Prova d'Avaluació Continuada 1

Joan Padrosa Pulido

Contents

1 Abstract

El present treball té com a objectiu determinar l'expressió diferencial de gens en cultius de càncer de pròstata en funció de si s'han incubat amb àcid araquidònic o no. Per tal de dur a terme l'anàlisi, despres de l'exposició, es va obtenir l'ARN total i es va analitzar l'expressió gènica mitjançant *microarrays*, 4 rèpliques per nivell del tractament. En el preprocessament de l'anàlisi s'han objectivat problemes importants, sobretot una sospita elevada de fonts de variabilitat importants alternatives al tractament a estudi, probablement en relació a *batch effect*, el que posa en dubte qualsevol condició extreta de l'estudi. Les mostres s'han comparat mitjançant un anàlisi basat en models lineals, i s'han analitzat els processos implicats en funció dels termes GO relacionats.

2 Objectius

L'objectiu del treball és realitzar un anàlisi de *microarrays* per determinar si l'addició d'àcid araquidònic, un àcid gras ω-6, indueix expressió diferencial de gens en cultius cel · lulars de càncer de pròstata (PC-3).

3 Materials i mètodes

S'han obtingut els fitxers .CEL continents dels valors d'expressió crus de l'anàlisi de *microarrays* titulat *Arachidonic acid effect on prostate cancer cells*, publicat per Hughes-Fulford *et al.* al 2006[1], descarregats des de la base de dades de *Gene Expression Omnibus* (GEO)[2], mitjançant l'identificador GDS1736.

En l'estudi, s'incubaren cèl · lules de càncer de pròstata PC-3 amb àcid araquidònic a 5 $\mu g/mL$ en un medi RPMI amb 0.25 mg/mL d'albúmina durant dues hores, mentre que el grup control s'incubà només amb albúmina. Es realitzaren 4 rèpliques per cada nivell, de les quals s'obtingué l'ARN total i l'expressió gènica relativa es va analitzar mitjançant *arrays* d'Affymetrix.

L'anàlisi s'ha dut a terme mitjançant el programari R 4.0.3[3] i les eines de l'entorn `Bioconductor`[4] .

Les dades crues, després d'un anàlisi de qualitat, han estat normalitzades mitjançant el mètode RNA, que consisteix en tres passes:

- Correcció del fons.
- Normalització.
- Resum de valors del grup en un únic valor d'expressió absoluta.

Les dades obtingudes mitjançant aquest procés han estat les utilitzades per a realitzar tot l'anàlisi. S'ha realitzat un filtratge no específic per eliminar els gens de baixa variabilitat, seleccionant aquells gens amb variabilitat (per rang interquartílic) superior al percentil 75 de la mostra o sense entrada a Entrez.

Posteriorment, mitjançant el mètode desenvolupat per Smyth *et al.*[5], s'han seleccionat els gens diferencialment expressats, utilitzant el mètode de Benjamini i Hochberg per a ajustar la significació estadística per les comparacions múltilpes[6]. La distribució d'aquests gens s'ha visualitzat mitjançant un *volcano plot*. Els gens diferencialment expressats s'han agrupat per a buscar patrons comuns d'expressió per a cada un dels grups, utilitzant mapes de colors.

Per a l'anotació de gens s'han utilitzat bases de dades com Entrez, i posteriorment s'ha realitzat un anàlisi d'enriquiment per a identificar els processos afectats amb més freqüència.

[1] Hughes-Fulford M, Li CF, Boonyaratanakornkit J, Sayyah S. Arachidonic acid activates phosphatidylinositol 3-kinase signaling and induces gene expression in prostate cancer. Cancer Res 2006 Feb 1;66(3):1427-33. PMID: 16452198

[2] https://www.ncbi.nlm.nih.gov/sites/GDSbrowser

[3] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[4] Orchestrating high-throughput genomic analysis with Bioconductor. W. Huber, V.J. Carey, R. Gentleman, . . . , M. Morgan Nature Methods, 2015:12, 115.

[5] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol, 3:Article3, 2004.

[6] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B, 57:289–300, 1995.

4 Resultats

A continuació es mostra, pas per pas, l'anàlisi realitzat i els resultats obtinguts.

4.1 Identificació de mostres i classificació en grups

Per a començar l'anàlisi, definim el directori de treball. Aquest directori conté una carpeta, `Dades`, amb el fitxer .tar que conté els fitxers .CEL descarregat prèviament, i una carpeta `Resultats`, buida. També conté aquest fitxer en format .pdf i en format .Rmd, així com l'enunciat de la prova d'avaluació continuada.

```
setwd("~/OneDrive/UOC/Assignatures/Anàlisi de dades òmiques/PAC1/omiques/")
```

Descomprimim el fitxer .tar, eliminem tots els fitxers que no siguin .CEL i mostrem els fitxers resultants:

```
tar -xf ./Dades/GSE3737_RAW.tar -C Dades
rm ./Dades/*EXP.gz
gzip -df ./Dades/*.gz
ls ./Dades | grep .CEL
```

```
## GSM86079.CEL
## GSM86080.CEL
## GSM86081.CEL
## GSM86082.CEL
## GSM86083.CEL
## GSM86084.CEL
## GSM86085.CEL
## GSM86086.CEL
```

Generem el fitxer `targets`, per a definir, per a cada fitxer .CEL, el grup al què pertany:

```
files<-paste("GSM860",79:86,sep="")
group<-c(rep("Control",4),rep("Tractament",4))
name<-c(paste("CTRL",1:4,sep=""),paste("AA",1:4,sep=""))
treatment<-c(rep("Cap",4),rep("Àcid Araquidònic",4))
targets<-data.frame(Fitxer=files,Grup=group,
                    Tractament=treatment,
                    Nom=name)
readr::write_csv(targets,file="./Dades/targets.csv")
knitr::kable(targets,
             caption="Fitxer targets")
```

Table 1: Fitxer targets

Fitxer	Grup	Tractament	Nom
GSM86079	Control	Cap	CTRL1
GSM86080	Control	Cap	CTRL2
GSM86081	Control	Cap	CTRL3
GSM86082	Control	Cap	CTRL4
GSM86083	Tractament	Àcid Araquidònic	AA1
GSM86084	Tractament	Àcid Araquidònic	AA2
GSM86085	Tractament	Àcid Araquidònic	AA3
GSM86086	Tractament	Àcid Araquidònic	AA4

A continuació, importem els fitxers .CEL, creant un objecte tipus `ExpressionSet` per a contenir les dades.

```
library(Biobase)
library(oligo)
# Importem les ubicacions dels fitxers:
celFiles<-list.celfiles("./Dades",full.names=T)
#Creem un ExpressionSet a partir del fitxer "targets"
my.targets<-read.AnnotatedDataFrame("./Dades/targets.csv",
                                    header=T, row.names=1,
                                    sep=",")
rawData<-read.celfiles(celFiles,phenoData = my.targets)
```

```
## Reading in : ./Dades/GSM86079.CEL
## Reading in : ./Dades/GSM86080.CEL
## Reading in : ./Dades/GSM86081.CEL
## Reading in : ./Dades/GSM86082.CEL
## Reading in : ./Dades/GSM86083.CEL
## Reading in : ./Dades/GSM86084.CEL
## Reading in : ./Dades/GSM86085.CEL
## Reading in : ./Dades/GSM86086.CEL
```

```
#Canviem els noms per simplicitat posterior:
rownames(pData(rawData))<-targets$Nom
colnames(rawData)<-rownames(pData(rawData))
```

4.2 Control de qualitat de les dades crues

Per al control de qualitat de les dades crues utilitzem la funció `arrayQualityMetrics()` sobre les dades crues.

```
library(arrayQualityMetrics)
arrayQualityMetrics(rawData,
                    outdir = "./Resultats/qualitat-crues",
                    force=T)
```

Els gràfics generats, així com l'informe, són emmagatzemats a la carpeta `Resultats/qualitat-dades-crues/`, disponible al repositori de `Github`.

	array	sampleNames	*1	*2	*3	Grup	Nom
☐	1	CTRL1			x	Control	CTRL1
☐	2	CTRL2			x	Control	CTRL2
☐	3	CTRL3				Control	CTRL3
☐	4	CTRL4				Control	CTRL4
☐	5	AA1			x	Tractament	AA1
☐	6	AA2		x	x	Tractament	AA2
☐	7	AA3			x	Tractament	AA3
☐	8	AA4			x	Tractament	AA4

Figure 1: Resum de l'anàlisi de les dades crues

Podem veure, en el resum, com no hi ha problemes significatius en les dades abans de la normalització, tot i que en algun xip es detecten *outliers* per dos de tres mètodes (Figura 1).

Podem fer l'anàlisi de components principals (els 2 primers) per a veure si els resultats s'agrupen per grup (seria esperable) o si no és així. Com que aquest anàlisi es repetirà, creem una funció que el faci directament, que anomenem `PCA()`.

Aplicant la funció es pot fer, visualment, un anàlisi de components principals (Figura 2).

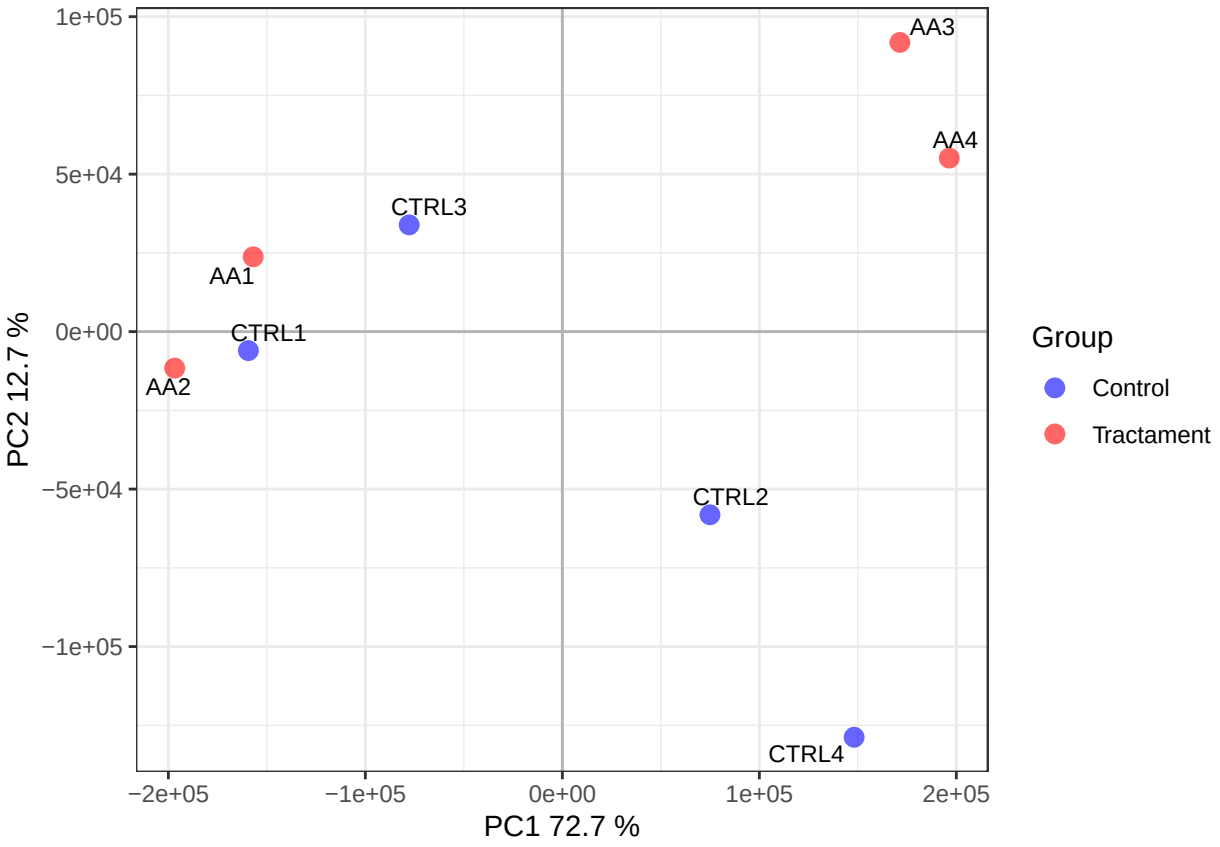

Figure 2: Anàlisi de components principals per les dades crues

Es pot intuir que hi ha una agrupació natural en l'eix vertical en funció del tractament amb àcid araquidònic, però el segon component principal només justifica el 12.7% de la variabilitat. La major part de la variabilitat, explicada pel primer component principal, depèn d'algun factor diferent, ja que no s'observa una agrupació per grup en l'eix horitzontal.

Representem cada *array* mitjançant *boxplot* (Figura 3). Observem variabiltiat entre mostres, esperable en les dades crues.

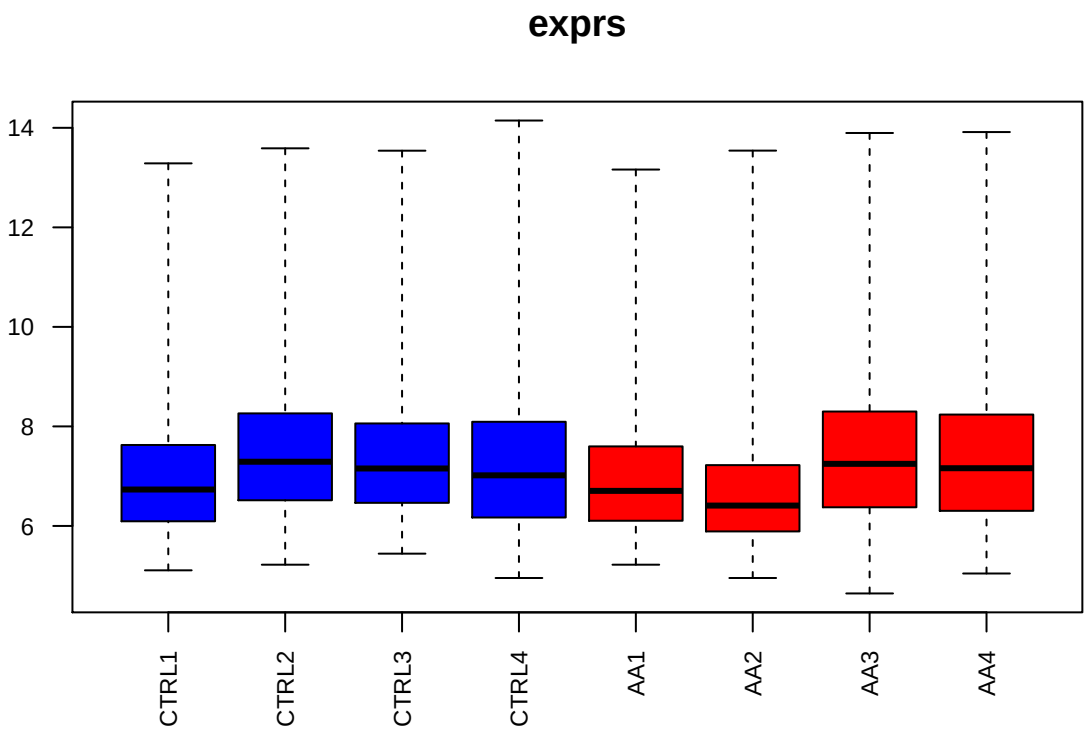

Figure 3: Boxplot d'expressió de cada mostra

4.3 Normalització de les dades

Un cop acceptada la qualitat de les dades, procedim a la seva normalització mitjançant la funció `rma()`.

```
normalitzades<-rma(rawData)
```

```
## Background correcting
## Normalizing
## Calculating Expression
```

4.4 Control de qualitat de les dades normalitzades

Tornem a fer el control de qualitat de les dades normalitzades.

```
arrayQualityMetrics(normalitzades,
                    outdir="./Resultats/qualitat-normalitzades",
                    force=T)
```

	array	sampleNames	*1	*2	*3	Grup	Nom
☐	1	CTRL1				Control	CTRL1
☐	2	CTRL2				Control	CTRL2
☐	3	CTRL3				Control	CTRL3
☐	4	CTRL4				Control	CTRL4
☐	5	AA1				Tractament	AA1
☐	6	AA2				Tractament	AA2
☐	7	AA3				Tractament	AA3
☐	8	AA4				Tractament	AA4

Figure 4: Resum de l'anàlisi de les dades normalitzades

En l'anàlisi de les dades normalitzades no s'observen problemes de qualitat de les mostres (Figura 4). Repetim l'anàlisi de components principals (Figura 5).

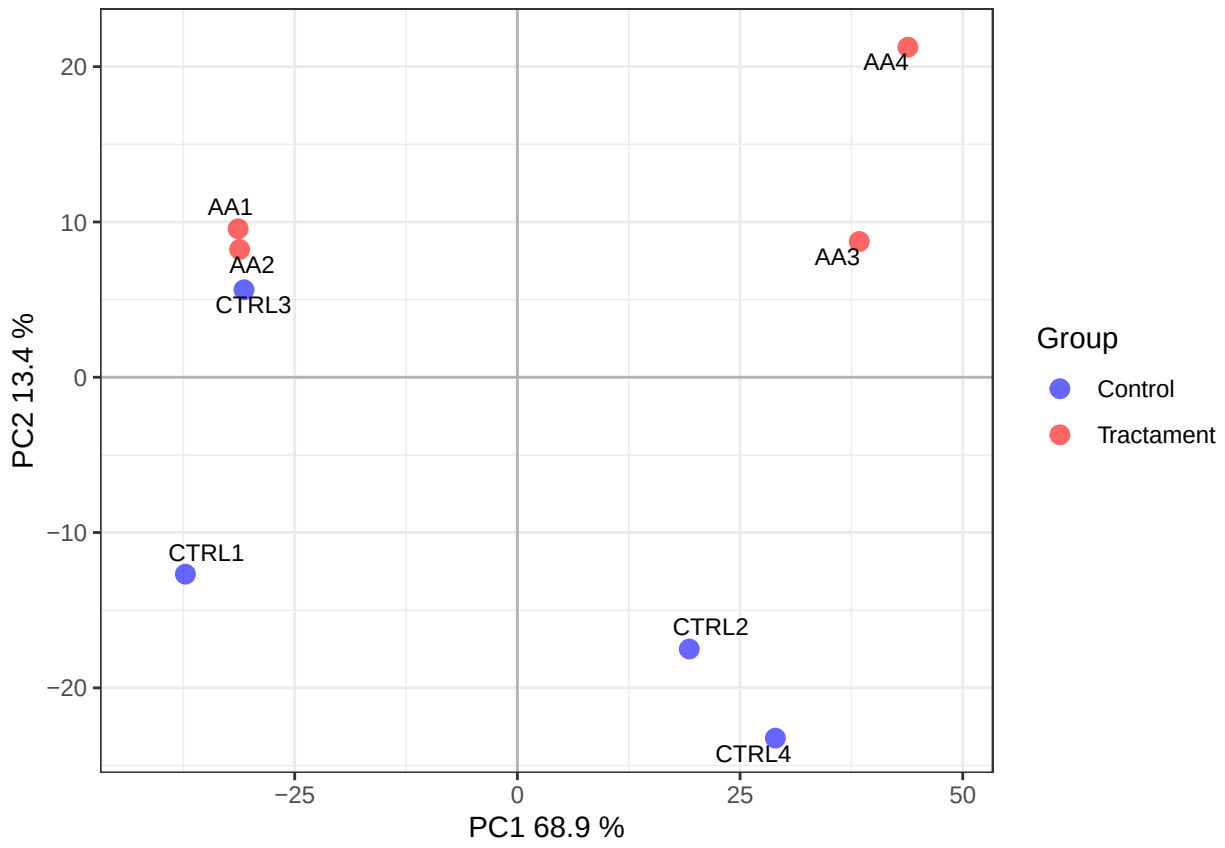

Figure 5: Anàlisi de components principals per les dades normalitzades

Un cop normalitzades les dades, el segon component explica una part discretament de la variabiltiat, i l'agrupació en l'eix vertical és més evident, però es segueix observant una absència d'agrupació en l'eix horitzontal. Quant als *boxplots*, després de la normalització mitjançant `RMA`, que utilitza normalització de quantils, són idèntics, com és esperable.

A continuació, comprovem, per a descartar algun efecte tipus *batch effect*, es poden analitzar les dates de processament de cada *array* mitjançant la funció `get.celfile.dates()`.

```
## [1] "2003-07-31" "2003-01-08" "2003-07-31" "2004-09-03" "2003-07-31"
## [6] "2003-07-31" "2004-09-03" "2004-09-03"
```

Les mostres han estat processades en dates diferents, pel que podem, fàcilment, tornar a realitzar l'anàlisi de components principals tenint en compte la data de processament com si fos un altre tractament. Primer de tot, creem un nou fitxer anomenat `targets1`, pque inclogui la data com a criteri de creació de grups.

Table 2: Fitxer targets modificat amb la data de processament com a factor

Fitxer	Grup	Tractament	Data	Nom
GSM86079	CTRL.2003	Cap	2003	CTRL1
GSM86080	CTRL.2003-1	Cap	2003-1	CTRL2

Fitxer	Grup	Tractament	Data	Nom
GSM86081	CTRL.2003	Cap	2003	CTRL3
GSM86082	CTRL.2004	Cap	2004	CTRL4
GSM86083	AA.2003	Àcid Araquidònic	2003	AA1
GSM86084	AA.2003	Àcid Araquidònic	2003	AA2
GSM86085	AA.2004	Àcid Araquidònic	2004	AA3
GSM86086	AA.2004	Àcid Araquidònic	2004	AA4

Creem el nou `ExpressionSet`:

```
#Creem un nou ExpressionSet:
my.targets1<-read.AnnotatedDataFrame("./Dades/targets1.csv",
                                     header=T, row.names=1,
                                     sep=",")
rawData_data<-read.celfiles(celFiles,phenoData = my.targets1)
```

```
## Reading in : ./Dades/GSM86079.CEL
## Reading in : ./Dades/GSM86080.CEL
## Reading in : ./Dades/GSM86081.CEL
## Reading in : ./Dades/GSM86082.CEL
## Reading in : ./Dades/GSM86083.CEL
## Reading in : ./Dades/GSM86084.CEL
## Reading in : ./Dades/GSM86085.CEL
## Reading in : ./Dades/GSM86086.CEL
```

```
#Canviem els noms per simplicitat posterior:
rownames(pData(rawData_data))<-targets1$Nom
colnames(rawData_data)<-rownames(pData(rawData_data))
#Normalitzem les dades:
normalitzades_data<-rma(rawData_data)
```

```
## Background correcting
## Normalizing
## Calculating Expression
```

Amb el nou fitxer, tornem a realitzar l'anàlisi de components principals (Figura 7).

Es pot observar com en l'eix horitzontal, a l'esquerra queden les mostres processades al juliol de 2003, al centre la mostra processada a l'agost del mateix any i a la dreta les mostres processades al setembre 2004, i que el primer component principal justifica, en aquest cas, el 68.9% de la variabilitat. Aquestes troballes suggereixen que la data podria jugar un paper rellevant en els resultats. Utilitzem un anàlisi tipus *Principal Variation Component Analysis* (PVCA) per a comprovar d'on prové la variabilitat, utilitzant la data com un factor (Figura 8).

Tal com es preveia, la major part de la variabilitat (>80%) ve explicada per la data en què s'ha cursat la mostra. Tot això és suggestiu d'un efecte tipus *batch effect* molt marcat, i podria invalidar els resultats.

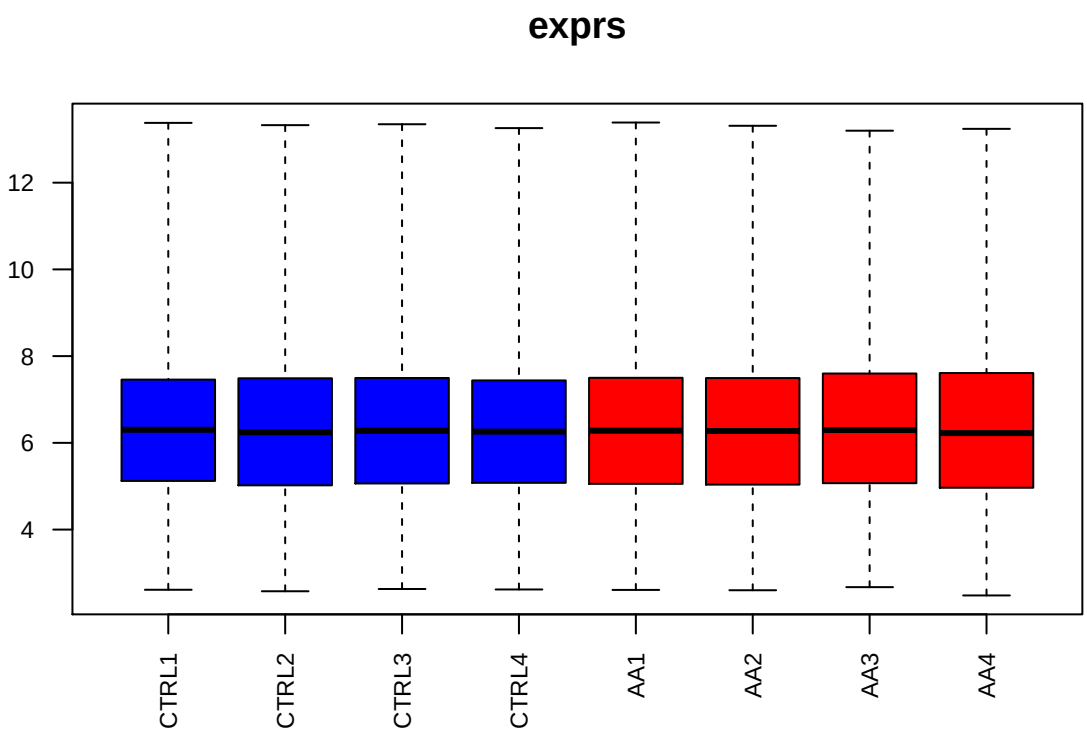

Figure 6: Boxplot de les dades normalitzades

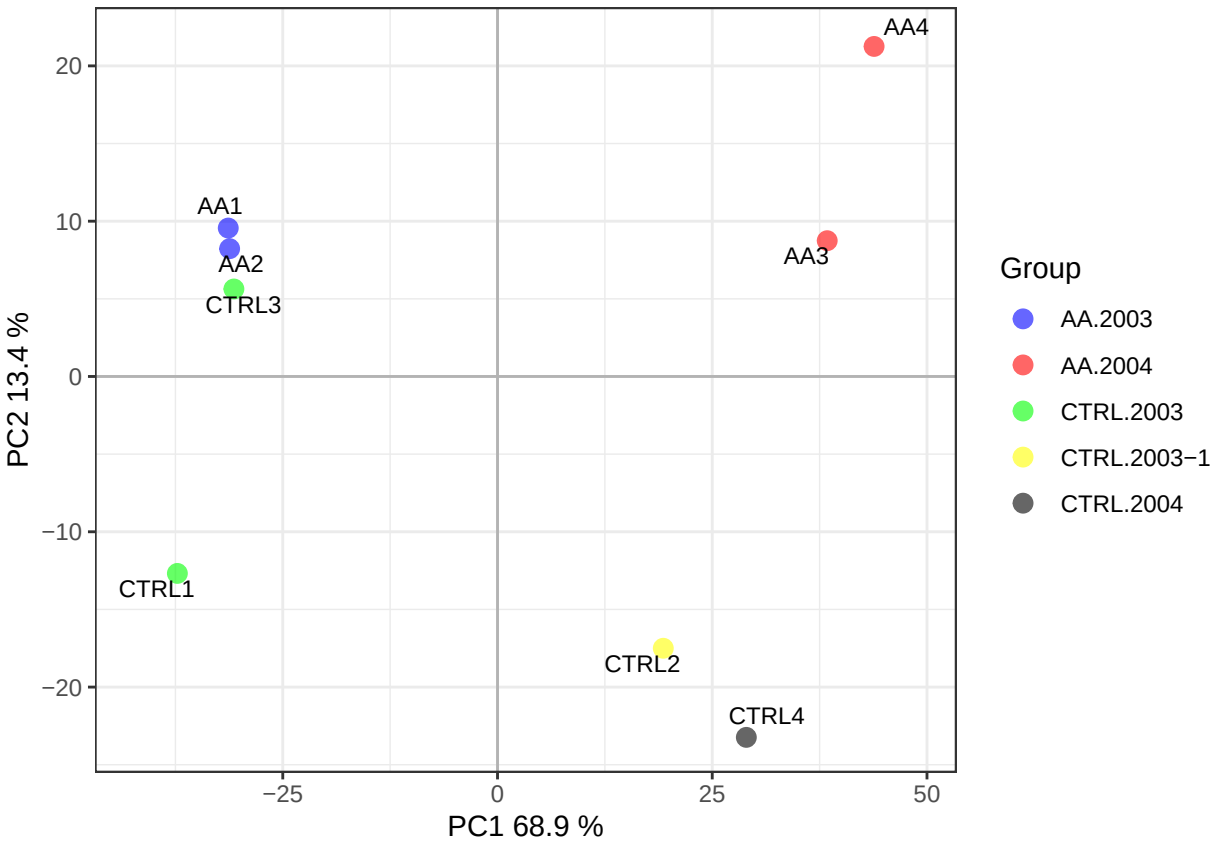

Figure 7: Anàlisi de components principals amb la data com a font de variabilitat

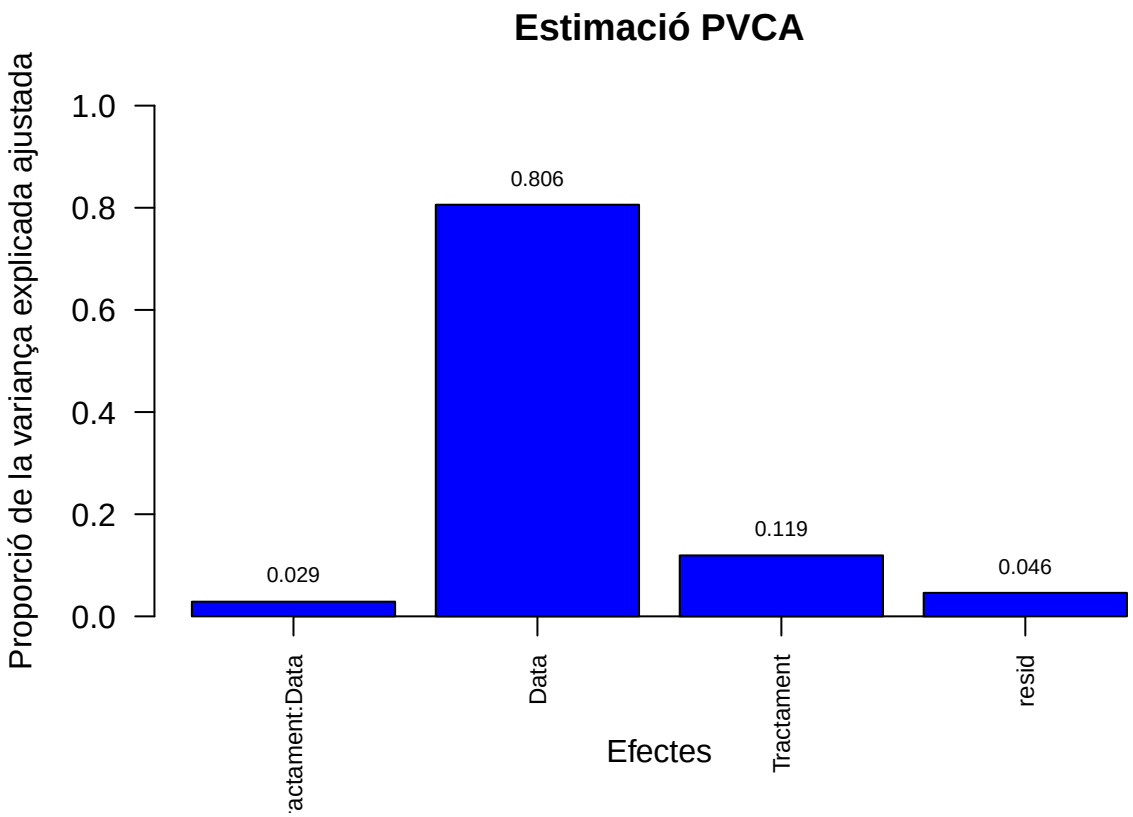

Figure 8: Anàlisi PVCA

4.5 Filtratge no específic

A continuació es busquen els gens amb més variabilitat, utilitzant les desviacions estàndar per files (cada columna representa una mostra, i cada fila un transcrit). Les representem gràficament, amb cada gen a l'eix d'ordenades i línies verticals marcant els percentils 90 i 95 (Figura 9).

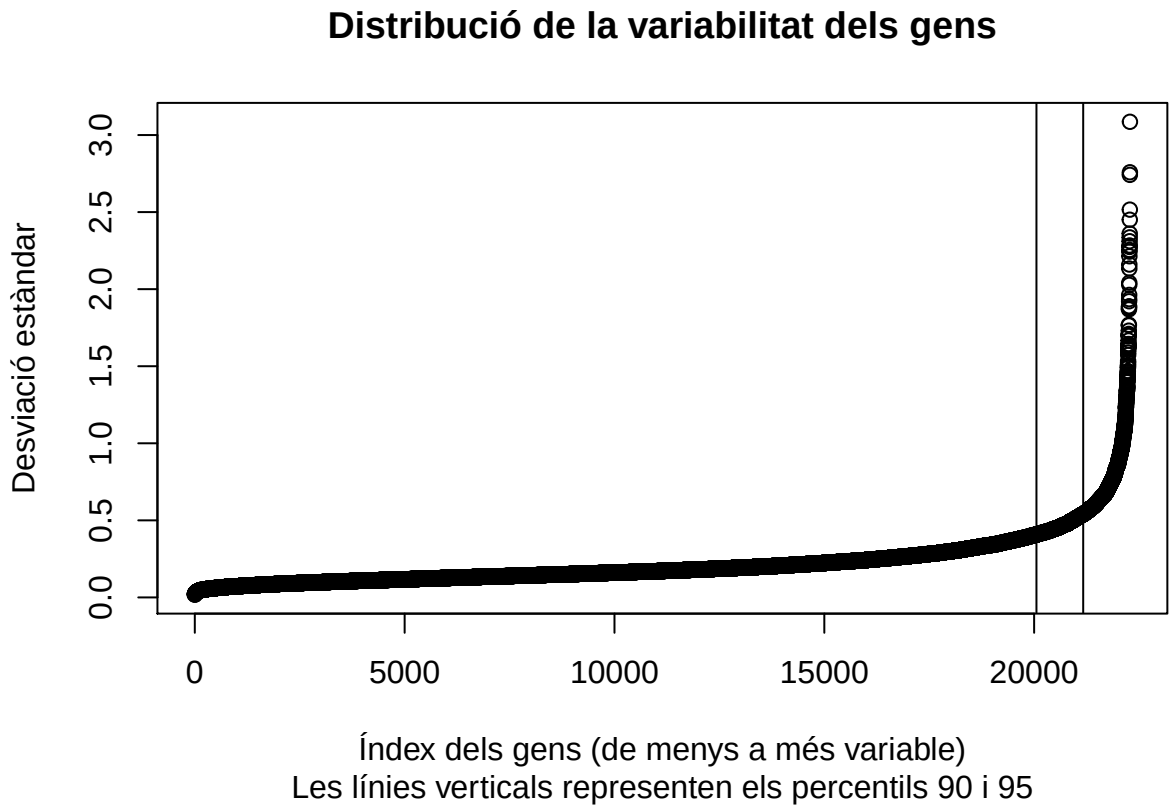

Figure 9: Variabilitat dels gens

Eliminem els gens amb menys variabilitat per tal de disminuir el nombre de comparacions posteriors. S'utilitza un paquet específic, `genefilter`. A part, podem traure aquells transcrits que no tinguin un identificador de gen associat.

```
library(genefilter);library(hgu133a.db)
#Marquem les anotacions:
annotation(normalitzades)<-"hgu133a.db"
#Filtrem per variança i ID a Entrez.
filtrades<-nsFilter(normalitzades,
                    require.entrez=T,remove.dupEntrez=T,
                    var.filter=T,var.func=IQR,var.cutoff=0.75,
                    filterByQuantile=T,feature.exclude="^AFFX")
print(filtrades$filter.log)
```

```
## $numDupsRemoved
## [1] 7399
##
## $numLowVar
```

```
## [1] 9301
##
## $numRemoved.ENTREZID
## [1] 2472
##
## $feature.exclude
## [1] 10
```

```
#Guardem l'expressionSet
gens_filtrats<-filtrades$eset
```

Es pot veure com hem retirat 7399 transcrits duplicats, 9301 transcrits amb variança baixa, 2472 transcrits sense entrada a *Entrez* i 10 registres que corresponen a detalls tècnics del *microarray*. Queden un total de 3101 transcrits per analitzar a la mostra.

A continuació guardem les dades a la carpeta `Resultats`, tant en forma de fitxer .csv com de fitxer .R per a consultar amb facilitat.

```
write.csv(exprs(normalitzades),file="./Resultats/dades_normalitzades.csv")
write.csv(exprs(gens_filtrats),file="./Resultats/dades_normalitzades_filtrades.csv")
save(normalitzades, gens_filtrats, file="./Resultats/normalized.data.Rda")
```

4.6 Identificació de gens diferencialment expressats

Per tal d'analitzar l'expressió diferencial de gens, s'utilitzarà un mètode basat en models lineals, utilitzant les funcions del paquet `limma`, implementat a `Bioconductor`, carregat prèviament.

4.6.1 Definició de la matriu de disseny:

Per a utilitzar-lo, és necessari definir la matriu de disseny i la matriu de contrasts, encara que en el nostre anàlisi només realitzarem una comparació. La matriu de disseny té tantes files com mostres i columnes com grups. Cada fila conté un 1 en la columna del grup a la qual pertany i un 0 en la resta. Utilitzem la variable "Grup" del factor `targets` que separa en funció de condició experimental (tractament i control).

```
disseny<-model.matrix(~0+Grup,pData(gens_filtrats))
colnames(disseny)<-c("CTRL","AA")
disseny
```

```
##       CTRL AA
## CTRL1    1  0
## CTRL2    1  0
## CTRL3    1  0
## CTRL4    1  0
## AA1      0  1
## AA2      0  1
## AA3      0  1
## AA4      0  1
## attr(,"assign")
## [1] 1 1
## attr(,"contrasts")
## attr(,"contrasts")$Grup
## [1] "contr.treatment"
```

4.6.2 Definició de la matriu de contrasts:

La matriu de contrasts descriu les comparacions entre grups, on cada fila és un grup i cada columna una comparació. En aquest cas només fem una comparació, i no hi ha grup interacció.

```
library(limma)
contrasts<-makeContrasts(AAvsCTRL = AA-CTRL,
                         levels=disseny)
contrasts
```

```
##       Contrasts
## Levels AAvsCTRL
##   CTRL       -1
##   AA          1
```

4.6.3 Estimació del model i selecció de gens:

Amb les matrius i les dades, estimem el model mitjançant el paquet `limma`.

```
fit<-lmFit(gens_filtrats,disseny)
fit.main<-contrasts.fit(fit,contrasts)
fit.main<-eBayes(fit.main)
```

La funció `topTable()`de `limma` permet veure, ordenats de manera descendent en funció del p-valor, els gens diferencialment expressats en cada comparació. Ajustem el p-valor mitjançant el mètode de Benjamini i Hochberg mitjançant l'argument `adjust`.

```
taula<-topTable(fit.main,number=nrow(fit.main),
                coef="AAvsCTRL",adjust="fdr")
knitr::kable(head(taula),
             caption="Expressió diferencial de gens")
```

Table 3: Expressió diferencial de gens

	logFC	AveExpr	t	P.Value	adj.P.Val	B
209774_x_at	5.072934	8.474089	16.189816	0.0e+00	0.0000024	11.316764
204470_at	4.211596	8.735476	14.780252	0.0e+00	0.0000036	10.592752
211506_s_at	3.505208	8.032616	13.337346	0.0e+00	0.0000082	9.729091
207850_at	4.537659	7.163907	10.796657	1.0e-07	0.0000732	7.827364
201502_s_at	2.945333	10.208619	8.678451	1.1e-06	0.0006827	5.756761
210229_s_at	1.957622	7.717277	8.211583	2.0e-06	0.0010381	5.226944

A continuació es dibuixa un *volcano plot* per tal a representar, de manera gràfica, la quantitat de gens diferencialment expressats i la magnitud de la diferència (Figura 10).

La línia horitzontal representa la significació estadística considerada com a p<0.05. Sorprenentment, s'observa que el gràfic és molt asimètric, el que no té massa sentit a nivell biològic, ja que implicaria que un dels grups té molts gens activats en relació a l'altre, sense tenir quasi gens menys expressats. Es pot veure, a més, com no hi ha una quantitat de gens massa gran expressats diferencialment. Per a una p no ajustada menor a 0.25 hi ha 231 gens diferencialment expressats, i si es selecciona una p menor de 0.1 només hi ha 68 gens.

De cara a il · lustrar l'anàlisi d'enriquiment, es seleccionen el màxim de gens possibles per al treball, utilitzant una p menor de 0.25, sense límit de significació biològica i sense mètode d'ajustament del p-valor.

```
res<-decideTests(fit.main,adjust.method="none",
                 p.value=0.25)
sum.res.rows<-apply(abs(res),1,sum)
res.selected<-res[sum.res.rows!=0,]
summary(res.selected)
```

```
##        AAvsCTRL
## Down         13
## NotSig        0
## Up          218
```

Com s'intuïa, hi ha 231 gens expressats diferencialment entre els dos grups utilitzant aquest criteri tan poc restrictiu, quasi tots ells més expressats en un dels grups.

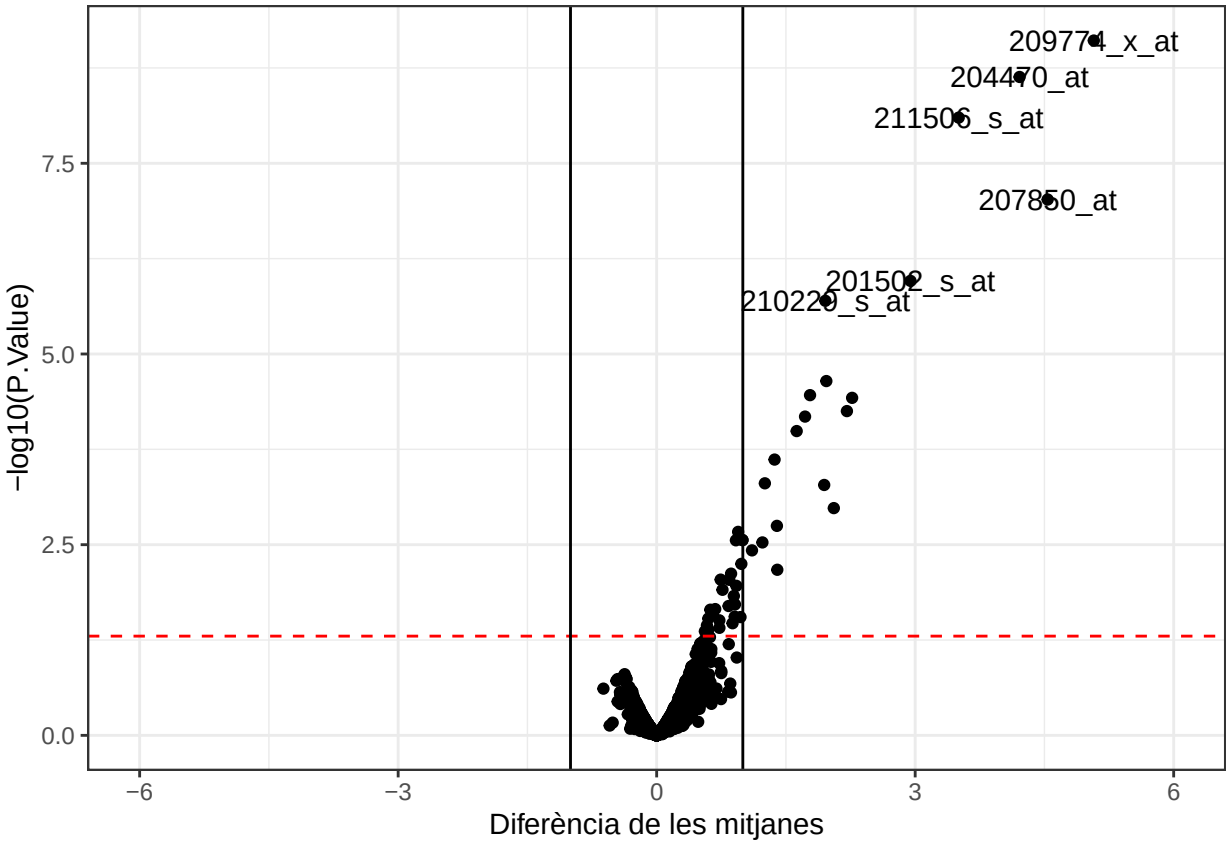

Figure 10: Volcano plot de les dades

4.7 Anotació dels resultats

A continuació afegim una columna a la taula generada que contingui el nom de l'identificador d'Affymetrix, que actualment correspon al nom de la fila, i la utilitzem per anotar la taula, de la qual seleccionem només els gens que considerem expressats diferencialment. S'utilitzen les dades del xip `hgu133a` (*Affymetrix Human Genome U133A Array*).

```
#Afegim la columna amb la PROBEID
taula_id<-cbind(PROBEID=rownames(taula),taula)
transcrits<-rownames(taula_id)
#Guardem les dades del paquet d'anotacions
paquet<-eval(parse(text="hgu133a.db"))
anotacions<-select(paquet,transcrits,c("SYMBOL",
                                       "ENTREZID",
                                       "GENENAME"))
taulaAnotada<-merge(x=anotacions,
                    y=taula_id,
                    by.x="PROBEID",
                    by.y="PROBEID")
seleccioAnotada<-taulaAnotada[taulaAnotada$PROBEID %in% rownames(res.selected),]
knitr::kable(head(seleccioAnotada[c(2,4,8)],10),
             caption="Gens seleccionats amb anotacions")
```

Table 4: Gens seleccionats amb anotacions

	SYMBOL	GENENAME	P.Value
71	MCL1	MCL1 apoptosis regulator, BCL2 family member	0.1130191
153	BCLAF1	BCL2 associated transcription factor 1	0.2321433
155	THBS1	thrombospondin 1	0.0361383
158	EIF5A	eukaryotic translation initiation factor 5A	0.1537334
164	EIF2S1	eukaryotic translation initiation factor 2 subunit alpha	0.1892972
175	AMD1	adenosylmethionine decarboxylase 1	0.0657511
209	ETS2	ETS proto-oncogene 2, transcription factor	0.0625341
241	JUN	Jun proto-oncogene, AP-1 transcription factor subunit	0.0148790
246	JUNB	JunB proto-oncogene, AP-1 transcription factor subunit	0.0281475
251	RCN2	reticulocalbin 2	0.2207257

Es mostra la taula anotada. De nou, es guarden els resultats en un fitxer .csv.

```
write.csv(taulaAnotada,file="./Resultats/taulaAnotada.csv")
write.csv(seleccioAnotada,file="./Resultats/seleccioAnotadaCTRLvsAA.csv")
```

4.8 Anàlisi de significació biològica

A continuació es realitza l'anàlisi de signifiació biològica per a veure els processos implicats. Es seleccionen els gens que s'han considerat diferencialment amb un criteri poc estricte de cara a augmentar la mida de la mostra.

```
probesInHeatMap<-rownames(res.selected)
HMdata<-exprs(gens_filtrats)[rownames(exprs(gens_filtrats)) %in% probesInHeatMap,]
geneSymbols<-select(hgu133a.db,rownames(HMdata),c("SYMBOL"))
SYMBOLS<-geneSymbols$SYMBOL
rownames(HMdata)<-SYMBOLS
knitr::kable(head(HMdata),
             caption="Expressió per mostra i gen",
             digits=3)
```

Table 5: Expressió per mostra i gen

	CTRL1	CTRL2	CTRL3	CTRL4	AA1	AA2	AA3	AA4
TSFM	4.239	4.923	4.390	4.541	4.594	4.694	5.209	5.395
FEM1B	6.038	5.703	6.115	5.708	6.278	6.396	6.331	6.463
RBM12	6.492	7.335	7.067	6.735	7.248	7.229	7.732	7.895
MBNL2	5.351	5.595	5.495	6.306	5.693	5.477	6.642	6.598
ADAM10	7.789	7.430	8.180	7.503	8.056	8.199	8.018	8.171
RXYLT1	5.007	4.241	5.235	4.660	5.488	5.094	5.467	5.255

```
write.csv(HMdata,file="./Resultats/data4Heatmap.csv")
```

Com es pot veure, s'ha creat una taula on cada fila és un gen diferencialment expressat i cada columna els seus valors d'expressió en cada mostra. Ara es realitza el mapa de color (Figura 11).

Es pot observar un perfil d'expressió diferenciat entre els controls i el grup tractat, tot i que també hi ha diferències aparents dins els grups.

Com a part de l'anàlisi de significació biològica, creem un informe HTML, consultable des de la carpeta `Resultats`, amb l'anàlisi associat a la base de dades GO.

```
library(GOstats);library(annotate)
# Seleccionem la llista de tots els gens de l'estudi.
entrezUniverse<-unique(getEG(taulaAnotada$PROBEID, "hgu133a.db"))
# Seleccionem els identificadors dels gens que hem escollit
quinsGens<-taulaAnotada["P.Value"]<0.25
genIDs<-unique(getEG(taulaAnotada$PROBEID[quinsGens],"hgu133a.db"))
# Creem els paràmetres:
GOparams<-new("GOHyperGParams",
              geneIds=genIDs, universeGeneIds=entrezUniverse,
              annotation="org.Hs.eg.db", ontology="BP",
              pvalueCutoff=0.001,
              conditional=FALSE,
              testDirection="over")
GOhyper<-hyperGTest(GOparams)
# Creem un informe amb els resultats.
htmlReport(GOhyper,file="./Resultats/Informe.html",summary.args=list("htmlLinks"=TRUE))
```

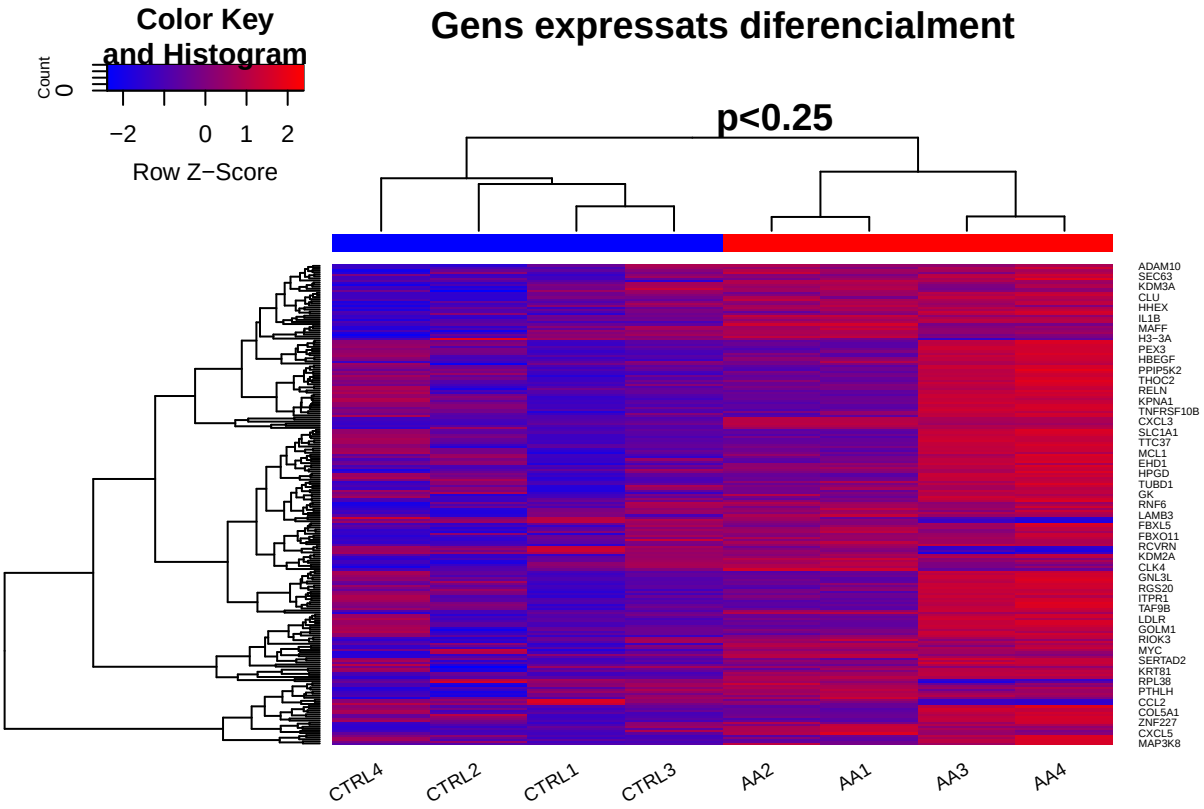

Figure 11: Mapa de color amb expressió diferencial entre mostres

5 Discussió

Malgrat que en l'anàlisi de qualitat inicial no sembla que s'haguessin detectat problemes importants amb les mostres, en analitzar les fonts de variabiltiat de les dades s'ha determinat que la principal font de variabilitat era la data de processament de la mostra. Aquesta troballa és altament suggestiva de l'existència d'un efecte mostra o *batch effect* molt marcat, que fa que les diferències que s'hagin pogut observar entre els grups puguin dependre meś de com s'ha processat la mostra que no de l'efecte del tractament que es volia estudiar. Per aquest motiu, qualsevol conclusió extreta d'aquest estudi s'hauria de tractar amb prudència, i fins i tot s'hauria de valorar repetir l'estudi intentant corregir aquest efecte per a poder arribar a cap conclusió.

En el present treball s'ha dut a terme únicament una comparació, en la què possiblement per algun motiu relacionat amb el processament de la mostra, s'ha objectivat expressió diferencial de relativament pocs gens i sorprenentment quasi en un únic sentit, tal i com es pot observar a la taula:

Table 6: Nombre de gens expressats diferencialment

Expressió	AAvsCTRL
Upregulated	218
Downregulated	13

Els gens seleccionats s'han desat al fitxer `seleccioAnotada.csv`, disponible a la carpeta de resultats. En funció de les entrades a Gene Ontology, s'ha generat un document HTML, el fitxer `informe.html` que permet veure les funcions i processos afectats per l'expressió diferencial de gens i per tant, si l'estudi no tingués les limitacions explicades, l'efecte del tractament sobre les cèl · lules neolpàsiques.

Cal tenir en compte que qualsevol dels gens seleccionats és un candidat possible, però la seva expressió s'hauria de verificar per tècniques d'amplificació, ja que l'anàlisi de *microarrays* només té com a objectiu descobrir possibles candidats.

6 Conclusió

No s'ha realitzat interpretació biològica dels resultats ja que l'objectiu era principalment l'anàlisi bioinformàtic. Totes les dades utilitzades inicialment, aquest informe i el fitxer Rmarkdown utilitzat per a generar-lo i les dades generades durant l'anàlisi bioinformàtic poden trobar-se, també, en aquest repositori de *Github*[7].

[7] https://github.com/padro89/omiques