem comprovar ja que hiha una gran diferència segons que mirem. Les mostres pertanyent al grup no tractat *NIT* i les *SFI* no es poden separar i, no obstant això, sí que n'hi ha una separació separació entre aquestes dues i el grup *ELI*. Això seria un primer indicí que el tractament d'*ELI* seria molt més efectiu.

També podem fer d'altra banda comparacions amb *heatmaps* diferents:

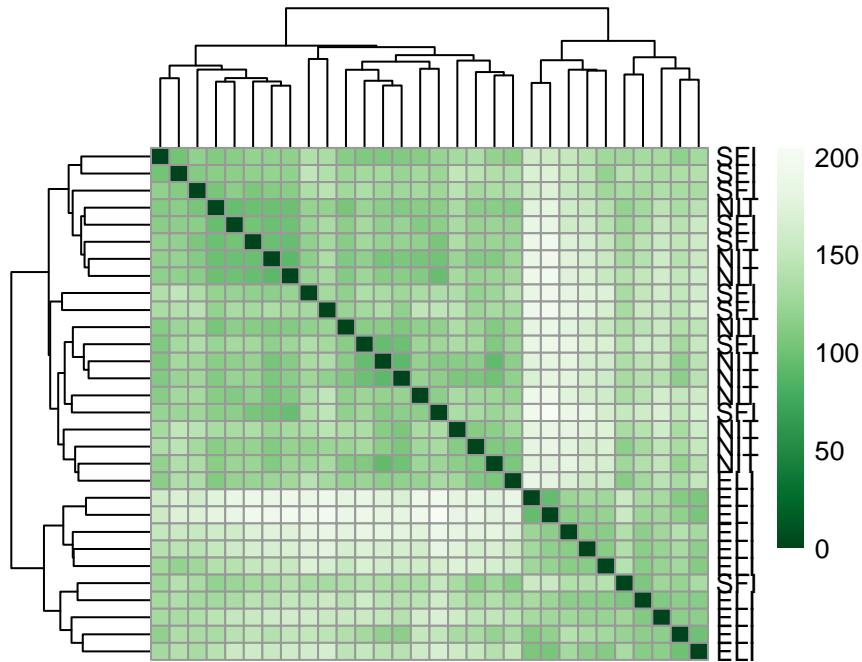


Figura 3: Heatmap de les distàncies entre mostres segons la matriu de distàncies

Aquest gràfic mostra amb claredat com hi ha una clara diferència visual d'agrupació entre el grup *ELI* (amb algun intrús del grup *SFI*) i la resta. Tanmateix, les altres dos grups (*SFI* i *NIT*) són incapços de distingir-se entre sí.

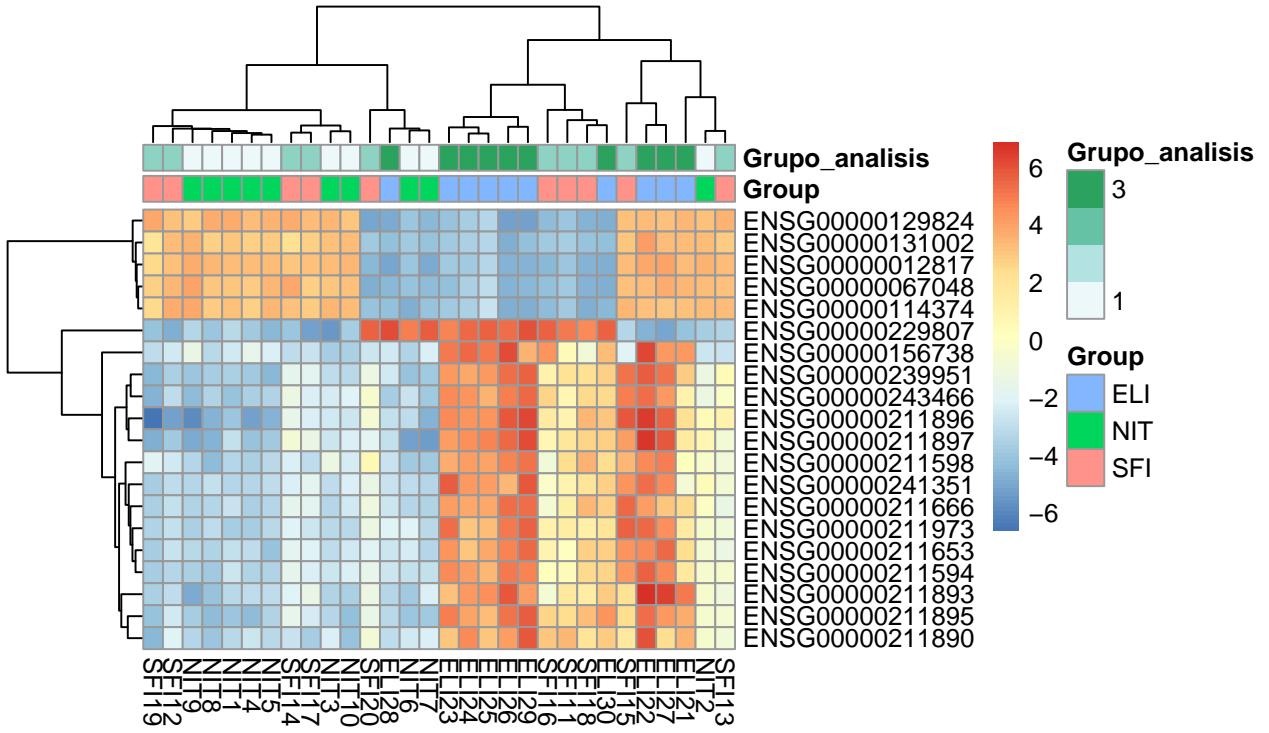


Figura 4: Heatmap amb les transformacions VST, amb els 20 gens més significatius

Sembla que hi ha patrons però que responden a alguna altra variable.

Per a cadascuna de les comparacions podem veure ara els *MA plot* per a veure com és la variació de l'expressió gènica als 3 casos.

4.3. Descobriment de gens diferencials

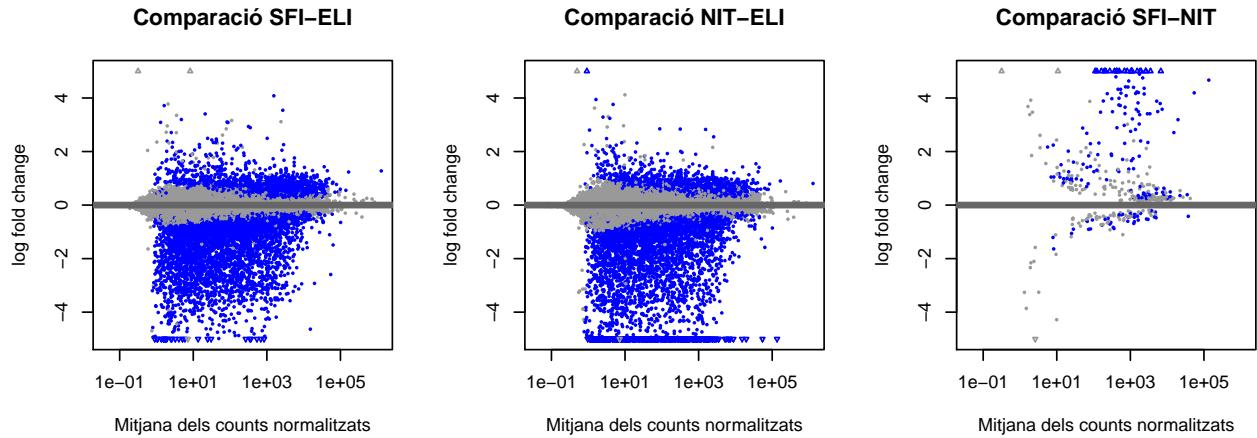


Figura 5: MA plots mostrant les diferències d'expressió gènica per a cadascuna de les 3 comparacions

És aquí on podem veure on està la diferència principal. Pel que fa a les comparacions SFI-ELI i NIT-ELI es veu com hi ha molts gens que es troben diferents, amb una clara inclinació cap a la sobreexpressió en les

mostres ELI. Tanmateix el contrari passa amb la comparació SFI-NIT, en què molt pocs gens estan diferents, la qual cosa es pot explicar per la baixa separació que veiem als gràfics anteriors.

Ho podem comprovar fent una comparació de comparacions amb un diagrama de Venn:

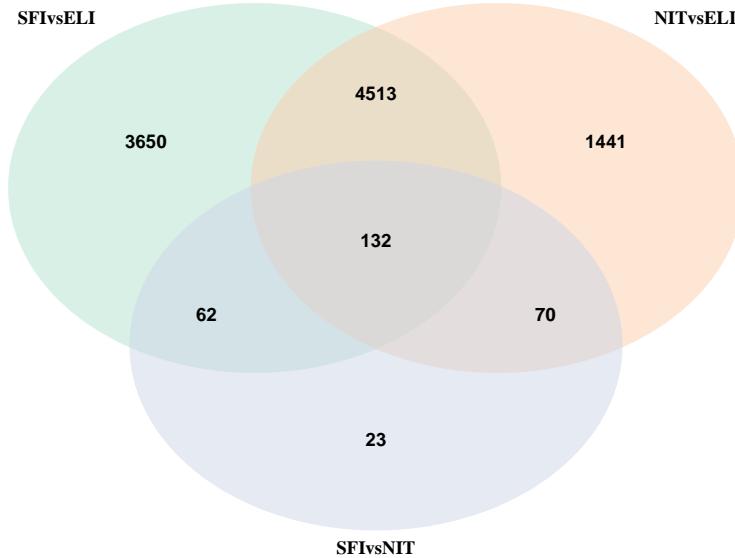


Figura 6: Diagrama de Venn representant els gens significatius ($p\text{-adj} < 0.1$) entre les tres comparacions

Clarament els gens que canvién significativament en la comparació SFI-NIT són molt pocs, el que pot significar què el tractament per SFI no té efecte sobre els no tractats. Cosa corroborada en veure que en les comparacions amb el tractament amb ELI, hi ha un gran canvi tant amb el tractament amb SFI com els no tractats.

Per exemple, podem veure com és la taula de resultats, que s'ha obtingut per la comparació SFI-ELI:

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol
ENSG00000204282	905.6322	-3.129541	0.2891076	-10.824831	0	0	TNRC6C-AS1
ENSG00000270164	203.8483	-3.676989	0.3519667	-10.446982	0	0	LINC01480
ENSG00000205744	1599.5176	-2.652373	0.2656252	-9.985396	0	0	DENND1C
ENSG00000185404	1092.6380	-1.658340	0.1755190	-9.448210	0	0	SP140L
ENSG00000068831	3468.8224	-3.063995	0.3265512	-9.382894	0	0	RASGRP2

Amb gens com, TNRC6C-aS1, LINC01480, DENND1C o RASGRP2 són els que surten més significatius per la comparació SFI-ELI.

Amb la comparació NIT-ELI, és a dir, no tractats amb el tractats amb ELI, en tindríem els següents 5:

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol
ENSG00000211899	7308.5316	-7.167011	0.6155085	-11.64405	0	0	NA
ENSG00000269404	395.0225	-7.228375	0.6254229	-11.55758	0	0	SPIB
ENSG00000205744	1599.5176	-2.977339	0.2659657	-11.19445	0	0	DENND1C
ENSG00000167483	839.0444	-6.922085	0.6185490	-11.19084	0	0	NIBAN3
ENSG00000083454	1066.9653	-6.437342	0.5773098	-11.15058	0	0	P2RX5

El primer de tots no està descrit, però la resta sí, i a primera vista semblen als de la primera comparació. Podem mirar també la comparació SFI-NIT, en la qual veiem que els més significatius no estan ni tan sols anotats per la base de dades de org.Hs.eg.org. Cosa que es repeteix al llarg de la comparació.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	symbol
ENSG00000211950	500.0219	8.469630	1.1835730	7.155985	0	0e+00	NA
ENSG00000211942	163.3674	9.889103	1.4156858	6.985380	0	0e+00	NA
ENSG00000239951	6735.1041	6.120618	0.8872957	6.898059	0	0e+00	NA
ENSG00000243264	122.9929	7.921815	1.1989846	6.607104	0	2e-07	NA
ENSG00000211595	1114.4988	6.040136	0.9497942	6.359416	0	9e-07	NA

4.4. Anàlisi de significació biològica

Una vegada fet un cop d'ull a alguns dels gens, el més important ara és veure en quines rutes pot estar implicats aquells gens que més varien. Ho farem amb un *enrichment analysis* dut a terme amb el paquet *clusterProfiler*. Podem veure els *dotplots* per a cadascuna de les comparacions, en ser molt útils perquè pots veure tant la significació com el nombre de gens que hi han alterats en un procés.

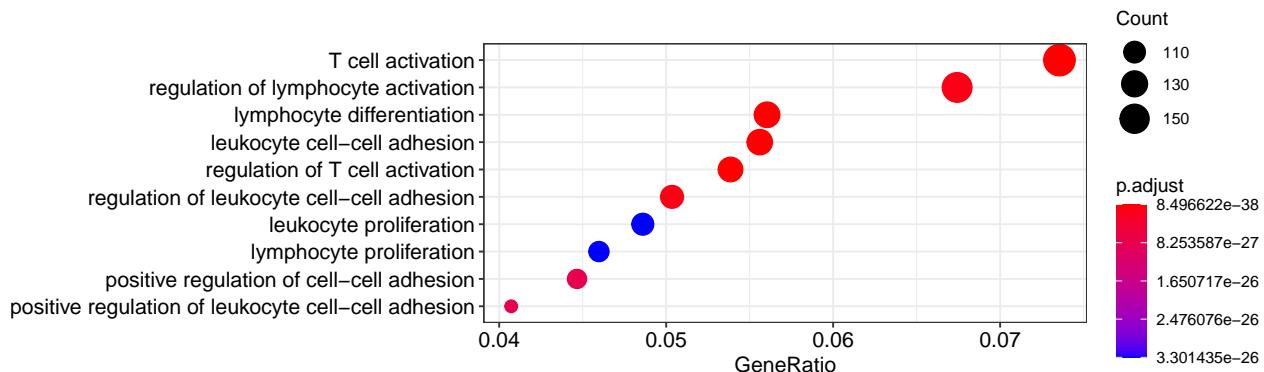


Figura 7: Dotplots per la comparació SFI-ELI

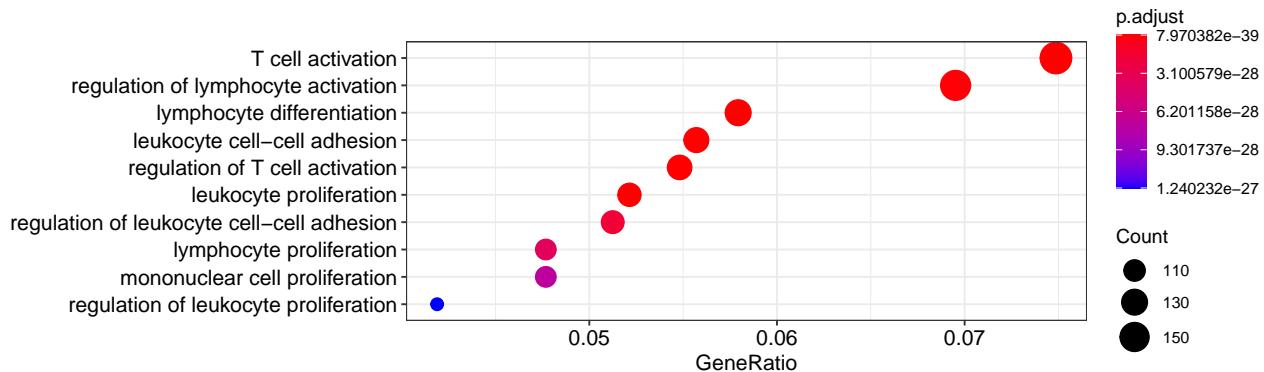


Figura 8: Dotplots per la comparació NIT-ELI

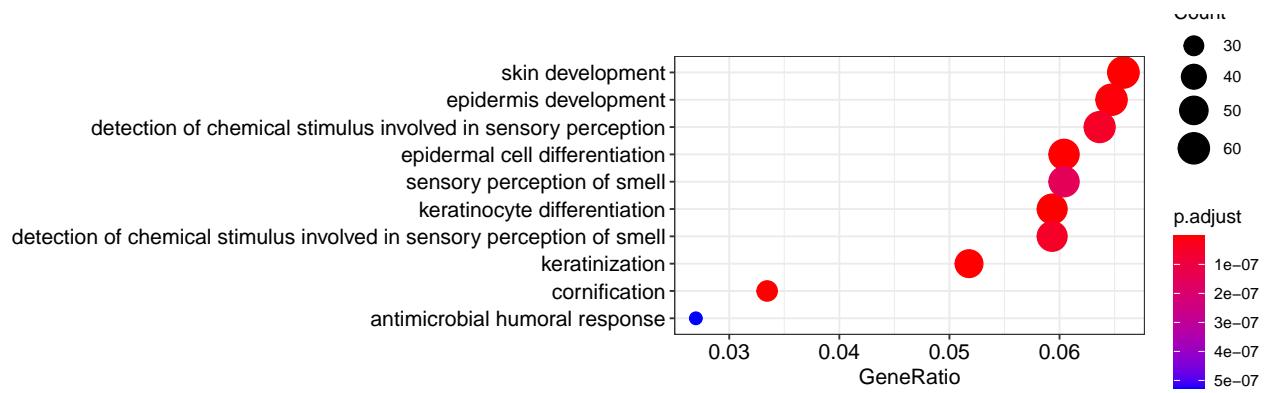


Figura 9: Dotplots per la comparació SFI-NIT

En aquest cas, veiem dues coses:

- La comparació SFI-ELI i la NIT-ELI són molt similars i ambdues mostren un mateix tipus de gens alterats, els relacionats amb el sistema immune i l'activació de limfòcits, cosa que es persegueix amb el tractament.
- Pel que fa a la comparació SFI-NIT, veiem que els gens que s'alteren són molt diferents, gairebé tots relacionats amb la pell i no pas amb l'activació immunitària.

Podem veure una xarxa d'igual manera que representi quins són els gens diferencials, la qual cosa ens pot resultar de molta utilitat per a trobar relacions entre gens compartits entre diversos processos (en les primeres comparacions ometo el nom dels gens, perquè limita la visibilitat)

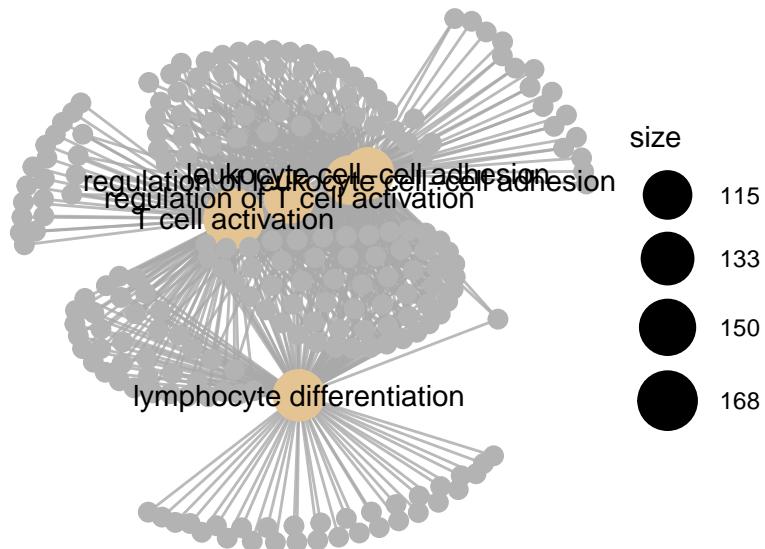


Figura 10: Xarxa de gens per les comparació SFI-ELI

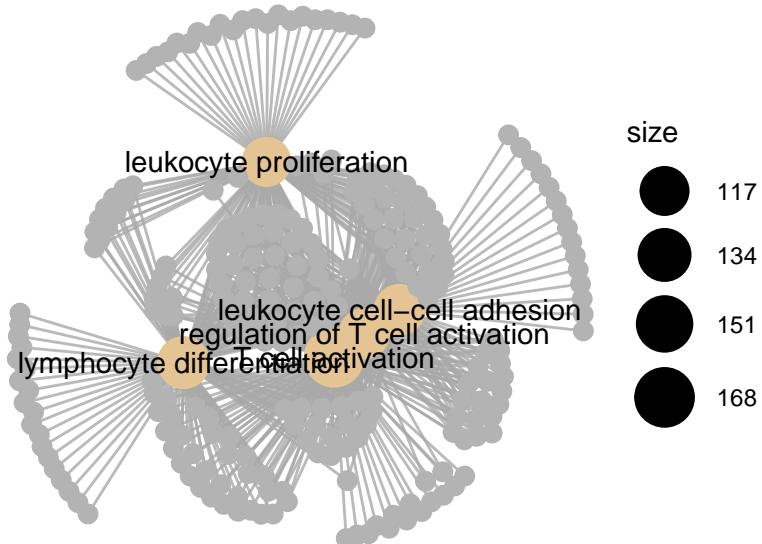


Figura 11: Xarxa de gens per les comparació NIT-ELI

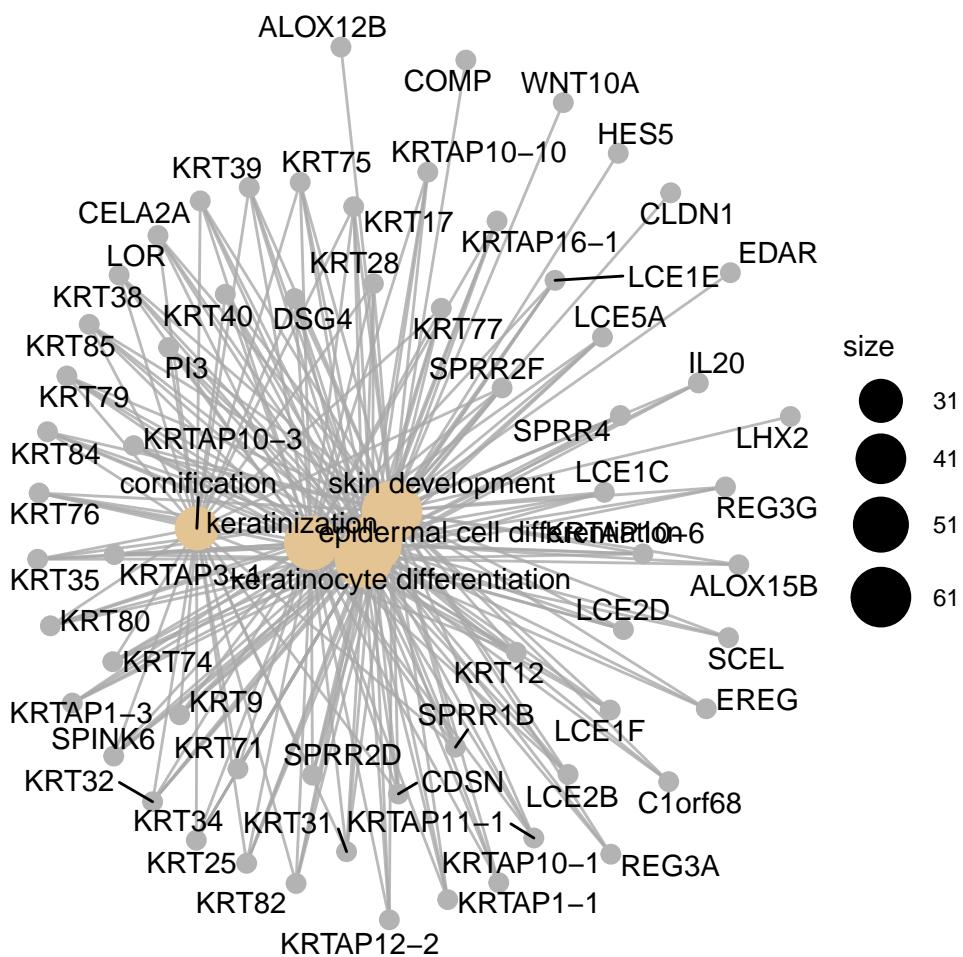


Figura 12: Xarxa de gens per les comparació SFI-NIT

S'acaba de comprovar com els gens que s'activen quan es tracta amb SFI respecte al no tractat, no té res a veure amb els altres dos, que estan involucrats en resposta per limfòcits.

5. Discusió

S'ha dut a terme l'estudi con molt poques mostres (10), tenint en compte la poca capacitat de agrupació que es tenen amb aquestes, potser seria molt més recomanable agafar-ne més, ja que les teníem disponibles, però també es gastaria molta capacitat computacional. D'altra banda, destacar la diferència de grandària tan enorme entre els diferents grups originals. Això fa que segons el seed que agafem les dades del grup no tractat puguin ser molt diferents les unes de les altres, mentre que al grup ELI, que només en són 14, les diferències seran mínimes. Aquesta tria pot afectar notablement a l'anàlisi. De fet, abans d'aconseguir que el seed fos el mateix per la consola i el RMarkdown, l'agrupació de mostres en cada cas era different, en algun cas he vist més separació entre els grups NIT-SFI i en d'altres no, tot i que ELI sempre resulta el més different. En tot cas, segons les que he obtingut, els resultats mostrarien que el tractament en forma de SFI no és eficaç (pel que fa a la resposta gènica) i que mitjançant l'ELI sí que hi hauria un canvi pertinent (tant amb les NIT com amb les SFI).

6. Conclusió

Sembla que hi ha diferències entre els grups NIT-ELI i SFI-ELI, però no així les que es podrien esperar entre SFI-NIT.

7. Apèndix

L'arxiu .Rmd executable que s'ha emprat per fer aquest informe es pot trobar a GitHub al repositori https://github.com/joaseki/PEC2_ADO

8. Referències

- Love MI, Anders S, Huber W (2020). “Analyzing RNA-seq data with DESeq2”. Disponible a: <http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html>
- Love MI, Anders S, Kim V, Huber W (2015). “RNA-Seq workflow: gene-level exploratory analysis and differential expression.” F1000Research. doi: 10.12688/f1000research.7035.
- Pagès H, Carlson M, Falcon S, Li N (2020). AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. R package version 1.50.0.
- Carlson M (2018). org.Hs.eg.db: Genome wide annotation for Human. R package version 3.7.0.
- Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics. 2019;35(12):2084-2092. doi:[10.1093/bioinformatics/bty895](https://doi.org/10.1093/bioinformatics/bty895)
- Huntley MA, Larson JL, Chaivorapol C, Becker G, Lawrence M, Hackney JA, Kaminker JS (in press). “ReportingTools: an automated result processing and presentation toolkit for high throughput genomic analyses.” Bioinformatics. doi: 10.1093/bioinformatics/btt551.
- 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.