PEC 2

Vicent Caselles Ballester

2023-12-17

Contents

1 Introducció

L'objectiu de la PAC 2 és la realització d'un anàlisi de dades de microarray d'un *dataset* obtingut del repositori públic d'NCBI ***Gene Expression Omnibus*** (***GEO***, (Edgar 2002)).

En aquesta PEC, intentaré demostrar que he anat assolint els coneixements que s'han exposat al contingut de l'assignatura Anàlisi de Dades Òmiques, concretament referent als mòduls 1 i 2 de l'assignatura. En el cas d'aquesta PEC, no realitzarem el pas de normalització de les dades ni de control de qualitat d'aquestes, tal i com vam fer a la prova anterior.

El exercici, doncs, consistirà en descarregar un conjunt de dades del repositori de GEO utilitzant el conjunt d'eines de `BioConductor`, el *fitteig* d'un model lineal amb `limma` que ens permeti realitzar contrasts entre els grups d'interès biològic, i l'obtenció a partir d'aquest model dels gens diferencialment expressats en un format **llista**.

A partir d'aquesta llista, podrem dur a terme el que es coneix com a l'annotació d'aquests gens, que ens permetrà extreure informació relativa als potencials processos biològics que poden veure afectats amb la presència o absència d'una determinada covariable. Això ho podem fer de diverses maneres, entre les quals destaquen el ***Over Representation Analysis*** (***ORA***) i el ***Gene Set Enrichment Analysis*** (***GSEA***).

```
require(GEOquery)

gds <- getGEO('GDS2107')
remove(gds)
```

A continuació, descarrego les dades amb les que treballaré en aquesta PEC. Aquestes corresponen a la sèrie `GSE3311`, que formen part de l'estudi de Kubisch et al. (2006). En aquest estudi, els autors van intentar mesurar els gens involucrats en la sensibilització del pàncrees que s'ha observat davant de consumició d'etanol a llarg termini.

```
require(GEOquery)
gse <- getGEO('GSE3311')
```

```
## Found 1 file(s)
```

```
## GSE3311_series_matrix.txt.gz
```

2 Materials i mètodes

2.1 *Dataset*

El *dataset* escollit per a la realització d'aquesta PEC és el dataset amb *Accession ID* `GDS2107` o, per altra banda, a la sèrie amb *Accession ID* `GSE3311` (Kubisch et al. (2006)).

Aquest estudi va separar dos grups de rates (*Rattus norvegicus*), a les quals se'ls hi va donar etanol durant 8 setmanes, moment en el qual van ser *eutanitzades* i el pàncrees extret per al seu anàlisi. Posteriorment, es va homogeneïtzar el teixit pàncreatic de 3/4 rates per grup experimental, i aquest homogeneïtzat és el que va ser analitzat en el *microarray*.

És a dir, entenc que les 6 mostres són rèpliques tècniques provenent 3 del homogeneïtzat del grup control i 3 del homogeneïtzat del grup tractat amb etanol.

Utilitzant el *slot* de l'`ExpressionSet` anomenat `phenoData`, podem extreure informació molt valuosa referent al *dataset* `GDS2107`. Per exemple, podem esbrinar el número de canals (i també si totes les mostres tenien el mateix número de canals).

```
unique(pData(gse[[1]])$channel_count)
```

```
## [1] "1"
```

Veiem que tots els canals tenen només un canal.

```
unique(pData(gse[[1]])$characteristics_ch1)
```

```
## [1] "pancreas, control diet, male Wistar rat"
## [2] "pancreas, ethanol diet, male Wistar rat"
```

Les mostres provenen totes de rates de tipus Wistar, i també interessant, totes eren mascles de sexe (no tenim sexe com a covariable).

```
knitr::kable(pData(gse[[1]])[, c(1,31)])
```

	title	data_row_count
GSM74493	pancreas, control diet, replicate 1	15923
GSM74494	pancreas, control diet, replicate 2	15923
GSM74495	pancreas, control diet, replicate 3	15923
GSM74496	pancreas, ethanol diet, replicate 1	15923
GSM74497	pancreas, ethanol diet, replicate 2	15923
GSM74498	pancreas, ethanol diet, replicate 3	15923

A més podem veure el tractament i rèplica al que pertany cada *sample*. Amb aquesta informació, canvio el nom de les columnes de l'`ExpressionSet` per tal de que siguin més informatives.

```
colnames(gse[[1]]) <- paste(c(rep("Ctrl", 3), rep('Eth', 3)), c("01", "02", "03"),
                            sep='')
```

2.2 Eines de BioConductor (Huber et al. 2015)

2.2.1 GEOquery (Davis and Meltzer 2007)

Com he mencionat abans, GEOquery és un paquet que permet la interacció amb el repositori de dades d'NCBI ***Gene Expression Omnibus***. D'aquesta manera, permet descarregar fàcilment conjunts de dades directament des de R, obtenint-les en formats compatibles amb els altres paquets de BioConductor.

2.2.2 limma (Ritchie et al. 2015)

Per a dur a terme els anàlisis estadístics de les dades del *microarray*, i.e. l'ànalisi de gens diferencialment expressats entre els grups experimentals, utilitzarem el paquet limma, que permet ajustar models lineals en gran conjunts de dades com els microarrays.

2.2.3 arrayQualityMetrics (Kauffmann, Gentleman, and Huber 2009)

El paquet arrayQualityMetrics permet realitzar un control de qualitat de dades de *microarray* de manera fàcil, mitjançant l'ús de bàsicament una única funció. Guarda els resultats (imatges i un fitxer html que facilita la comprensió del QC amb interpretacions dels gràfics que es generen) a un directori que l'usuari especifica.

2.2.4 genefilter (Gentleman et al. 2023)

Per a realitzar el filtratge preliminar de gens, utilitzarem el paquet genefilter. Aquest paquet permet utilitzar diferents criteris per a descartar gens que potencialment no ens interessen. Aquests criteris solen estar relacionats a la variabilitat que mostren els gens, si tenen una annotació a ENTREZ, o altres.

Aquest tipus de filtratge es sol dir no específic. Es defineix com a filtratge específic aquell que està relacionat amb els grups experimentals (i.e. que no està diferencialment expressat (DE) en els dos – o més – grups experimentals). En canvi l'inespecífic és el que no està relacionat amb aquests criteris (d'acord a la vignette del paquet).

3 Resultats i discussió

3.1 Control de qualitat

Primer de tot, recullo les dades d'expressió de manera que sigui còmode treballar amb elles.

```
eset_eth <- gse[[1]]
class(eset_eth)
```

```
## [1] "ExpressionSet"
## attr(,"package")
## [1] "Biobase"
```

Veiem que aquest objecte és un ExpressionSet com déu mana.

Per a dur a terme el control de qualitat, faig servir la funció arrayQualityMetrics per a fer el control de qualitat de les dades, però no mostro el codi. He seleccionat quatre gràfics que contenen informació rellevant referent al control de qualitat.

```
require(arrayQualityMetrics)
arrayQualityMetrics(eset_eth, outdir='report_exprsdata', force=T)
```

Com podem veure a la figura 1, els boxplots clarament demostren que les dades estan normalitzades. També he realitzat un *boxplot* "manualment" (fig. 2), però com veieu em surten molts *outliers* (cercles). No sé fins a quin punt això pot resultar preocupant, però ja que totes les mostres mostren un comportament similar, entenc que aquest problema no és sistèmic. Sembla sorprenent que això només em passi utilitzant la funció boxplot, mentre que utilitzant arrayQualityMetrics això no s'observi.

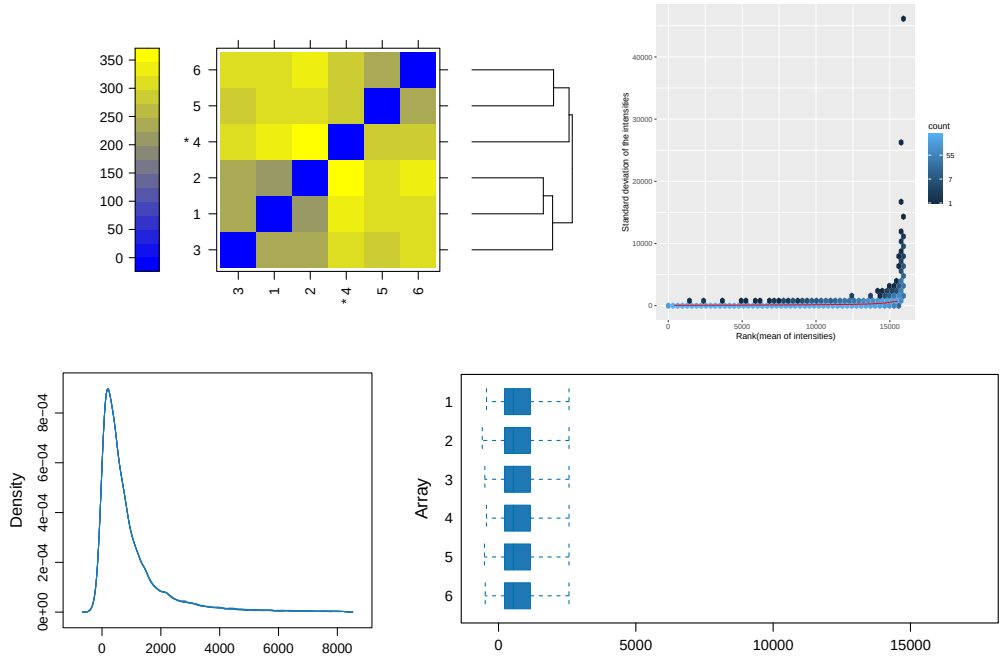


Figure 1: Control de qualitat mitjançant ArrayQualityMetrics

```
boxplot(exprs(eset_eth), which='all', cex.axis=0.6,
        names = colnames(eset_eth), las=1, horizontal=TRUE)
```

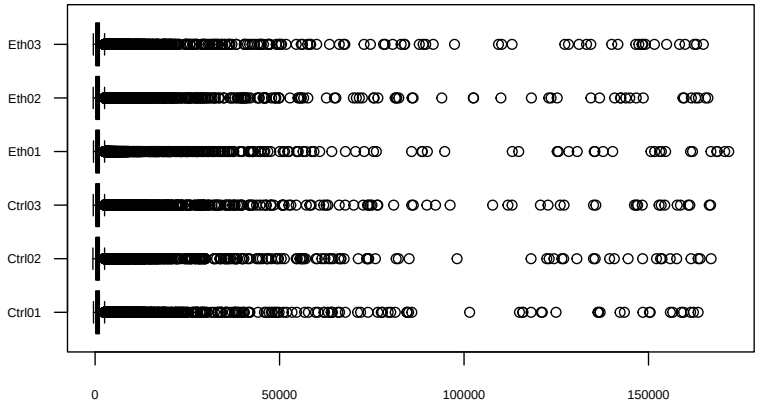


Figure 2: Boxplot de les intensitats de les diferents mostres

Cal destacar també que, al gràfic de dalt a la esquerra de la figura 1, es pot observar que la mostra 4, que correspon a la primera rèplica del grup etanol, presenta una suma de les distàncies ($S_i = \sum_j d_{ij}$, on dij és la distància L_1 entre les mostres i i j) excepcionalment gran.

Utilitzant el codi del professor de l'assignatura (concretament, el de https://github.com/ASPteaching/Analisis_de_datos_omicos-Ejemplo_0-Microarrays), realitzaré un gràfic de les dues components principals.

```
plotPCA(exprs(eset_eth), labels=colnames(eset_eth))
```

Aquest ens permet comprovar si les mostres (o l'expressió que podem observar en aquestes) es comporta com esperaríem, mitjançant un potencial clustering d'acord a les dues components principals (les que expliquen

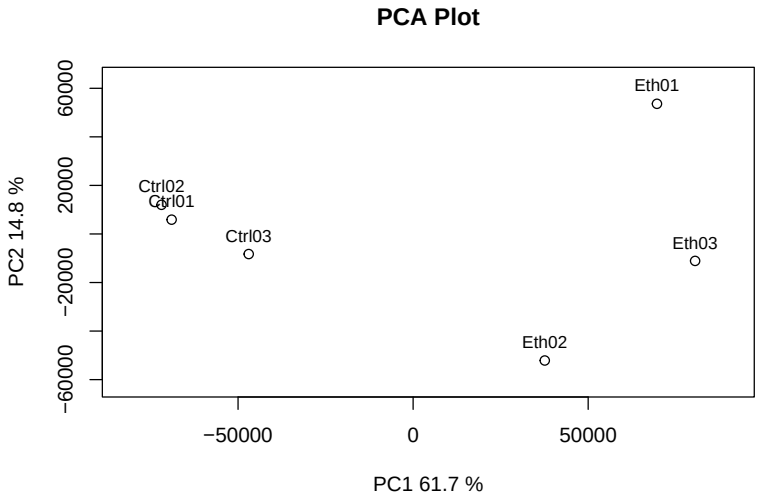


Figure 3: Resultat de graficar les dues principals components de les dades del 'nostre' experiment

més variabilitat del conjunt de dades). Com podem observar, les mostres es disposen segons el seu grup o tractament principalment a través de la component 1, que explica la majoria de la variança de les dades (61.7%). Per altra banda, veiem que les dades control es troben molt més properes entre elles, mentre que les dades del grup etanol es separen principalment al llarg de la PC2.

3.2 Filtrat de gens no específic

Per a realitzar el filtratge de gens no específic, utilitzaré la funció `nsFilter` de `genefilter`. Dintre dels criteris que podem triar, utilitzaré aquells que he vist en els materials de l'assignatura. Quasi sempre es filtren aquells gens que presenten una variabilitat baixa (mesurada amb el `IQR`, o *Inter Quantile Range*), i també aquells gens que no estan annotats a ENTREZ.

```
require(genefilter)
annotation(eset_eth) <- 'rae230a.db'
if (!require(annotation(eset_eth), character.only = T))
  BiocManager::install(annotation(eset_eth))
filtered_eset = nsFilter(eset_eth, var.func=IQR, var.cutoff=0.75, var.filter=TRUE,
                         require.entrez=TRUE, filterByQuantile=TRUE)
```

Com veieu, he marcat com a annotació per al filtratge el paquet `rae230a.db`, que és el tipus de *chip* que s'ha utilitzat per a les mostres d'aquest experiment, segons la pàgina web de GEO. Com veieu, filtrem els gens (o *probes*, *features*) que presenten un IQR (el quantil 0.75 menys el 0.25) menor al IQR que deixa un 75% de IQRs per sota d'aquest. Així doncs, estem descartant un 75% de les dades amb menor variabilitat (definida per IQR).

```
filtered_eset$filter.log
```

```
## $numDupsRemoved
## [1] 2537
##
## $numLowVar
## [1] 8070
##
## $numRemoved.ENTREZID
## [1] 2621
##
## $feature.exclude
```

```
## [1] 6
```

Com podem veure, el número de *probes* filtrades per "LowVar" és 8070. Això inicialment m'ha fet pensar que algo havia fet malament, ja que 8070/nrow(eset_eth) == 0.51. Però, llegint més atentament la documentació de `nsFilter`, trobem que el filtratge de gens degut a la variança es duu a terme en últim lloc, així que el càlcul hauria de ser el següent: 8070/(nrow(eset_eth) - filtered_eset$filter.log$numDupsRemoved, numRemoved.ENTREZID, feature.exclude)). És a dir, hauríem de fer la divisió amb el denominador resultant de fer el filtratge d'acord a tots els altres criteris excloent la variança. Això dóna: 0.75.

Així doncs, s'ha fet el filtratge de manera satisfactòria. Veiem que hem perdut un total de 0. Concretament, degut al criteri de variança baixa s'han filtrat 8070; pel criteri de filtratge de sondes conegudes com a sondes de control de qualitat d'Affymetrix s'han exclòs 6; pel criteri d'exclusió de sondes no anotades per ENTREZ s'han deixat enrere 2621; i, finalment, en quant a sondes duplicades s'han filtrat 2537 sondes.

3.3 *Fitteig* del model lineal

Ara, procedeixo a generar l'objecte corresponent al model lineal que utilitzaré per a trobar els gens diferencialment expressats. Per a això utilitzem el paquet popular `limma`. Creo una matriu de disseny molt facileta, amb dos columnes (una per a cada coeficient corresponent als nivells del factor "tractament" (Control, i tractat amb etanol)), i el mateix número de files com mostres hi ha a l'`ExpressionSet`.

```
require(stringr); require(limma)
myeset <- filtered_eset$eset
groups <- str_replace_all(colnames(myeset), "[:digit:]", "")

design <- model.matrix(~0 + factor(c(1,1,1,2,2,2)))
colnames(design) <- unique(groups)
rownames(design) <- colnames(exprs(myeset))
show(design)
```

```
##        Ctrl Eth
## Ctrl01    1   0
## Ctrl02    1   0
## Ctrl03    1   0
## Eth01     0   1
## Eth02     0   1
## Eth03     0   1
## attr(,"assign")
## [1] 1 1
## attr(,"contrasts")
## attr(,"contrasts")$`factor(c(1, 1, 1, 2, 2, 2))`
## [1] "contr.treatment"
```

Un cop preparada la matriu de disseny, podem realitzar el *fit* del model, amb la funció `lmFit`.

```
require(limma)
fit <- lmFit(myeset, design)
```

Ara, ja que ens interessa comprovar en quins gens hi ha diferències significatives entre els dos grups experimentals, crearem una matriu de contrast que faci aquesta comparació. Com que només tenim un factor amb dos nivells, el número de contrasts només serà 1.

```
contrast.matrix <- makeContrasts(Eth - Ctrl, levels=design)
show(contrast.matrix)
```

```
##       Contrasts
## Levels Eth - Ctrl
##   Ctrl         -1
```

```
##   Eth           1
```

Un cop ho tenim tot preparat, podem procedir a realitzar els contrasts per al model que hem ajustat prèviament. Aplicarem la funció `eBayes`, que permet obtenir *t-stats*, *F-stats* i *log-odds* "moderats" mitjançant tècniques d'estadística Bayesian, tenint en compte la variança dels gens a tot el microarray.

```
require(limma)
fit2 <- contrasts.fit(fit, contrast.matrix)
fit2 <- eBayes(fit2)
```

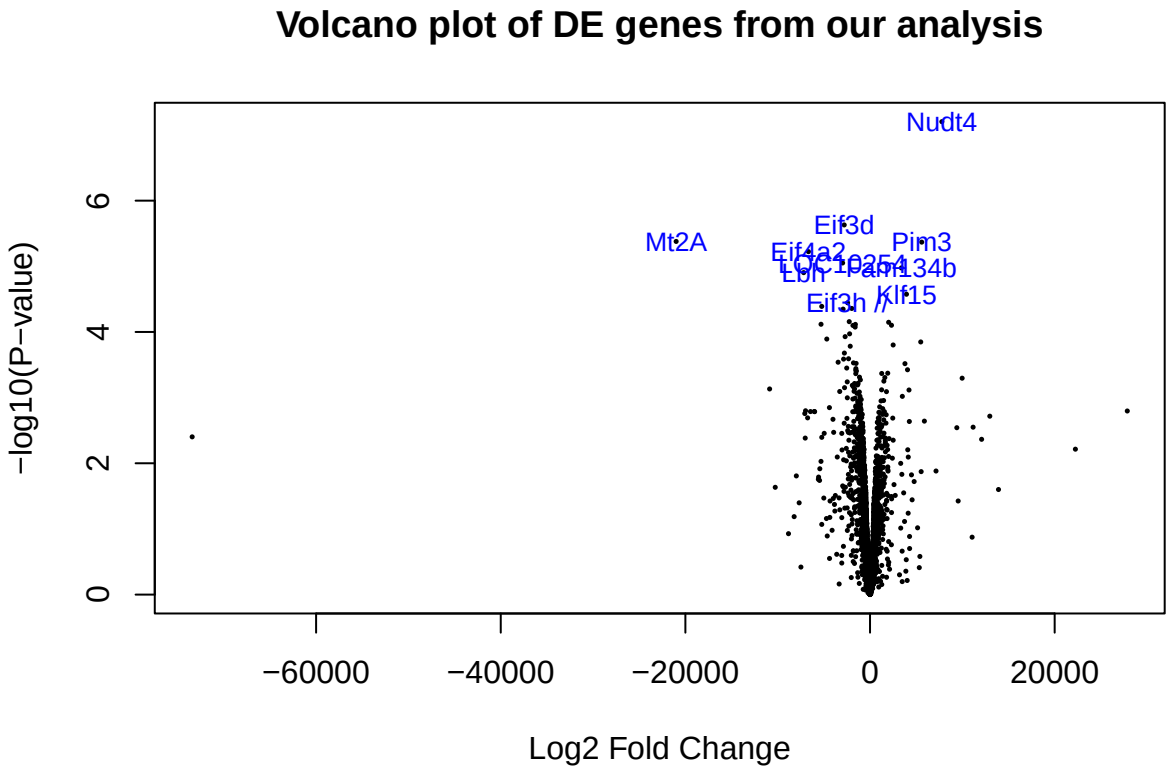
Un cop fet això, podem obtenir la taula amb els gens ordenats segons la seva expressió diferencial, donada pel *B-statistic* o el *p-value*. Com veiem, el *gene symbol* del gen "més diferencialment expressat" correspon a Nudt4.

```
topTabCtrlvsEth <- topTable(fit2, number=nrow(fit2), coef="Eth - Ctrl", adjust="fdr")
head(topTabCtrlvsEth[,c("Gene.Symbol", 'logFC', 't', 'B', 'adj.P.Val')])
```

```
##             Gene.Symbol      logFC         t         B    adj.P.Val
## 1370180_at        Nudt4   7779.667  30.52131 -4.594398 0.0001685876
## 1388568_at        Eif3d  -2795.333 -16.79448 -4.594410 0.0028813172
## 1388271_at         Mt2A -20999.667 -15.25066 -4.594414 0.0028813172
## 1367725_at         Pim3   5599.667  15.18671 -4.594414 0.0028813172
## 1371642_at       Eif4a2  -6670.000 -14.36399 -4.594417 0.0032163563
## 1388900_at LOC102549726  -2967.333 -13.43967 -4.594420 0.0039850084
```

Podem fer una primera visualització dels resultats amb la confecció d'un *volcano plot*.

```
volcanoplot(fit2, highlight=10, names=fit2$genes$`Gene Symbol`,
            main="Volcano plot of DE genes from our analysis")
```



Com és obvi a la figura **??**, hi ha un gen que despunta clarament en quant al seu Log2 FC, amb un valor negatiu < -60000. Podem trobar més informació sobre aquest gen de la manera següent.

```
which.max(abs(topTabCtrlvsEth$logFC))
```

```
## [1] 185
```

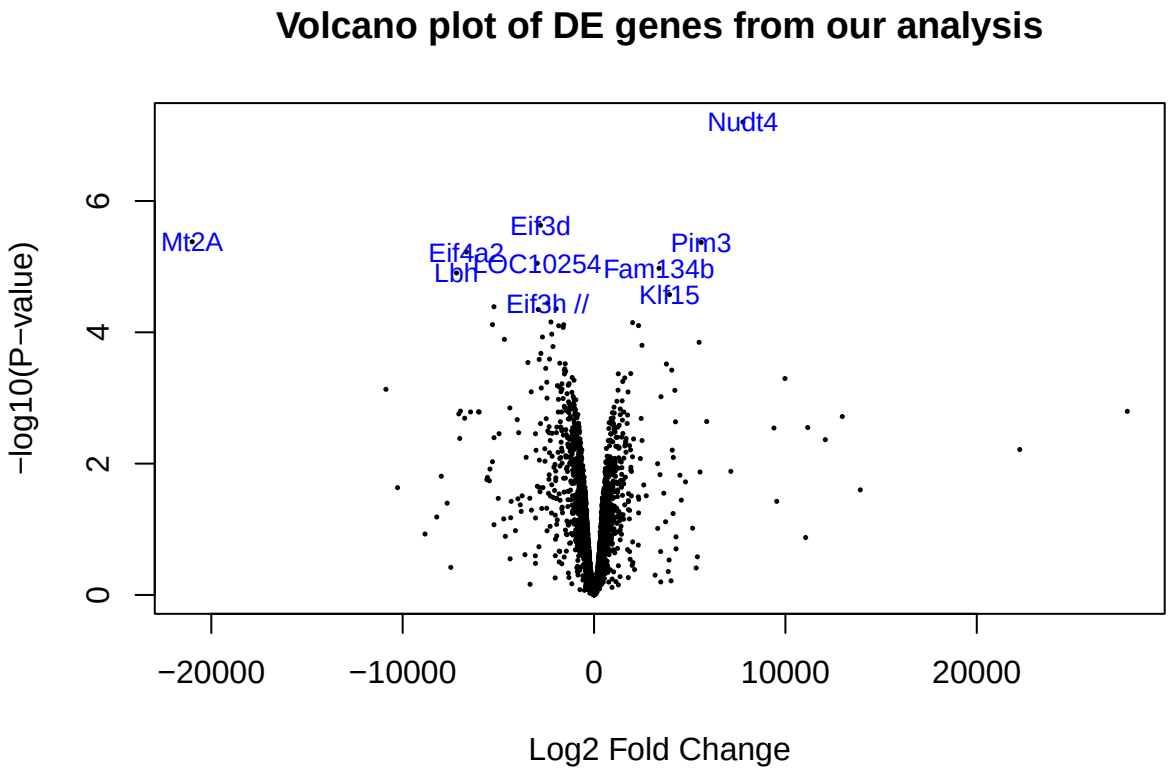
```
topTabCtrlvsEth[185, c('Gene.Symbol', 'logFC', 't', 'adj.P.Val', 'B')]
```

```
##            Gene.Symbol     logFC         t  adj.P.Val        B
## 1368629_at       Reg1a -73422.33 -4.485255 0.05619508 -4.59459
```

Podem veure que aquest gen correspon a la sonda 1368629_at, amb el *Gene Symbol* Reg1a, i que no assoleix un p-valor significatiu.

Si el treiem del *volcano plot*:

```
ix = which(fit2$genes$ID=="1368629_at")
fit2withoutoutlier <- fit2[-ix,]
volcanoplot(fit2withoutoutlier, highlight=10, names=fit2withoutoutlier$genes$`Gene Symbol`,
            main="Volcano plot of DE genes from our analysis")
```



D'aquesta manera podem veure els 10 gens més diferencialment expressats. Només per curiositat, vull veure el SE del gen que té el |logFC| tan gran, per esbrinar el motiu de que no aparegui amb un p-valor significatiu.

```
fit2$s2.post['1368629_at']
```

```
## 1368629_at
##  401951206
```

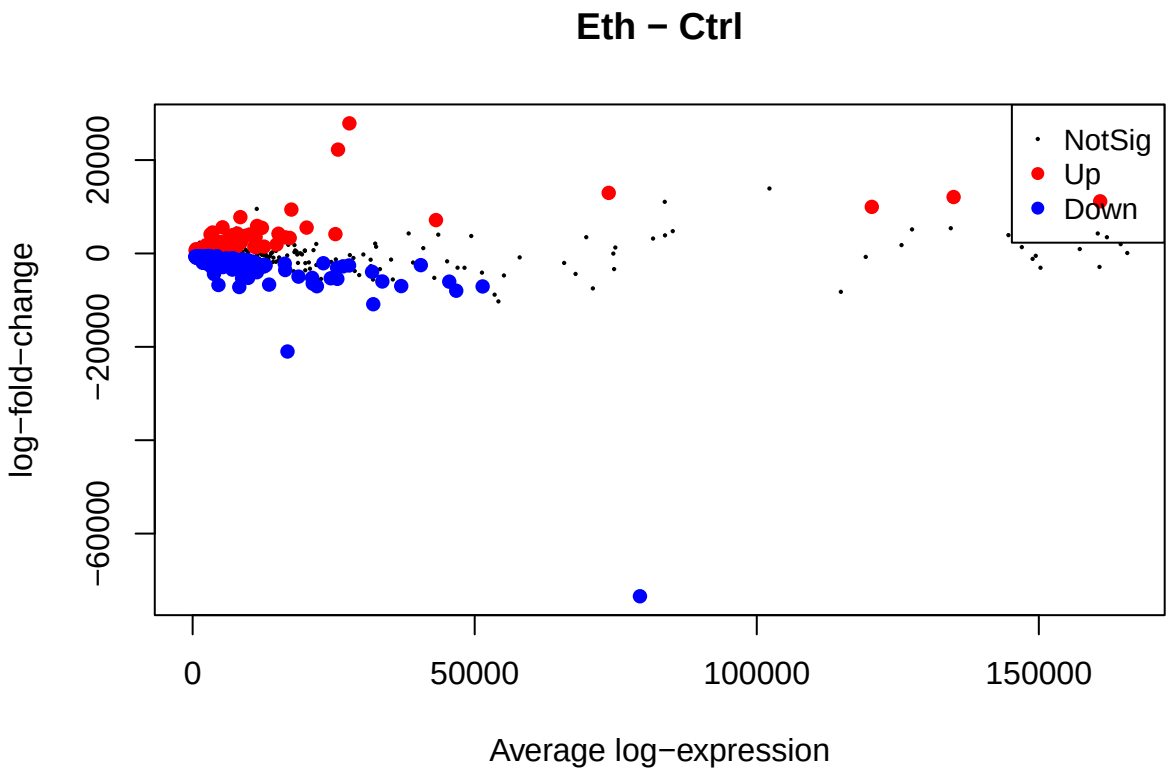
```
which.max(fit2$s2.post)
```

```
## 1368629_at
##        435
```

Com veiem, té un valor del error estàndard corresponent a la distribució posterior per a la variança bastant gran (el major de tot el *fit*).

Després d'obtenir aquesta taula, podem obtenir un vector o llista amb els gens expressats diferencialment mitjançant la funció `decideTests` del paquet `limma`. Podem utilitzar aquest

```
results<-decideTests(fit2, adjust.method="fdr", p.value=0.1, lfc=1)
plotMD(fit2, status = results)
```




```
topTabCtrlvsEth[185, ]
```

```
##                     ID    GB_ACC SPOT_ID Species.Scientific.Name Annotation.Date
## 1368629_at 1368629_at NM_012641               Rattus norvegicus     Oct 6, 2014
##                 Sequence.Type Sequence.Source
## 1368629_at Consensus sequence         GenBank
##
## 1368629_at gb:NM_012641.1 /DB_XREF=gi:6981469 /GEN=Reg1 /FEA=FLmRNA /CNT=5 /TID=Rn.11332.1 /TIER=FL /
##            Representative.Public.ID                          Gene.Title
## 1368629_at                NM_012641 regenerating islet-derived 1 alpha
##            Gene.Symbol ENTREZ_GENE_ID RefSeq.Transcript.ID
## 1368629_at       Reg1a          24714            NM_012641
##            Gene.Ontology.Biological.Process
## 1368629_at
##
## 1368629_at 0005576 // extracellular region // inferred from electronic annotation /// 0070062 // extr
##
## 1368629_at 0008083 // growth factor activity // inferred from direct assay /// 0008083 // growth fact
##                 logFC  AveExpr         t     P.Value  adj.P.Val        B
## 1368629_at -73422.33 79294.17 -4.485255 0.003957507 0.05619508 -4.59459
```

Referències

Davis, Sean, and Paul Meltzer. 2007. "GEOquery: A Bridge Between the Gene Expression Omnibus (GEO) and BioConductor." *Bioinformatics* 14: 1846–47.

Edgar, R. 2002. "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository." *Nucleic Acids Research* 30 (1): 207–10. https://doi.org/10.1093/nar/30.1.207.

Gentleman, Robert, Vincent J. Carey, Wolfgang Huber, and Florian Hahne. 2023. *Genefilter: Genefilter: Methods for Filtering Genes from High-Throughput Experiments*. https://doi.org/10.18129/B9.bioc.genefilter.

Huber, W., V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, et al. 2015. "Orchestrating High-Throughput Genomic Analysis with Bioconductor." *Nature Methods* 12 (2): 115–21. http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html.

Kauffmann, Audrey, Robert Gentleman, and Wolfgang Huber. 2009. "arrayQualityMetrics–a Bioconductor Package for Quality Assessment of Microarray Data." *Bioinformatics* 25 (3): 415–16.

Kubisch, Constanze H., Ilya Gukovsky, Aurelia Lugea, Stephen J. Pandol, Rork Kuick, David E. Misek, Samir M. Hanash, and Craig D. Logsdon. 2006. "Long-Term Ethanol Consumption Alters Pancreatic Gene Expression in Rats: A Possible Connection to Pancreatic Injury." *Pancreas* 33 (1): 68–76. https://doi.org/10.1097/01.mpa.0000226878.81377.94.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015. "limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.