Prova d'avaluació continuada 2

Anàlisi de dades òmiques

Núria Mercadé Besora

21-12-2023

Taula de continguts

Introducció i objectius

La dermatitis per contacte al · lèrgic (DCA) es caracteritza per una inflamació local a la pell. Al · lèrgens comuns inclouen metalls, perfums, colorants i preservatius. Quan una persona al · lèrgica està exposada a l'al · lergen, les quimioquines recluten limfòcits T específics a la pell, on les cèl · lules experimenten una proliferació extensa (Rustemeyer et al. (2020)). Les cèl · lules T activades produeixen i alliberen posteriorment alts nivells de citocines, i així va provocar un procés inflamatori que condueix a l'èczema.

Pedersen et al. (2007) van dur a terme el primer estudi dels canvis en l'expressió genètica des de l'exposició a l'al · lergen fins a l'aparició de l'èczema en DCA. Van dur a terme un estudi de microarrays d'ADN on van participar 12 dones (5 al · lèrgiques al níquel i 5 no al · lèrgiques). Es va exposar la pell que cobreix les natges superiors a pegats amb un contingut de sulfat de níquel del 5%. Tots els participants van ser exposats a 3 pegats amb temps d'exposició diferents: 7, 48 i 96 hores. Es va realitzar una biòpsia cutània per a l'anàlisis de microarrays abans de l'exposició, i just en acabar l'exposició a cada un dels pegats de níquel. Tots els al · lèrgics al níquel van reaccionar amb èczema a les 48 i 96h, en canvi, no es va desenvolupar èczema després de les 7h d'exposició. En el grup control (no al · lèrgics) no es va veure èczema en cap moment.

De les 48 biòpsies, 34 van resultar en RNA de suficient qualitat i quantitat per procedir amb l'anàlisi de microarrays. Les dades resultants s'han dipositat al repositori públic Gene Expresion Omnibus (GEO) amb l'identificador "GSE6281".

L'objectiu d'aquest treball, és utilitzar les dades de seqüenciació d'aquest estudi per respondre les preguntes següents:

1) Hi ha diferències en l'expressió de gens entre pacients de DCA i el grup control abans de l'exposició?
2) Hi ha diferències en l'expressió de gens entre pacients de DCA i el grup control a les 7h d'exposició, tot i no haver-hi diferències visibles (no hi ha èczema encara)?
3) Quines diferències en l'expressió genètica es produeixen entre al · lèrgics i no al · lèrgics al cap de 48 hores d'exposició, un cop l'èczema ja és visible? Hi ha les 96h d'exposició?

Material i mètodes

L'anàlisi de dades es durà a termen en R, versió 4.2.2, fent ús de paquets d'anàlisis de dades bioinformàtiques del repositori de Bioconductor.

Font de dades

Aquest estudi conté 34 mostres, 18 pertanyen al grup d'al · lèrgics al níquel (AN) i 16 al grup control (C). 7 de les mostres són a temps 0h (abans de l'exposició), 7 més són a les 48h, i a les 7h i 96h disposem de 10 mostres per cada temps d'exposició. L'identificador serie d'aquest conjunt de dades en el portal GEO és GSE6281, i el del conjunt de dades és GDS2935. En el portal GEO del dataset a part d'informació sobre com s'han recollit les mostres, també trobem informació sobre la seqüenciació: s'ha utilitzat la plataforma Affymetrix Human Genome U133 Plus 2.0 Array (HG-U133_Plus_2).

Lectura i preprocessat de dades

Accedirem a les dades publicades a GEO fent ús del paquet `GEOquery` de Bioconductor. Aquest paquet permet de manera senzilla descarregar-se dades de GEO i convertir-les en objectes de classe

expressionSets, típics per l'anàlisi de dades de seqüenciació. Per fer-ho, necessitem l'identificador del dataset o be de la serie. A continuació és mostra com s'ha accedit a les dades d'estudi i s'han convertit a expressionSet fent servir funcions de `GEOquery`.

```
gds <- getGEO("GDS2935")
eset <- GDS2eSet(gds, do.log2 = FALSE)
```

Anotació

Bioconductor conté una gran llista de paquets d'anotacions que permeten convertir els identificadors que utilitza cada plataforma de microarray a una anotació estándard. En el nostre cas, sabem que les dades s'han obtingut fent ús de microarrays Affymetrix Human Genome U133 Plus 2.0 Array (HG-U133_Plus_2). Busquem a la llista quin paquet conté les anotacions per aquesta plataforma, i trobquem que és `hgu133plus2.db`. Afegim aquesta infromació al expressionSet de la següent manera:

```
annotation(eset) <-"hgu133plus2.db"
```

Filtratge

Tot seguit utilitzarem el paquet de Bioconductor `genefilter` per tal d'eliminar categories amb poca variació o amb un senyal baixa de manera consistent en les mostres, ja que eliminar-les millorarà l'anàlisi de les dades Bourgon, Gentleman, and Huber (2010). A més, també exclourem de les dades aquelles que no estiguin anotades. Les comandes per obtenir l'expressionSet filtrat són les següents:

```
filter_result <- nsFilter(eset)
eset_filtered <- filter_result$eset

# Create variables with the slots pData and exprs
pData <- pData(phenoData(eset_filtered))
exprs <- exprs(eset_filtered)

# Visualize
pData %>% glimpse()
```

```
Rows: 34
Columns: 6
$ sample        <chr> "GSM144434", "GSM144437", "GSM144441", "GSM144444", "GSM~
$ time          <fct> control, control, control, control, control, control, co~
$ disease.state <fct> nickel allergy, nickel allergy, nickel allergy, nickel a~
$ agent         <fct> unexposed, unexposed, unexposed, unexposed, unexposed, u~
$ individual    <fct> patient2, patient4, patient5, patient6, control2, contro~
$ description   <chr> "Value for GSM144434: Skinbiopsy_0hoursnickel_nickel_all~
```

```
head(exprs)
```

```
           GSM144434 GSM144437 GSM144441 GSM144444 GSM144362 GSM144371 GSM144376
204639_at    6.24106   6.23889   6.57046   6.09209   6.78104   7.27359   6.27430
```

```
212607_at   5.73765   6.06047   5.67713   6.22969   7.42071   7.20218   6.68921
207078_at   5.96469   5.80767   6.45481   6.18406   5.87972   5.12872   5.76547
222161_at   3.35121   3.75537   3.33554   3.90450   3.98788   4.20155   3.70140
228647_at   6.56896   5.72418   6.83687   6.87060   6.83564   7.32018   7.05956
236514_at   4.74197   4.38581   4.17624   4.26629   4.17583   4.58546   5.21462
          GSM144435 GSM144438 GSM144442 GSM144445 GSM144447 GSM144309 GSM144366
204639_at   6.18129   7.31989   6.94369   6.43897   7.04201   6.22802   7.26160
212607_at   6.89481   6.50499   6.02609   5.78982   6.13465   6.93979   7.24565
207078_at   5.54353   6.01920   6.14915   5.80827   5.61342   6.02845   5.71553
222161_at   4.07895   3.77239   3.68202   3.44813   3.61618   3.99273   4.30384
228647_at   6.74953   6.93848   6.88798   6.46994   6.32149   7.10709   6.58335
236514_at   4.32581   4.20404   5.05790   4.31318   4.43925   4.69639   4.40354
          GSM144368 GSM144372 GSM144375 GSM144432 GSM144439 GSM144448 GSM144311
204639_at   6.78731   7.39886   6.51218   6.07728   5.93750   6.93081   6.69638
212607_at   7.03545   7.47729   6.85316   6.74364   6.28447   6.58367   7.00821
207078_at   6.00195   5.54906   5.76323   4.86630   4.95799   5.54647   5.69940
222161_at   4.08160   3.98707   3.36525   3.30894   2.94997   3.86869   4.16976
228647_at   6.84511   7.40944   7.00648   6.04199   5.06362   6.65158   7.01298
236514_at   4.84374   4.98347   5.02785   3.74830   3.93604   4.09838   4.87305
          GSM144369 GSM144373 GSM144419 GSM144433 GSM144436 GSM144440 GSM144443
204639_at   6.40853   7.06195   6.16773   6.52109   6.49284   6.23763   6.35588
212607_at   6.80333   7.14607   6.95668   6.23826   6.37164   5.43120   6.25941
207078_at   6.16299   5.28838   5.19782   6.13315   5.76370   5.23188   6.06327
222161_at   4.34127   3.97096   3.71251   3.19518   3.51242   3.01180   3.24181
228647_at   6.69136   7.04039   7.13357   6.50229   5.99724   5.31542   6.33774
236514_at   5.07931   4.63761   5.01211   3.98632   4.58163   3.51213   4.83571
          GSM144446 GSM144449 GSM144347 GSM144367 GSM144370 GSM144374
204639_at   6.42046   6.74154   6.17154   7.06811   6.92962   6.51462
212607_at   6.31408   6.83312   7.00294   7.43074   7.09883   7.46260
207078_at   5.35560   5.58896   5.81898   5.75374   5.69426   5.18662
222161_at   3.12971   3.05360   3.90679   4.20301   3.96997   3.25765
228647_at   5.26928   5.94080   7.00367   6.77216   6.45556   7.13930
236514_at   3.89929   3.70493   4.30360   4.35272   4.42966   4.70705
```

Exploració de les dades

Utilitzarem gràfics de diagrames de caixes i l'anàlisis de components principals per comprovar la normalització de les dades, o si bé és necessària alguna transformació.

Selecció de gens diferencialment expressats

Volem comparar les mostres del grup AN amb el C en quatre situacions diferents: abans de ser exposats a níquel (temps 0h), a les 7h d'exposició (temps 7), a les 48h (temps 48) i a les 96h (temps 96). Per fer-ho, utilitzarem models lineals generals i corregirem la variància resultant amb models bayesians empírics per a l'anàlisi de microarrays, tal com descriu Smyth (2004).

Matrius de diseny i contrast

La matriu de diseny descriu com es distribueixen les mostres segons condicions i/o grups experimentals. Definir aquesta matriu és el primer pas per ajustar models lineals. Cada fila representa una mostra, i les

columnes de la matriu dependran de les comparacions que es volen fer. En el nostre cas ens interessa tenir 8 columnes diferenciades corresponents als grups AN i C, en els diferents temps d'exposició: AN_0, C_0, AN_7, C_7, AN_48, C_48, AN_96, i C_96. Els valors de la matriu de disseny seràn 1 si la mostra pertany al grup i condicions indicats a la columna, o 0 si no és així. El codi per crear la matriu és el següent:

```
# informative name:
time <- gsub(" ", "", pData$time)
time <- gsub("control", "0h", time)

disease <- gsub("nickel allergy", "AN", pData$disease.state)
disease <- gsub("non-allergic control", "C", disease)

group <- paste(disease, time, sep="_" )

# design matrix
design  <-  model.matrix(~ 0 + group)
colnames(design) <- gsub("group", "", colnames(design))
rownames(design) <- pData$sample

design
```

	AN_0h	AN_48h	AN_7h	AN_96h	C_0h	C_48h	C_7h	C_96h
GSM144434	1	0	0	0	0	0	0	0
GSM144437	1	0	0	0	0	0	0	0
GSM144441	1	0	0	0	0	0	0	0
GSM144444	1	0	0	0	0	0	0	0
GSM144362	0	0	0	0	1	0	0	0
GSM144371	0	0	0	0	1	0	0	0
GSM144376	0	0	0	0	1	0	0	0
GSM144435	0	0	1	0	0	0	0	0
GSM144438	0	0	1	0	0	0	0	0
GSM144442	0	0	1	0	0	0	0	0
GSM144445	0	0	1	0	0	0	0	0
GSM144447	0	0	1	0	0	0	0	0
GSM144309	0	0	0	0	0	0	1	0
GSM144366	0	0	0	0	0	0	1	0
GSM144368	0	0	0	0	0	0	1	0
GSM144372	0	0	0	0	0	0	1	0
GSM144375	0	0	0	0	0	0	1	0
GSM144432	0	1	0	0	0	0	0	0
GSM144439	0	1	0	0	0	0	0	0
GSM144448	0	1	0	0	0	0	0	0
GSM144311	0	0	0	0	0	1	0	0
GSM144369	0	0	0	0	0	1	0	0
GSM144373	0	0	0	0	0	1	0	0
GSM144419	0	0	0	0	0	1	0	0
GSM144433	0	0	0	1	0	0	0	0
GSM144436	0	0	0	1	0	0	0	0
GSM144440	0	0	0	1	0	0	0	0
GSM144443	0	0	0	1	0	0	0	0

```
GSM144446     0      0     0      1    0     0    0     0
GSM144449     0      0     0      1    0     0    0     0
GSM144347     0      0     0      0    0     0    0     1
GSM144367     0      0     0      0    0     0    0     1
GSM144370     0      0     0      0    0     0    0     1
GSM144374     0      0     0      0    0     0    0     1
attr(,"assign")
[1] 1 1 1 1 1 1 1 1
attr(,"contrasts")
attr(,"contrasts")$group
[1] "contr.treatment"
```

Per altra banda, la matriu de contrastos es basa en la matriu de disseny per definir les comparacions que es volen dur a terme. Cada columna d'aquesta matriu és una comparació, i les files són els grups i/o condicions experimentals. El valors, seràn positius i negatius entre les files que representen que es volen comparar, amb la condició que la seva sigui 0. Així doncs, per a les nostres comparacions, per exemple la primera (AN vs. C a temps 0h) la fila AN_0 tindrà valor 1 en la columna corresponent a aquesta comparació, mentre que la fila C_0 tindrà valor -1. La resta de files tindran valor 0 en aquesta columna ja que no intervenen en la comparació. La matriu de contrastos la construim amb el paquet (limma) de Biocondcutor, tal com es mostra a continuació:

```
cont.matrix <- makeContrasts (
  time0  = AN_0h - C_0h,
  time7  = AN_7h - C_7h,
  time48 = AN_48h - C_48h,
  time96 = AN_96h - C_96h,
  levels = design
)

cont.matrix
```

```
        Contrasts
Levels   time0 time7 time48 time96
  AN_0h      1     0      0      0
  AN_48h     0     0      1      0
  AN_7h      0     1      0      0
  AN_96h     0     0      0      1
  C_0h      -1     0      0      0
  C_48h      0     0     -1      0
  C_7h       0    -1      0      0
  C_96h      0     0      0     -1
```

Estimació del model i selecció de gens

Per tal d'estimar el model lineal general que utilitza models de Bayes empírics farem ús de funcions del paquet `limma`. D'aquest anàlisis obtenim estimadors estadístics com el t-moderat i els p-valors ajustats.

Per seleccionar els gens que mostren significació estadística corregirem els p-valors utilitzant el mètode de Benjamini & Hochberg (BH) que permet controlar el nombre de falsos positius quan es duen a

terme múltiples comparacions. Seleccionem els gens estadísticament significatius, utilitzant un nivell de significació del 5%. A més, utilitzem el paquet d'anotacions per saber de quins gens es tracta.

```
# Estmació del model ----
fit <- lmFit(eset_filtered, design)
fit.cont <- contrasts.fit(fit, cont.matrix)
fit.cont <- eBayes(fit.cont)

# Selecció de gens ----
# anotacions
anotations <- AnnotationDbi::select(hgu133plus2.db,
                                    keys = rownames(exprs),
                                    columns = c("ENTREZID", "SYMBOL"))

# functió per anotar les taules
anotateTable <- function(x) {
  x %>%
    as_tibble() %>%
    dplyr::rename("PROBEID" = "ID") %>%
    dplyr::inner_join(anotations, by = "PROBEID") %>%
    arrange(adj.P.Val)
}

# Top tables anotades per cada comparació
topTab_time0  <- fit.cont %>%
  topTable(number = nrow(fit.cont), coef= "time0") %>%
  anotateTable()
topTab_time7  <- fit.cont %>%
  topTable(number = nrow(fit.cont), coef= "time7") %>%
  anotateTable()
topTab_time48 <- fit.cont %>%
  topTable(number = nrow(fit.cont), coef= "time48") %>%
  anotateTable()
topTab_time96 <- fit.cont %>%
  topTable(number = nrow(fit.cont), coef= "time96") %>%
  anotateTable()

# Selecció gens estadísticament singificiatius per cada comparació
summary.fit <- decideTests(fit.cont, p.value = 0.05)
```

Aquests resultats es reportaran en forma de gràfics com el diagrama de Vann i volcanoplots, a més d'un mapa de color per visualitzar el perfil d'expressió i detectar diferents patrons.

Anàlisis de significació biological

Un cop s'han identificat quins son els gens diferencialment expressats, comencem l'anàlisi de significació biologica per tal d'interpretar els resultats. Durem a terme dos tipus d'analisis, el d'enriquiment (o anàlisi de sobre representació), i l'anàlisi *Gene Set Expresion Analysis* (GSEA).

Anàlisis d'enriquiment

Aquesta anàlisi ens permetrà identificar quins processos biològics estan sobrerepresentats en el nostre llistat de gens que han mostrat significació estadística en l'anàlisi de gens diferencialment expressats. Durem a terme aquesta anàlisi basant-nos en anotacions ontològiques dels gens. Per tenir en compte els múltiples testos utilitzarem BH per ajustar els p-valors. En l'anàlisi a part de diferenciar per comparacions, també diferenciarem si els gens estan sobre-expressats o bé regulats, fixant un canvi mínim en el nivell d'expressió de 0.5 en escala logarítmica en base 2. A continuació es mostra un exemple del codi fet servir per dur a terme aquesta anàlisi en els gens sobre expressats a temps 0.

```
selectedEntrezs <- rownames(subset(topTab_time0, (logFC > 0.5) & (adj.P.Val < 0.05)))
ego <- enrichGO(gene = selectedEntrezs,
                universe = rownames(topTab_time0),
                keyType = "ENTREZID",
                OrgDb = hgu133plus2.db,
                ont = "BP",
                pAdjustMethod = "BH",
                qvalueCutoff = 0.05,
                readable = TRUE)
```

Gene Set Expresion Analysis

El GSEA es basa en el canvi en l'expressió dels gens entre dos grups i/o condicions experimentals. L'anàlisi utilitza el llistat de tots els gens que s'estan estudiant, independentment de si han resultat mostrar diferències en la seva expressió en anàlisis anteriors. Els gens es classifiquen tenint en compte les seves funcions biològiques. Hi ha diferents classificacions, per l'anàlisi nosaltres farem servir base de dades *Kyoto Encyclopedia of Genes and Genomes* (KEGG). En el GSEA s'assignarà puntuacions d'enriquiment als diferents conjunt de gens segons si s'observa una coordinació en la seva sobre-expressió o regulació. S'utilitzen tests de permutació per determinar la significació estadística i s'ajusten els p-valors utilitzant BH per tenir en compte les múltiples comparacions. A continuació es mostra com a exemple el codi utilitzat per fer aquest anàlisis en la comparació a temps 0:

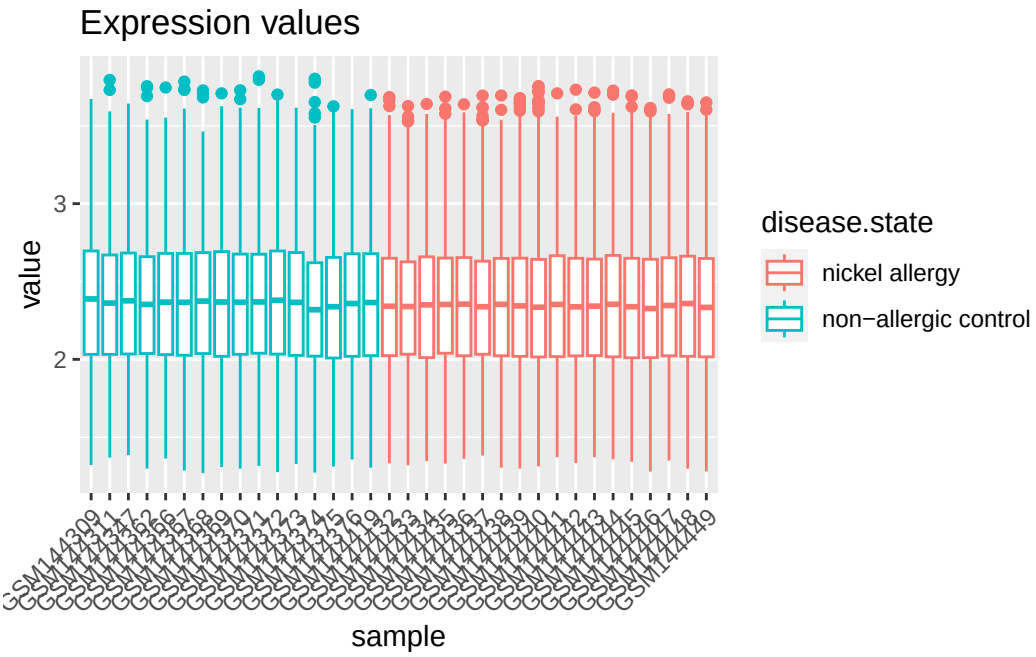
```
# sort by absolute logFC to remove duplicates with smallest absolute logFC
geneList <- topTab_time0[order(abs(topTab_time0$logFC), decreasing = TRUE),]
geneList <- geneList[!duplicated(geneList$ENTREZID), ]

# re-order based on logFC to be GSEA ready
geneList <- geneList[order(geneList$logFC, decreasing = TRUE),]
genesVector <- geneList$logFC
names(genesVector) <- geneList$ENTREZID

# GSEA
set.seed(123)
gseResults <- gseKEGG(geneList = genesVector)

gsea.result <- setReadable(gseResults, OrgDb = hgu133plus2.db,
                           keyType ="ENTREZID")
gsea.result.df <- as.data.frame(gsea.result)
gsea.result.df[,c("Description","setSize","NES","p.adjust")]
```

Resultats

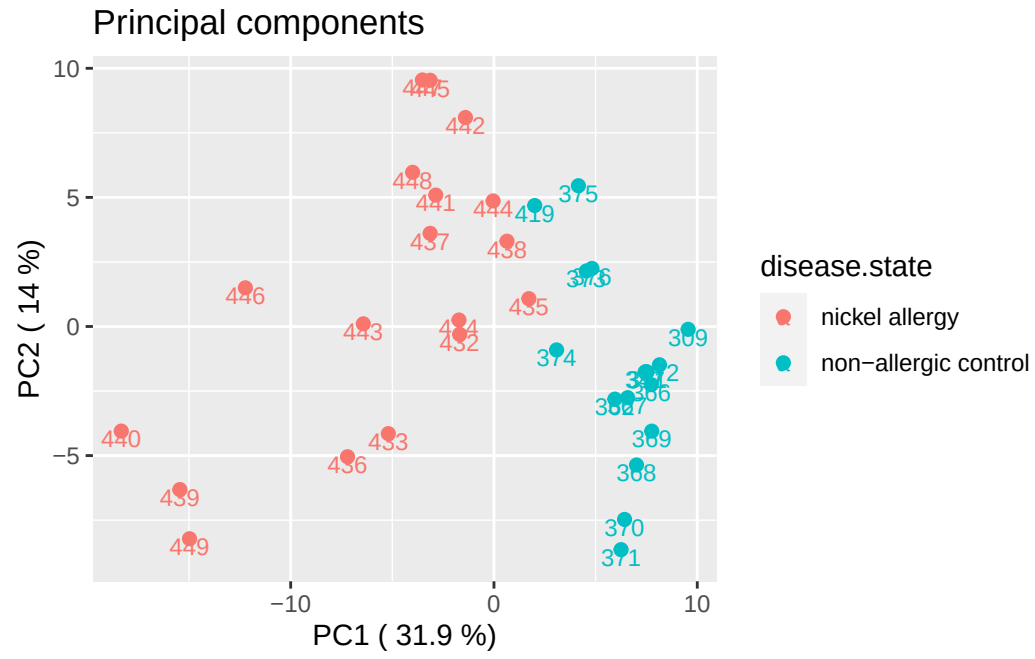
Preprocessat de dades

El conjunt de dades consta de 34 mostres i 54675 gens, els quals es redueixen a 10412 després del filtratge. El següent gràfic mostra els boxplots de la distribució de l'expressió per cada mostra.

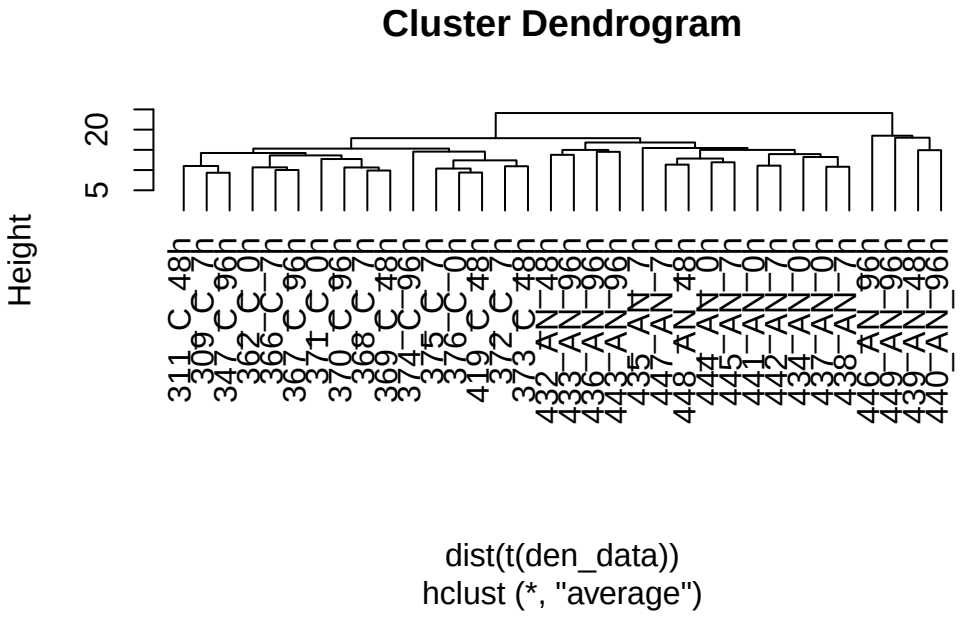
Com s'observa els valors no estan normalitzats. Apliquem una transformació logarítmica de base 2 a les dades i representem de nou el gràfic anterior. Com s'observa en el gràfic següent les dades ara es mostren més normalitzades que en el cas anterior.



El següent gràfic mostra els dues components principals del conjunt de dades, en el qual s'observa una separació entre els grups AN i C.

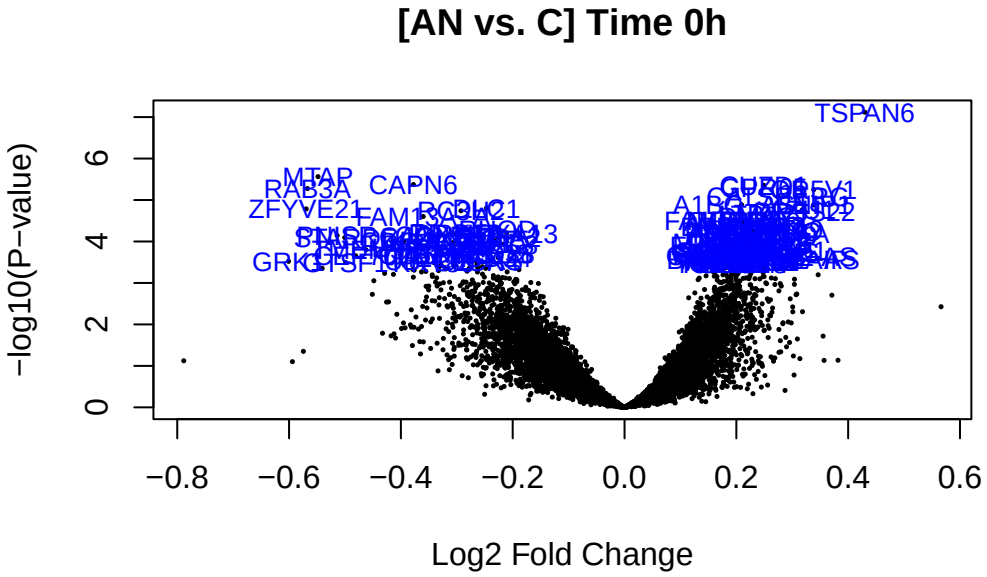


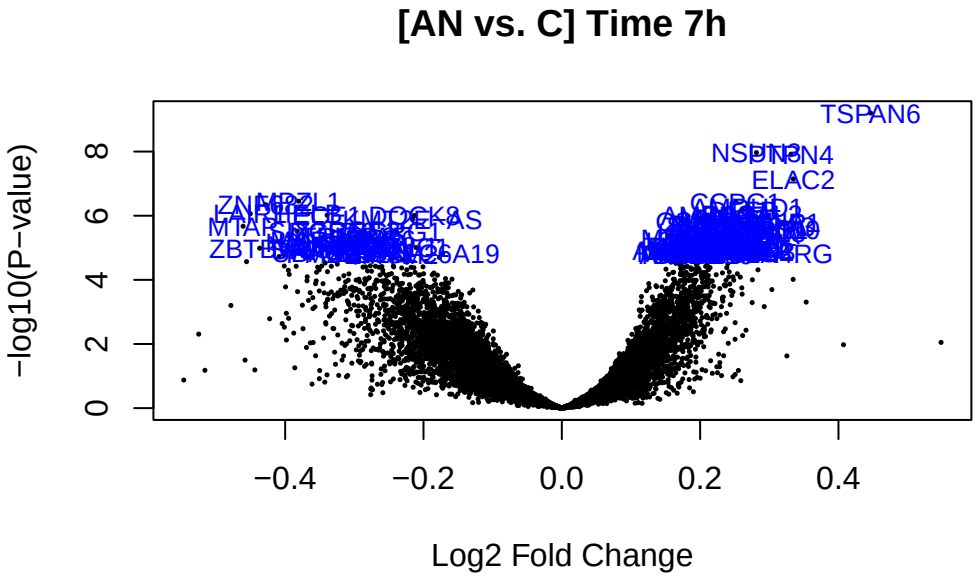
En representar un diagrama de grups jeràrquic, observem una separació diferenciada entre les mostres del grup AN a les 96h d'exposició i la resta. Però també, veiem dos grups diferents en la resta que separa les control dels al · lèrgics al níquel.

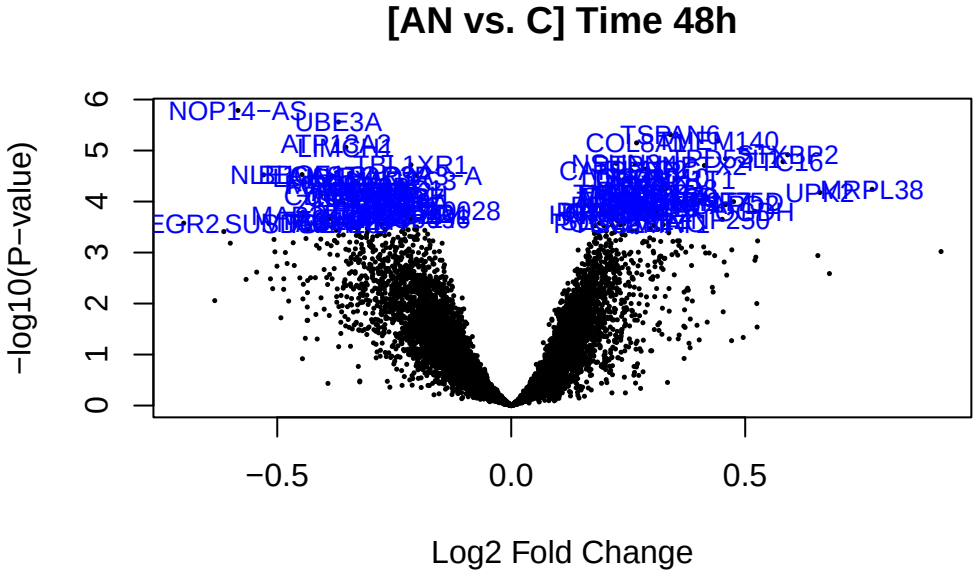


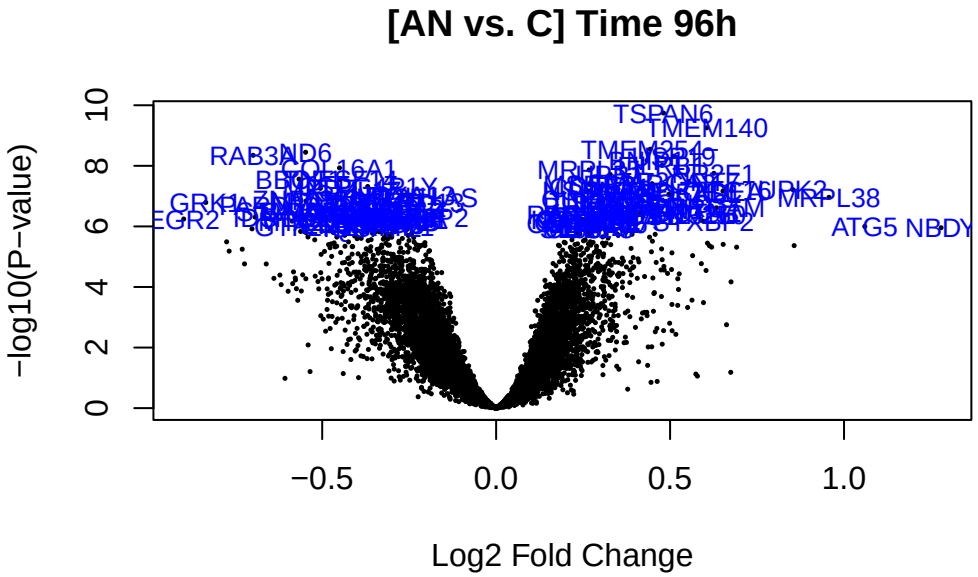
Gens diferencialment expressats

La següent imatge mostra un gràfic tipus volcanoplot per a cada una de les comparacions, on podem observar en blau els gens diferencialment expressats que s'han trobat en l'anàlisi, amb les anotacions SYMBOL.

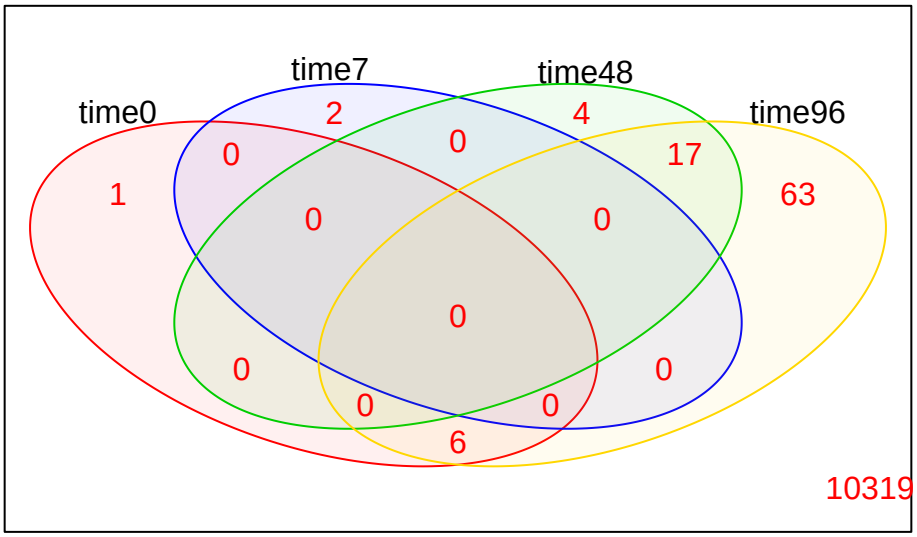



[AN vs. C] Time 7h
−log10(P−value)
0
2
4
6
8
−0.4
−0.2
0.0
0.2
0.4
Log2 Fold Change
TSPAN6
NSUN4
ELAC2

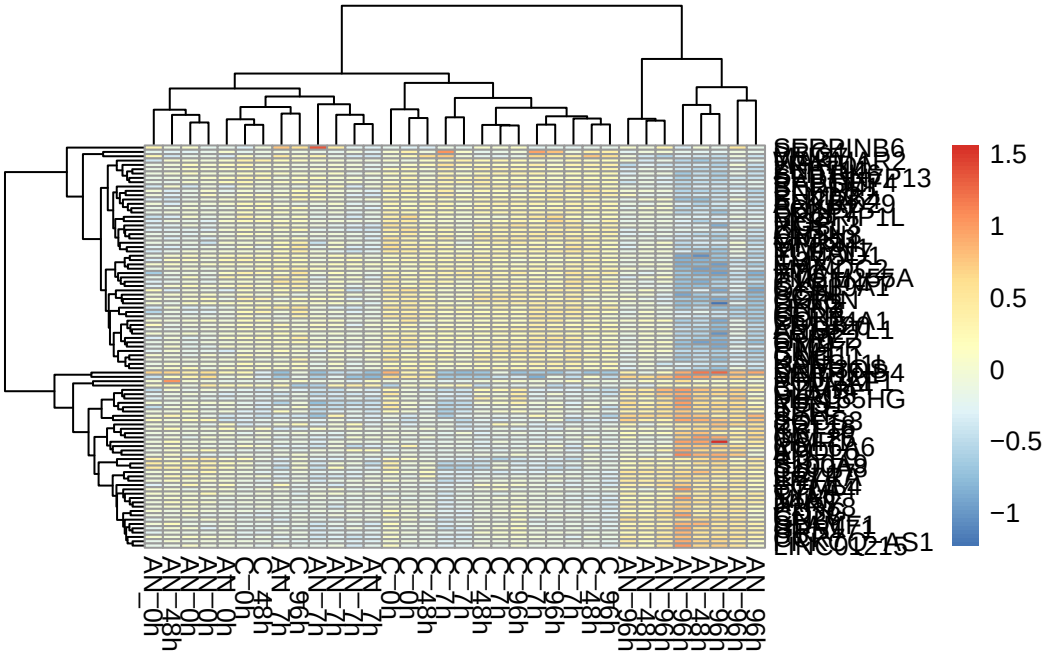

[AN vs. C] Time 48h
−log10(P−value)
0
1
2
3
4
5
6
−0.5
0.0
0.5
Log2 Fold Change
NOP14−AS1
UBE3A
TSPAN6
MRPL38



A conitnuació es mostra el diagrama de Vann on podem veure elnombre de gens regulats “up” i “down” per cada comparació, i quins són comuns entre comparacions. Observem com a temps 7 no trobem cap diferència, mentre que a temps 0 n’observem per dos gens. En canvi, a temps 48h i 96h, coincidint amb l’aparició d’èczema en AN, veien un nombre major de gens diferencialment expressats.



Finalment, fem la representació dels pefils d’expressió utilitzant un mapa de colors. Els resultats també suggereixen un canvi en l’expressió de gens en AN a partir de les 48 hores d’exposició.



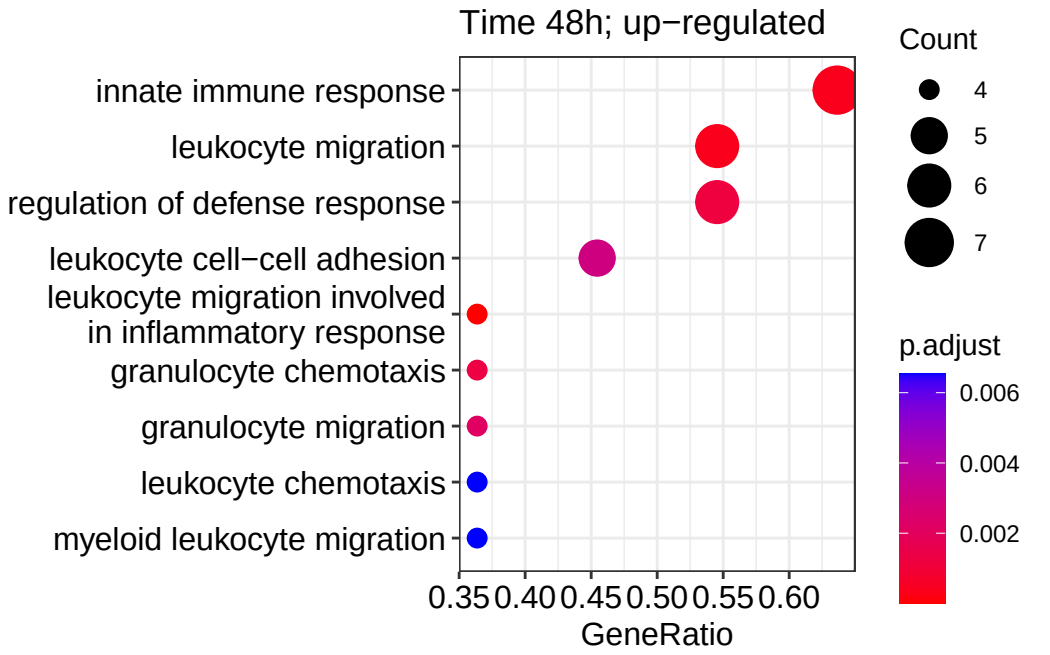
Significació bioloògica

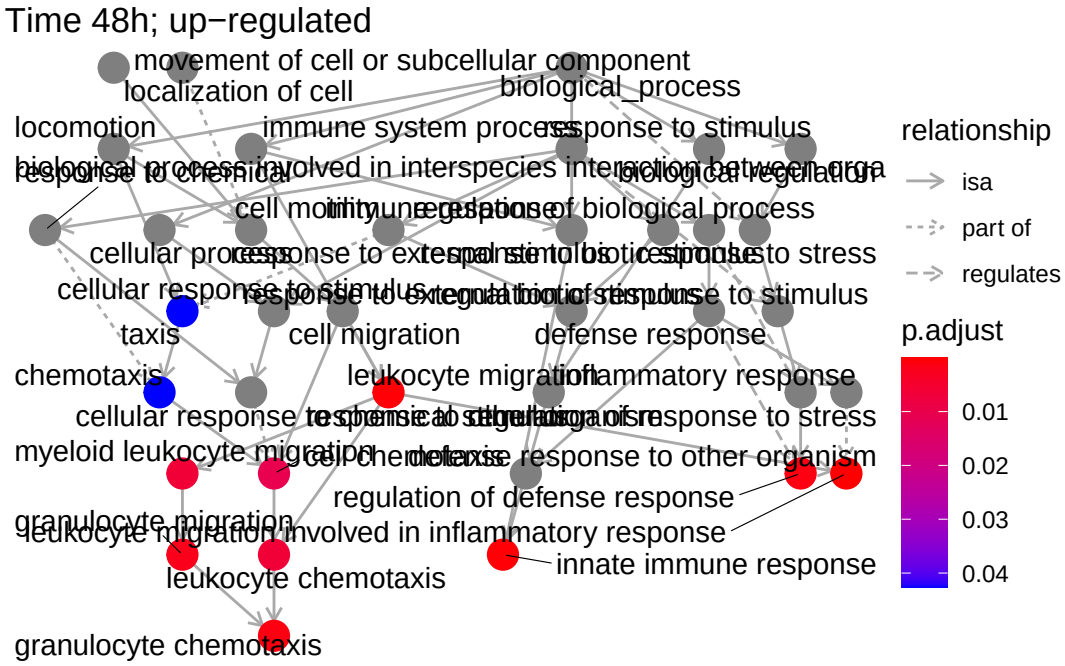
Degut als resultats de les comparacions prèvies, on s'ha vist que només hi ha un diferència entre l'expressió de gens a les 48 i 96h, els anàlisis de significació biològica només es duran a terme per aquestes dues comparacions.

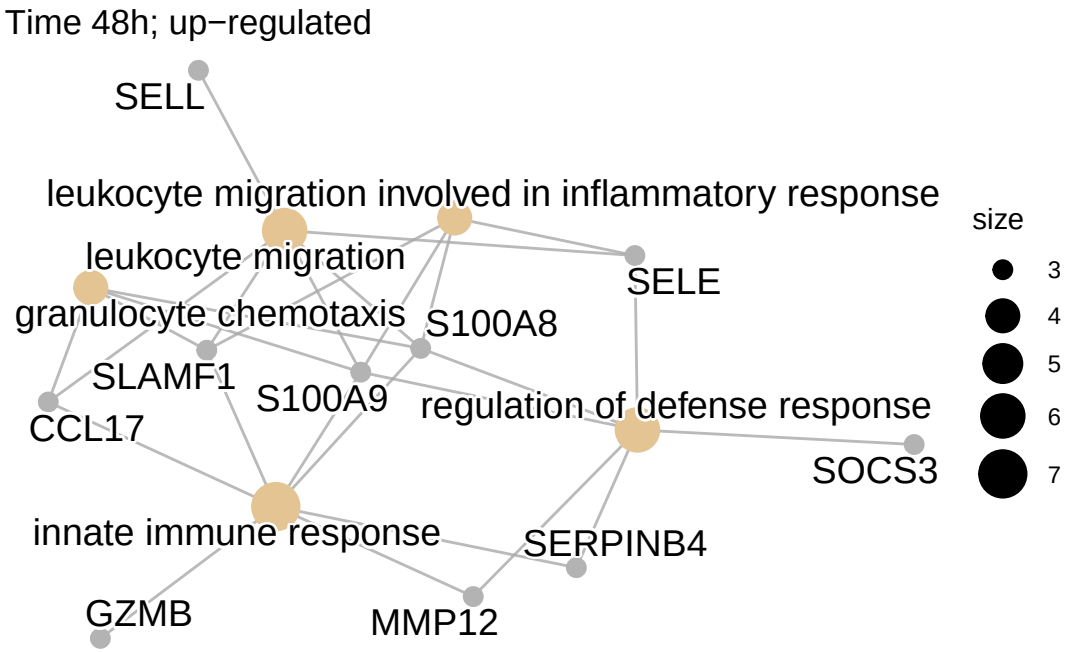
Anàlisis d'enriquiment

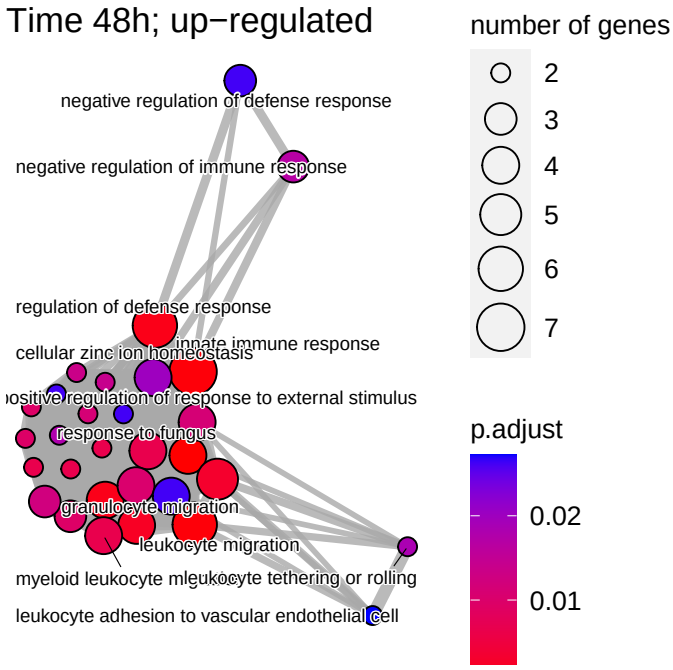
Comparació AN vs. C a les 48h d'exposició

A continuació es mostren els resultats de l'anàlisi d'enriquiment a les 48h per els gens sobre expressats. Els resultats es represent amb un gràfic de punts, una visualització jeràrquica de les ontologies genètiques significatives, gene network, i un mapa d'enriquiment.


Time 48h; up−regulated
innate immune response
leukocyte migration
regulation of defense response
leukocyte cell−cell adhesion
leukocyte migration involved in inflammatory response
granulocyte chemotaxis
granulocyte migration
leukocyte chemotaxis
myeloid leukocyte migration
0.35 0.40 0.45 0.50 0.55 0.60
GeneRatio
Count
4
5
6
7
p.adjust
0.006
0.004
0.002


Time 48h; up−regulated
movement of cell or subcellular component
localization of cell
biological_process
locomotion
immune system process
response to stimulus
response to chemical
biological regulation
cellular process
cell motility
regulation of biological process
response to stress
taxis
cell migration
defense response
chemotaxis
leukocyte migration
inflammatory response
myeloid leukocyte migration
regulation of defense response
granulocyte migration
leukocyte migration involved in inflammatory response
innate immune response
leukocyte chemotaxis
granulocyte chemotaxis
relationship
isa
part of
regulates
p.adjust
0.01
0.02
0.03
0.04

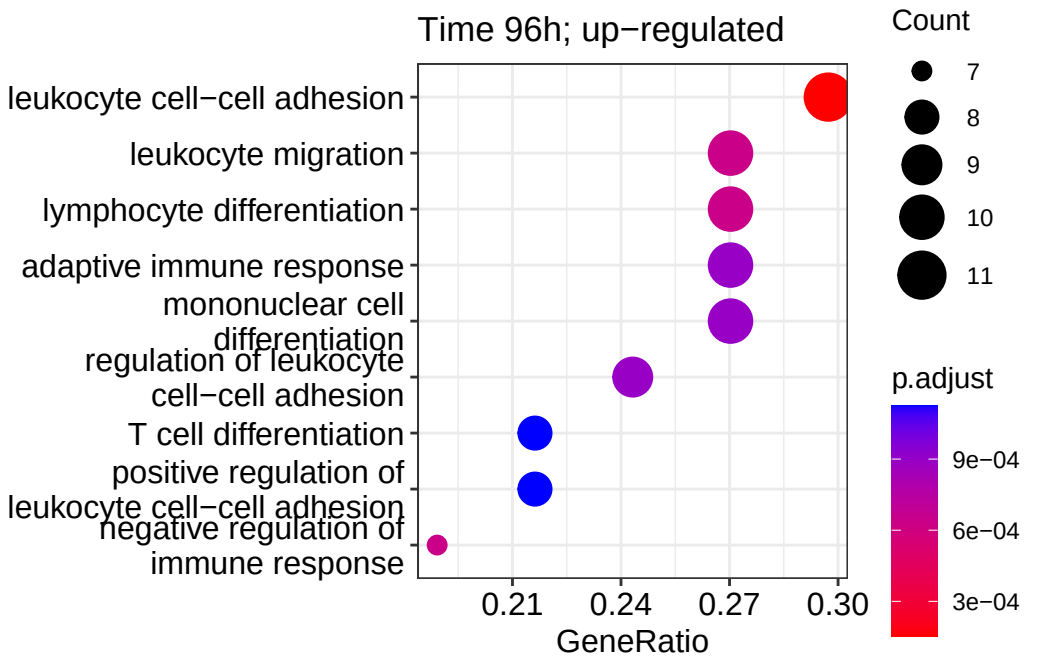


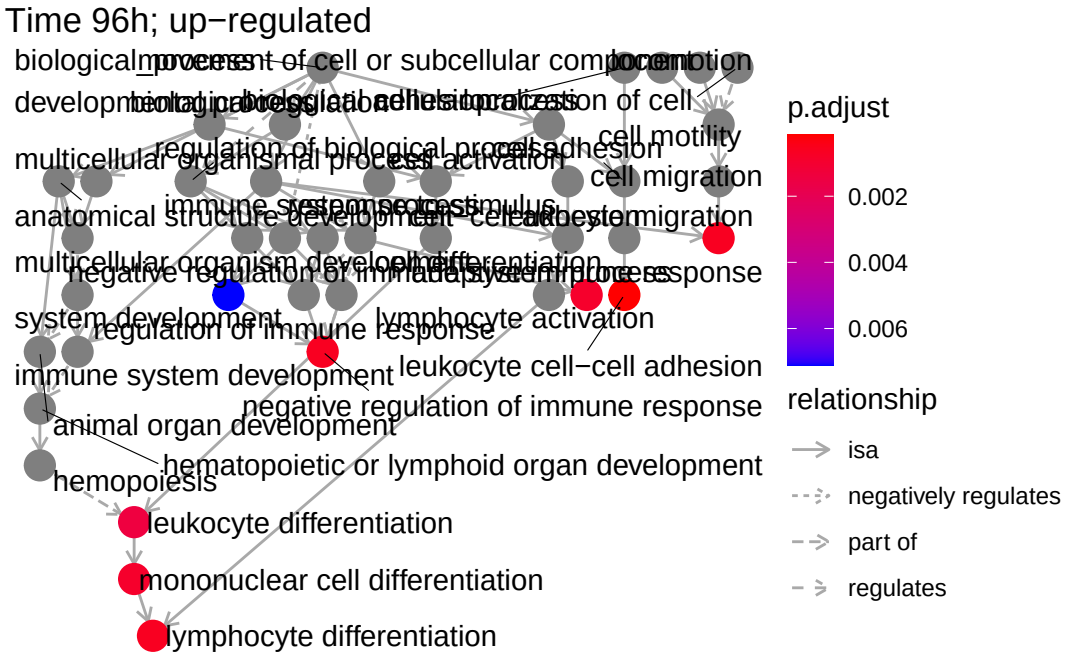


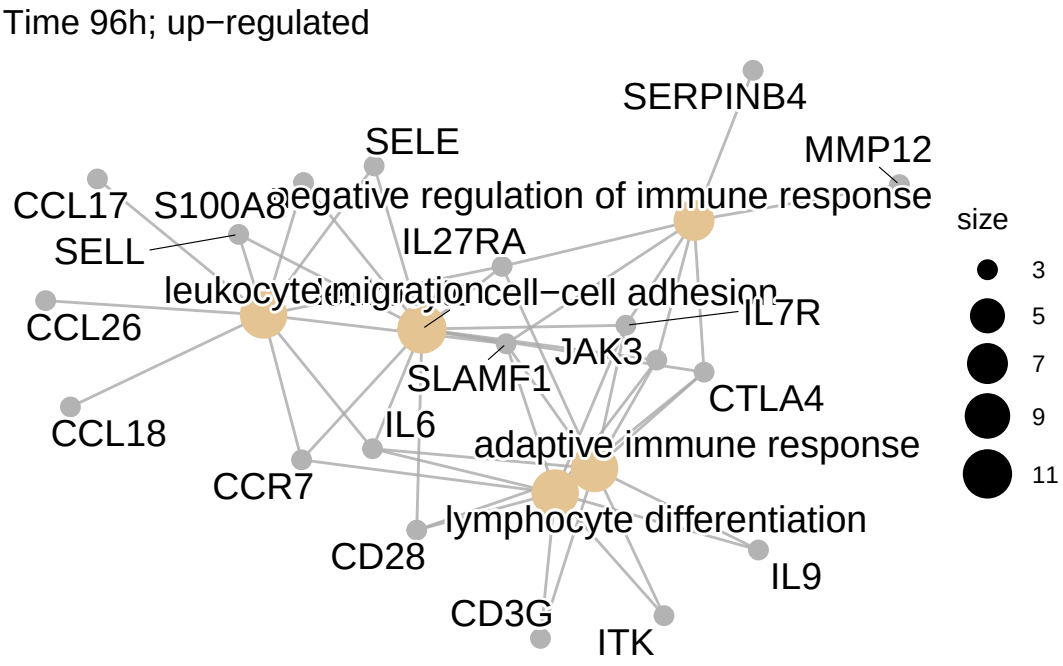
Per els gens que s'han regulat no s'han trobat termes enriquits.

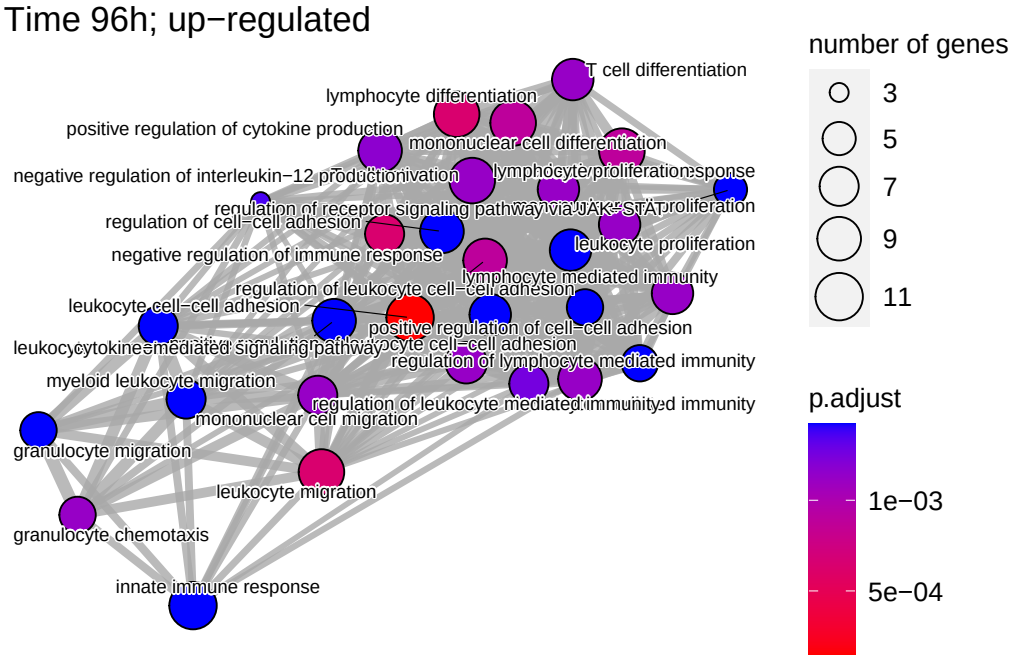
Comparació AN vs. C a les 96h d'exposició

A continuació mostrem els resultats per els gens sobre-expressats.


Time 96h; up−regulated
leukocyte cell−cell adhesion
leukocyte migration
lymphocyte differentiation
adaptive immune response
mononuclear cell differentiation
regulation of leukocyte cell−cell adhesion
T cell differentiation
positive regulation of leukocyte cell−cell adhesion
negative regulation of immune response
0.21
0.24
0.27
0.30
GeneRatio
Count
7
8
9
10
11
p.adjust
9e−04
6e−04
3e−04


Time 96h; up−regulated
biological_process
cell motility
cell migration
immune system development
animal organ development
hemopoiesis
leukocyte differentiation
mononuclear cell differentiation
lymphocyte differentiation
leukocyte cell−cell adhesion
negative regulation of immune response
hematopoietic or lymphoid organ development
p.adjust
0.002
0.004
0.006
relationship
isa
negatively regulates
part of
regulates





Igual que en les 48h, els resultats de l'ànalisi per els gens regulats, resulta en que no hi ha termes significativament enriquits.

Gene Set Expresion Analysis

Els reultats del anàlisi GSEA abans de l'exposició no són significatius, i a temps 7h obtenim el següent:

	Description	setSize	NES	p.adjust
hsa05168	Herpes simplex virus 1 infection	283	-1.904844	1.558191e-06
hsa03010	Ribosome	39	2.000353	2.683779e-02

El nombre de *pathways* biologics significatius augmenta considerablement a les 48h:

	Description	setSize
hsa04061	Viral protein interaction with cytokine and cytokine receptor	70
hsa04657	IL-17 signaling pathway	55
hsa04060	Cytokine-cytokine receptor interaction	178
hsa04064	NF-kappa B signaling pathway	70
hsa04062	Chemokine signaling pathway	127
hsa04668	TNF signaling pathway	75
hsa05169	Epstein-Barr virus infection	115
hsa04659	Th17 cell differentiation	67
hsa05417	Lipid and atherosclerosis	133
hsa04380	Osteoclast differentiation	95
hsa05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	57
hsa04660	T cell receptor signaling pathway	82
hsa04658	Th1 and Th2 cell differentiation	56
hsa05340	Primary immunodeficiency	26
hsa05323	Rheumatoid arthritis	56
hsa04630	JAK-STAT signaling pathway	100
hsa04650	Natural killer cell mediated cytotoxicity	76
hsa04621	NOD-like receptor signaling pathway	105
hsa04640	Hematopoietic cell lineage	66
hsa05162	Measles	82
hsa05167	Kaposi sarcoma-associated herpesvirus infection	108
hsa05163	Human cytomegalovirus infection	128
hsa05135	Yersinia infection	86
hsa05144	Malaria	32
hsa04625	C-type lectin receptor signaling pathway	69
hsa05171	Coronavirus disease - COVID-19	113
hsa04210	Apoptosis	78
hsa05166	Human T-cell leukemia virus 1 infection	133
hsa05164	Influenza A	95
hsa00982	Drug metabolism - cytochrome P450	27
hsa05150	Staphylococcus aureus infection	47
hsa04514	Cell adhesion molecules	97
hsa04620	Toll-like receptor signaling pathway	62
hsa05170	Human immunodeficiency virus 1 infection	120
hsa04672	Intestinal immune network for IgA production	29
hsa04217	Necroptosis	73
hsa05134	Legionellosis	33
hsa04623	Cytosolic DNA-sensing pathway	34
hsa04670	Leukocyte transendothelial migration	73
hsa04622	RIG-I-like receptor signaling pathway	36
hsa05418	Fluid shear stress and atherosclerosis	84
hsa04066	HIF-1 signaling pathway	71
hsa01250	Biosynthesis of nucleotide sugars	22
hsa05146	Amoebiasis	77
hsa05161	Hepatitis B	101

hsa04933	AGE-RAGE signaling pathway in diabetic complications	77
hsa05142	Chagas disease	63
hsa05332	Graft-versus-host disease	20
hsa04662	B cell receptor signaling pathway	53
hsa03008	Ribosome biogenesis in eukaryotes	29
hsa04151	PI3K-Akt signaling pathway	218
hsa05320	Autoimmune thyroid disease	24
hsa05200	Pathways in cancer	324
hsa05219	Bladder cancer	29
hsa05152	Tuberculosis	93
hsa00052	Galactose metabolism	14
hsa05143	African trypanosomiasis	25
hsa05133	Pertussis	36
hsa04710	Circadian rhythm	20
hsa04940	Type I diabetes mellitus	25
hsa00640	Propanoate metabolism	15
hsa05321	Inflammatory bowel disease	43
hsa05132	Salmonella infection	128
hsa04930	Type II diabetes mellitus	33
hsa05330	Allograft rejection	20
hsa00980	Metabolism of xenobiotics by cytochrome P450	26
hsa04310	Wnt signaling pathway	111
hsa04936	Alcoholic liver disease	75
hsa00650	Butanoate metabolism	14
hsa04924	Renin secretion	45
hsa04664	Fc epsilon RI signaling pathway	48

	NES	p.adjust
hsa04061	3.134458	6.480000e-09
hsa04657	3.088354	6.480000e-09
hsa04060	2.942565	6.480000e-09
hsa04064	2.860860	6.480000e-09
hsa04062	2.650331	6.480000e-09
hsa04668	2.710622	7.952262e-09
hsa05169	2.437776	1.215235e-08
hsa04659	2.594664	3.609801e-08
hsa05417	2.304976	8.000341e-08
hsa04380	2.440695	1.536623e-07
hsa05235	2.484186	1.994214e-07
hsa04660	2.398148	5.068760e-07
hsa04658	2.510378	5.863988e-07
hsa05340	2.626159	7.583974e-07
hsa05323	2.454130	1.517277e-06
hsa04630	2.269311	2.021132e-06
hsa04650	2.327650	2.729948e-06
hsa04621	2.181822	7.395551e-06
hsa04640	2.328276	9.465932e-06
hsa05162	2.211464	9.465932e-06
hsa05167	2.106289	1.300601e-05
hsa05163	2.050296	1.507667e-05
hsa05135	2.133359	2.346737e-05
hsa05144	2.251358	6.916899e-05
hsa04625	2.074081	2.006773e-04

```
hsa05171  1.942740 2.006773e-04
hsa04210  2.098758 2.051701e-04
hsa05166  1.866379 2.245919e-04
hsa05164  2.008619 2.436960e-04
hsa00982 -2.090124 2.656863e-04
hsa05150  2.113108 2.995436e-04
hsa04514  1.965855 3.408945e-04
hsa04620  1.991552 5.337177e-04
hsa05170  1.875696 5.337177e-04
hsa04672  2.121227 5.481060e-04
hsa04217  1.882098 1.701670e-03
hsa05134  1.997614 2.781012e-03
hsa04623  2.024589 3.670971e-03
hsa04670  1.800224 4.097279e-03
hsa04622  1.971037 4.893181e-03
hsa05418  1.825632 4.915126e-03
hsa04066  1.784283 4.915126e-03
hsa01250  1.929134 9.428581e-03
hsa05146  1.786513 9.949992e-03
hsa05161  1.676885 1.047699e-02
hsa04933  1.759938 1.133210e-02
hsa05142  1.728866 1.133210e-02
hsa05332  1.928764 1.193830e-02
hsa04662  1.727724 1.222706e-02
hsa03008  1.776376 1.395336e-02
hsa04151  1.460880 1.559627e-02
hsa05320  1.797560 1.819428e-02
hsa05200  1.351829 1.819428e-02
hsa05219  1.728976 1.959494e-02
hsa05152  1.646131 2.056898e-02
hsa00052  1.908606 2.058957e-02
hsa05143  1.808695 2.179661e-02
hsa05133  1.780148 2.285760e-02
hsa04710 -1.746269 2.505403e-02
hsa04940  1.776147 2.765528e-02
hsa00640 -1.775431 3.035064e-02
hsa05321  1.636601 3.421027e-02
hsa05132  1.475495 3.421027e-02
hsa04930  1.661515 3.487563e-02
hsa05330  1.739925 3.706661e-02
hsa00980 -1.666102 3.904977e-02
hsa04310 -1.538650 4.156950e-02
hsa04936  1.640509 4.314243e-02
hsa00650 -1.680648 4.817399e-02
hsa04924 -1.652048 4.817399e-02
hsa04664  1.599064 4.887902e-02
```

A continuació es mostra el resultat de GSEA a les 96h, que és molt similar al de les 48h.

```
                                                  Description setSize
hsa04657                                IL-17 signaling pathway      55
hsa04064                             NF-kappa B signaling pathway      70
```

hsa04061	Viral protein interaction with cytokine and cytokine receptor	70
hsa04060	Cytokine-cytokine receptor interaction	178
hsa04062	Chemokine signaling pathway	127
hsa05169	Epstein-Barr virus infection	115
hsa04659	Th17 cell differentiation	67
hsa04660	T cell receptor signaling pathway	82
hsa05417	Lipid and atherosclerosis	133
hsa04668	TNF signaling pathway	75
hsa04658	Th1 and Th2 cell differentiation	56
hsa05235	PD-L1 expression and PD-1 checkpoint pathway in cancer	57
hsa04650	Natural killer cell mediated cytotoxicity	76
hsa05162	Measles	82
hsa05340	Primary immunodeficiency	26
hsa05323	Rheumatoid arthritis	56
hsa04380	Osteoclast differentiation	95
hsa04640	Hematopoietic cell lineage	66
hsa05171	Coronavirus disease - COVID-19	113
hsa04630	JAK-STAT signaling pathway	100
hsa04210	Apoptosis	78
hsa05135	Yersinia infection	86
hsa05170	Human immunodeficiency virus 1 infection	120
hsa04621	NOD-like receptor signaling pathway	105
hsa05163	Human cytomegalovirus infection	128
hsa05146	Amoebiasis	77
hsa05164	Influenza A	95
hsa05150	Staphylococcus aureus infection	47
hsa05166	Human T-cell leukemia virus 1 infection	133
hsa04514	Cell adhesion molecules	97
hsa04625	C-type lectin receptor signaling pathway	69
hsa00052	Galactose metabolism	14
hsa04623	Cytosolic DNA-sensing pathway	34
hsa04217	Necroptosis	73
hsa05144	Malaria	32
hsa05134	Legionellosis	33
hsa05161	Hepatitis B	101
hsa05167	Kaposi sarcoma-associated herpesvirus infection	108
hsa04622	RIG-I-like receptor signaling pathway	36
hsa04310	Wnt signaling pathway	111
hsa04672	Intestinal immune network for IgA production	29
hsa04152	AMPK signaling pathway	78
hsa00051	Fructose and mannose metabolism	19
hsa04620	Toll-like receptor signaling pathway	62
hsa04923	Regulation of lipolysis in adipocytes	40
hsa04670	Leukocyte transendothelial migration	73
hsa00982	Drug metabolism - cytochrome P450	27
hsa05200	Pathways in cancer	324
hsa04924	Renin secretion	45
hsa05152	Tuberculosis	93
hsa05145	Toxoplasmosis	65
hsa03010	Ribosome	39
hsa05219	Bladder cancer	29
hsa05160	Hepatitis C	90

hsa04662	B cell receptor signaling pathway	53
hsa04664	Fc epsilon RI signaling pathway	48
hsa05230	Central carbon metabolism in cancer	42
hsa05142	Chagas disease	63
hsa01250	Biosynthesis of nucleotide sugars	22
hsa05320	Autoimmune thyroid disease	24
hsa05203	Viral carcinogenesis	102
hsa04933	AGE-RAGE signaling pathway in diabetic complications	77
hsa04115	p53 signaling pathway	60
hsa05332	Graft-versus-host disease	20

	NES	p.adjust
hsa04657	2.834292	1.050007e-08
hsa04064	2.820281	1.050007e-08
hsa04061	2.795306	1.050007e-08
hsa04060	2.688476	1.050007e-08
hsa04062	2.500463	2.487721e-08
hsa05169	2.532838	3.911508e-08
hsa04659	2.568700	1.201140e-07
hsa04660	2.504209	1.201140e-07
hsa05417	2.334053	1.360929e-07
hsa04668	2.474962	6.176719e-07
hsa04658	2.529263	8.303134e-07
hsa05235	2.488184	2.405505e-06
hsa04650	2.359399	2.405505e-06
hsa05162	2.333604	2.405505e-06
hsa05340	2.578622	5.205441e-06
hsa05323	2.383729	1.113618e-05
hsa04380	2.188138	1.636295e-05
hsa04640	2.289695	2.690993e-05
hsa05171	2.106335	4.289060e-05
hsa04630	2.079204	7.719620e-05
hsa04210	2.133545	1.669946e-04
hsa05135	2.085353	1.669946e-04
hsa05170	1.995255	2.356409e-04
hsa04621	2.011079	5.927326e-04
hsa05163	1.868846	1.142881e-03
hsa05146	1.933371	1.243109e-03
hsa05164	1.912608	1.243109e-03
hsa05150	1.992002	2.386803e-03
hsa05166	1.785270	2.386803e-03
hsa04514	1.875229	2.495490e-03
hsa04625	1.967988	2.641378e-03
hsa00052	2.146935	3.286732e-03
hsa04623	2.020464	3.286732e-03
hsa04217	1.904664	3.332411e-03
hsa05144	2.009720	3.470331e-03
hsa05134	2.001047	3.606360e-03
hsa05161	1.767934	4.452030e-03
hsa05167	1.777529	4.679386e-03
hsa04622	2.002633	6.082688e-03
hsa04310	-1.710302	7.325076e-03
hsa04672	1.982569	9.299835e-03

```
hsa04152 -1.771350 9.299835e-03
hsa00051  1.907115 1.110518e-02
hsa04620  1.793495 1.110518e-02
hsa04923 -1.821685 1.245046e-02
hsa04670  1.761873 1.374176e-02
hsa00982 -1.880044 1.449225e-02
hsa05200  1.378422 1.570293e-02
hsa04924 -1.801277 1.628239e-02
hsa05152  1.648152 1.628239e-02
hsa05145  1.753991 1.735932e-02
hsa03010  1.760633 2.223842e-02
hsa05219  1.852232 2.365561e-02
hsa05160  1.636096 2.371811e-02
hsa04662  1.683683 2.618367e-02
hsa04664  1.695904 2.792193e-02
hsa05230  1.694243 3.188724e-02
hsa05142  1.673254 3.358029e-02
hsa01250  1.804635 3.593760e-02
hsa05320  1.794757 3.688350e-02
hsa05203  1.582182 3.761704e-02
hsa04933  1.611740 4.018677e-02
hsa04115  1.670294 4.103818e-02
hsa05332  1.824579 4.638684e-02
```

Discusió

Els resultats presentats en aquest document estan molt en línia amb els van publicar Pedersen et al. (2007). A les 7h no s'observa diferencies significatives entre els dos grups, en canvi, a les 48h i 96h, quan ja es mostra l'èczema, si que en veiem. A més, de l'anàlisi de significació biològica els resultats a temps 48h i 96h mostren resultats molt plausibles amb el coneixement que és té dels processos biologics que es desenvolupen quan s'entra en contacte amb un al · lergen (e.g. leukocyte migration involved in inflammatory response, regulation of defense response, chemokine signaling pathway…).

Aquests estudi té algunes limitacions. Per començar, ens trobem amb les limitacions habituals en estudis de microarrays que inclouen l'efecte batch. A més, la mida de la mostra és molt petita (12 persones, 34 mostres en 4 cursos temporals per dos grups de població). Aquesta limitació pot portar a tenir un nombre major de falsos positius i falsos negatius, a més de limitar la reproductibilitat de l'estudi. Un altra limitació és que les mostres corresponen a dones de mitjana edat, el que en limita la generalització dels resultats a la població general. Punts forts de l'estudi és que els resultats obtinguts estan en línia amb evidencies anteriors @… . A més l'estudi utilitza paquets de Bioconductor per a l'anàlisi de dades, els quals han estat validats anteriorment i són àmpliament utilitzats en el camp de la bioinformàtica.

Aquest treball és un primer pas per al descobriment dels gens que intervenen en el desenvolupament d'èczema per DCA. Cal seguir investigant sobre les bases d'aquest estudi per donar com a definitius els resultats obtinguts, i proporcionar més evidencies científiques sobre els processos biològiques que tenen lloc en DCA.

Codi

```
## Packages ----
library(Biobase)
library(GEOquery)
library(dplyr)
library(tidyr)
library(hgu133plus2.db)
library(ggplot2)
library(limma)
library(pheatmap)
library(clusterProfiler)
library(enrichplot)
library(genefilter)

# Use GEOquery to download the dataset with dataset accesion ID GDS2935
gds <- getGEO("GDS2935")
eset <- GDS2eSet(gds, do.log2=FALSE)

# Annotation
annotation(eset) <-"hgu133plus2.db"

# filter genes with little variation or low signal across samples, and those
# with inssuficient annotations
filter_result <- nsFilter(eset)
eset_filtered <- filter_result$eset


# Create variables with the slots pData and exprs
pData <- pData(phenoData(eset_filtered))
exprs <- exprs(eset_filtered)

# EXPLORATORY DATA ANALYSIS ----
# prepare data
data_to_plot <- exprs %>%
  as_tibble() %>%
  pivot_longer(cols = colnames(exprs), names_to = "sample") %>%
  inner_join(pData %>%
               as_tibble() %>%
               dplyr::select(sample, time, disease.state, agent, individual)) %>%
  mutate(value_log = log(value))

## Boxplot ----
ggplot(data_to_plot,
       aes(x = sample, y = value, color = disease.state)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  ggtitle("Expression values")

## Boxplot - log
```

```
ggplot(data_to_plot,
       aes(x = sample, y = value_log, color = disease.state)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  ggtitle("Log expression values")

# Logarithmic transform data shows better results regarding normalisation,
# therefore we work with log2 data:
eset <- GDS2eSet(gds, do.log2=TRUE)
annotation(eset) <-"hgu133plus2.db"
filter_result <- nsFilter(eset)
eset_filtered <- filter_result$eset
pData <- pData(phenoData(eset_filtered))
exprs <- exprs(eset_filtered)
data_to_plot <- exprs %>%
  as_tibble() %>%
  pivot_longer(cols = colnames(exprs), names_to = "sample") %>%
  inner_join(pData %>%
               as_tibble() %>%
               dplyr::select(sample, time, disease.state, agent, individual)) %>%
  mutate(value_log = log(value))


## PCA ----
# pca with log data
pcs <- prcomp(t(exprs), scale = FALSE) # Log data

# nice axisa labels
loads <- round(pcs$sdev^2/sum(pcs$sdev^2)*100, 1)
xlab <- c(paste("PC1", "(", loads[1],"%)"))
ylab <- c(paste("PC2", "(", loads[2],"%)"))

# plot
pcs$x %>% as_tibble(rownames = "sample") %>%
  inner_join( data_to_plot %>% distinct(sample, disease.state)) %>%
  ggplot(aes(x = PC1, y = PC2, col = disease.state)) +
  geom_point(size = 2) +
  ylab(ylab) +
  xlab(xlab) +
  geom_text(aes(y = PC2-0.3, label = substr(sample, 7, 9)), size = 3) +
  ggtitle("Principal components")


## Hierarchichal clustering ----
# informative name:
time <- gsub(" ", "", pData$time)
time <- gsub("control", "0h", time)

disease <- gsub("nickel allergy", "AN", pData$disease.state)
disease <- gsub("non-allergic control", "C", disease)
```

```r
inf_name <- paste(substr(rownames(pData), 7, 9), disease, time, sep="_" )

den_data <- exprs
colnames(den_data) <- inf_name
clust.euclid.average <- hclust(dist(t(den_data)), method = "average")
plot(clust.euclid.average, hang = -1)


# DIFFERENTIALY EXPRESSED GENS ----
## Desing ----
group <- paste(disease, time, sep="_" )
design  <-  model.matrix(~0+group)
colnames(design) <- gsub("group", "", colnames(design))
rownames(design) <- pData$sample

## Comparisons
# 1) 0h: allergic vs. control
# 2) 7h: allergic vs. control
# 3) 48h: allergic vs. control
# 4) 96h: allergic vs. control

## Contrasts ----
cont.matrix <- makeContrasts (
  time0 = AN_0h - C_0h,
  time7 = AN_7h - C_7h,
  time48 = AN_48h - C_48h,
  time96 = AN_96h - C_96h,
  levels = design)

## Fit the linear model ----
fit <- lmFit(eset_filtered, design)
fit.cont <- contrasts.fit(fit, cont.matrix)
fit.cont <- eBayes(fit.cont)

## Anotation ----
# get gene anotation
anotations <- AnnotationDbi::select(hgu133plus2.db,
                                    keys = rownames(exprs),
                                    columns = c("ENTREZID", "SYMBOL"))

# function to add annotations to result tables
anotateTable <- function(x) {
  x %>%
    as_tibble() %>%
    dplyr::rename("PROBEID" = "ID") %>%
    dplyr::inner_join(anotations, by = "PROBEID") %>%
    arrange(adj.P.Val)
}

## Get top tables anotated ----
```

```r
topTab_time0  <- fit.cont %>%
  topTable(number = nrow(fit.cont), coef= "time0") %>%
  anotateTable()
topTab_time7  <- fit.cont %>%
  topTable(number = nrow(fit.cont), coef= "time7") %>%
  anotateTable()
topTab_time48 <- fit.cont %>%
  topTable(number = nrow(fit.cont), coef= "time48") %>%
  anotateTable()
topTab_time96 <- fit.cont %>%
  topTable(number = nrow(fit.cont), coef= "time96") %>%
  anotateTable()

# head(topTab_time0) %>%
#  dplyr::select(PROBEID, GO.Function, adj.P.Val, ENTREZID, SYMBOL)

# Volcano plots per comparsion ----
# TIME 0
volcanoplot(fit.cont, coef = "time0", highlight = 100,
            names = topTab_time0$SYMBOL,
            main="[Time 0] nickel allergy vs. control")
abline(v = c(-1, 1))

# TIME 7
volcanoplot(fit.cont, coef = "time7", highlight = 100,
            names = topTab_time0$SYMBOL,
            main="[Time 0] nickel allergy vs. control")
abline(v = c(-1, 1))

# TIME 48
volcanoplot(fit.cont, coef = "time48", highlight = 100,
            names = topTab_time0$SYMBOL,
            main="[Time 0] nickel allergy vs. control")
abline(v = c(-1, 1))

# TIME 96
volcanoplot(fit.cont, coef = "time96", highlight = 100,
            names = topTab_time0$SYMBOL,
            main="[Time 0] nickel allergy vs. control")
abline(v = c(-1, 1))


## Multiple comparisons ----
summary.fit <- decideTests(fit.cont, p.value = 0.05)

# ven diagram
vc <- vennCounts(summa.fit)
vennDiagram(vc,
            include=c("up", "down"),
            counts.col=c("red", "blue"),
```

```
            circle.col = c("red", "blue", "green3", "gold"),
            cex=c(1, 1, 1, 1))


## Expression profile ----
topGenes0 <- subset(topTab_time0, (abs(logFC) > 0.5) & (adj.P.Val < 0.05))$SYMBOL
topGenes7 <- subset(topTab_time7, (abs(logFC) > 0.5) & (adj.P.Val < 0.05))$SYMBOL
topGenes48 <- subset(topTab_time48, (abs(logFC) > 0.5) & (adj.P.Val < 0.05))$SYMBOL
topGenes96 <- subset(topTab_time96, (abs(logFC) > 0.5) & (adj.P.Val < 0.05))$SYMBOL

topGenes <- unique(c(topGenes0, topGenes7, topGenes48, topGenes96))

# affy to symbol:
exprs_symb <- exprs
rownames(exprs_symb) <- anotations$SYMBOL

mat  <- exprs_symb[topGenes, ]
mat  <- mat - rowMeans(mat)
colnames(mat) <- group
pheatmap(mat)


# BIOLOGICAL SIGNIFICANCE ANALYSES ----
## Over-Representation Analysis (ORA) ----
### Time 0 ----
# no significative

### Time 7 ----
#### up ----
# no significative


### Time 48 ----
#### up ----
selectedEntrezs <- subset(topTab_time48, (logFC > 0.5) & (adj.P.Val < 0.05))$ENTREZID
ego <- enrichGO(gene = selectedEntrezs,
                universe = rownames(topTab_time48),
                keyType = "ENTREZID",
                OrgDb = hgu133plus2.db,
                ont = "BP",
                pAdjustMethod = "BH",
                qvalueCutoff = 0.05,
                readable = TRUE)

title <- "Time 48h; overexpressed"
# dot plot of more enriched categories
dotplot(ego, showCategory = 9) + ggtitle(title)

# herarchical visualization of GO terms
goplot(ego, showCategory = 5, cex = 0.5)  + ggtitle(title)
```

```r
# gene network
cnetplot(ego) + ggtitle(title)

# Enrichment map
ego_sim <- pairwise_termsim(ego)
emapplot(ego_sim, cex_label_category=0.5) + ggtitle(title)


#### down ----
selectedEntrezs <- rownames(subset(topTab_time48, (logFC < -0.5) & (adj.P.Val < 0.05)))
ego <- enrichGO(gene = selectedEntrezs,
                universe = rownames(topTab_time48),
                keyType = "ENTREZID",
                OrgDb = hgu133plus2.db,
                ont = "BP",
                pAdjustMethod = "BH",
                qvalueCutoff = 0.05,
                readable = TRUE)

title <- "Time 48h; regulated"
# dot plot of more enriched categories
dotplot(ego, showCategory = 9) + ggtitle(title)

# gene network
cnetplot(ego) + ggtitle(title)

### Time 96 ----
#### up ----
selectedEntrezs <- rownames(subset(topTab_time96, (logFC > 0.5) & (adj.P.Val < 0.05)))
ego <- enrichGO(gene = selectedEntrezs,
                universe = rownames(topTab_time96),
                keyType = "ENTREZID",
                OrgDb = hgu133plus2.db,
                ont = "BP",
                pAdjustMethod = "BH",
                qvalueCutoff = 0.05,
                readable = TRUE)

title <- "Time 96h; overexpressed"
# dot plot of more enriched categories
dotplot(ego, showCategory = 9) + ggtitle(title)

# herarchical visualization of GO terms
goplot(ego, showCategory = 5, cex = 0.5)  + ggtitle(title)

# gene network
cnetplot(ego) + ggtitle(title)

# Enrichment map
ego_sim <- pairwise_termsim(ego)
```

```
emapplot(ego_sim, cex_label_category=0.5) + ggtitle(title)


#### down ----
selectedEntrezs <- rownames(subset(topTab_time48, (logFC < -0.5) & (adj.P.Val < 0.05)))
ego <- enrichGO(gene = selectedEntrezs,
                universe = rownames(topTab_time48),
                keyType = "ENTREZID",
                OrgDb = hgu133plus2.db,
                ont = "BP",
                pAdjustMethod = "BH",
                qvalueCutoff = 0.05,
                readable = TRUE)

title <- "Time 96h; regulated"
# dot plot of more enriched categories
dotplot(ego, showCategory = 9) + ggtitle(title)

# gene network
cnetplot(ego) + ggtitle(title)


## Gene Set Enrichment Analysis (GSEA) ----
### time 0 ----
# sort by absolute logFC to remove duplicates with smallest absolute logFC
geneList <- topTab_time0[order(abs(topTab_time0$logFC), decreasing = TRUE),]
geneList <- geneList[!duplicated(geneList$ENTREZID), ]  ### Keep highest
# re-order based on logFC to be GSEA ready
geneList <- geneList[order(geneList$logFC, decreasing = TRUE),]
genesVector <- geneList$logFC
names(genesVector) <- geneList$ENTREZID

set.seed(123)
gseResults <- gseKEGG(geneList = genesVector)

gsea.result <- setReadable(gseResults, OrgDb = hgu133plus2.db, keyType ="ENTREZID")
gsea.result.df <- as.data.frame(gsea.result)
gsea.result.df[,c("Description","setSize","NES","p.adjust")]

### time 7 ----
# sort by absolute logFC to remove duplicates with smallest absolute logFC
geneList <- topTab_time7[order(abs(topTab_time7$logFC), decreasing = TRUE),]
geneList <- geneList[!duplicated(geneList$ENTREZID), ]  ### Keep highest
# re-order based on logFC to be GSEA ready
geneList <- geneList[order(geneList$logFC, decreasing = TRUE),]
genesVector <- geneList$logFC
names(genesVector) <- geneList$ENTREZID

gseResults <- gseKEGG(geneList = genesVector)
```

```
gsea.result <- setReadable(gseResults, OrgDb = hgu133plus2.db, keyType ="ENTREZID")
gsea.result.df <- as.data.frame(gsea.result)
gsea.result.df[,c("Description","setSize","NES","p.adjust")]

### time 48 ----
# sort by absolute logFC to remove duplicates with smallest absolute logFC
geneList <- topTab_time48[order(abs(topTab_time48$logFC), decreasing = TRUE),]
geneList <- geneList[!duplicated(geneList$ENTREZID), ]  ### Keep highest
# re-order based on logFC to be GSEA ready
geneList <- geneList[order(geneList$logFC, decreasing = TRUE),]
genesVector <- geneList$logFC
names(genesVector) <- geneList$ENTREZID

gseResults <- gseKEGG(geneList = genesVector)

gsea.result <- setReadable(gseResults, OrgDb = hgu133plus2.db, keyType ="ENTREZID")
gsea.result.df <- as.data.frame(gsea.result)
gsea.result.df[,c("Description","setSize","NES","p.adjust")]

### time 96 ----
# sort by absolute logFC to remove duplicates with smallest absolute logFC
geneList <- topTab_time96[order(abs(topTab_time96$logFC), decreasing = TRUE),]
geneList <- geneList[!duplicated(geneList$ENTREZID), ]  ### Keep highest

# re-order based on logFC to be GSEA ready
geneList <- geneList[order(geneList$logFC, decreasing = TRUE),]
genesVector <- geneList$logFC
names(genesVector) <- geneList$ENTREZID

gseResults <- gseKEGG(geneList = genesVector)

gsea.result <- setReadable(gseResults, OrgDb = hgu133plus2.db, keyType ="ENTREZID")
gsea.result.df <- as.data.frame(gsea.result)
gsea.result.df[,c("Description","setSize","NES","p.adjust")]
```

Referències

Bourgon, Richard, Robert Gentleman, and Wolfgang Huber. 2010. "Independent Filtering Increases Detection Power for High-Throughput Experiments." *Proceedings of the National Academy of Sciences* 107 (21): 9546–51. https://doi.org/10.1073/pnas.0914005107.

Pedersen, Malene B., Lone Skov, Torkil Menné, Jeanne D. Johansen, and Jørgen Olsen. 2007. "Gene Expression Time Course in the Human Skin During Elicitation of Allergic Contact Dermatitis." *Journal of Investigative Dermatology* 127 (11): 2585–95. https://doi.org/10.1038/sj.jid.5700902.

Rustemeyer, Thomas, Ingrid M. W. van Hoogstraten, B. Mary E. von Blomberg, and Rik J. Scheper. 2020. "Mechanisms of Allergic Contact Dermatitis." In *Kanerva's Occupational Dermatology*, edited by Swen Malte John, Jeanne Duus Johansen, Thomas Rustemeyer, Peter Elsner, and Howard I. Maibach, 151–90. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-68617-2_14.

Smyth, Gordon K. 2004. "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments." *Statistical Applications in Genetics and Molecular Biology*

3 (1): 1–25. https://doi.org/10.2202/1544-6115.1027.