Statistical Analysis of Microarray data.

Magda Pla Montferrer

April 15, 2020

Contents

El present informe i les seves dades es pot consultar a https://github.com/MagdaPla/ADO-PEC1.git

1. Abstract

En aquest document es proposa una metodologia amb el codi en R complet per a realitzar un procés d'anàlisis de microarrays (tipus Affimetrix) a partir de les dades brutes descarregables des del repositori GEO. Es realitza en mode d'exemple l'anàlisis de les dades de l'estudi GSE131667, que analitza l'expressió genètica de la glicoproteïna transmembrana CD248 i el seu impacte en el metabolisme del teixít adipós blanc (WAT) i malalties metabòliques. El document també conté informació complementària per si es desitga ampliar els anàlisis amb un altre grup de dades òmiques.

2. Objectius

L'objectiu principal del document és tenir una proposta d'anàlisis a realitzar amb el codi corresponent, correctament documentat d'informació bruta del repositori GEO en format `.CEL` utilitzant Bioconductor amb R i altres eines pròpies d'anàlisis de informació genètica.

3. Material i mètodes

3.1.Naturalesa de les dades

Les dades utilitzades s'han descarregat del repositori web de lliure accés GEO (Gene Expression Omnibus a https://www.ncbi.nlm.nih.gov/geo/browse/?view=series&zsort=date&display=100&page=4), concretament s'ha descarregat el treball amb codi GSE131667 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131667), publicat per Petrus et al. (2019). És conegut que un balanç energètic positiu afavoreix l'expansió del teixit adipós blanc (WAT) que es caracteritza per l'activació d'un repertori d'esdeveniments que inclouen hipòxia, inflamació i remodelació de la matriu extracel · lular. La glicoproteïna transmembrana CD248 està implicada en tots aquests processos en diferents malalties malignes i inflamatòries, però no s'ha explorat el seu impacte potencial en el WAT i en les malalties metabòliques. En el treball publicat a Petrus et al. (2019) s'evalua el paper de CD248 en la funció d'adipòcits i el metabolisme de la glucosa mitjançant anàlisis d'omics en WAT humà, knockdowns de gens en adipòcits diferenciats in vitro humans. A nivell experimental, les cèl · lules mare derivades del teixit adipós humà es van diferenciar per adipòcits in vitro. Al final de la diferenciació, es van tractar les cèl · lules amb siRNA dirigit al CD248 seguit d'exposició a nivells d'oxigen de l'1%. Es va realitzar una anàlisi de microarrays per identificar gens regulats de manera diferent. L'experiment concret amb aquestes dades proporciona informació sobre els gens expressats diferencialment per a la disminució de CD248 i com la resposta a la hipòxia està influenciada per un nivell reduït de CD248. Al final de la diferenciació, es van fixar les cèl · lules per obtenir ARN.

S'ha utilitzat el microarrays tipus Affymetrix [Clariom_S_Human] i es comparen un grups control amb siCD248 en condicions normals d'oxigen i un altre grup control amb siCD248 en condicions d'hipòxia. De cada grup s'han realitzat 3 rèpliques.

3.2 Mètodes utilitzats en l'anàlisis

Per a començar es preparen els directoris on s'emmagatzemaran les dades brutes descarregades de la web, directoris on es desaran els resultats i figures necessàries per a realitzar l'informe final. També es realitzarà la instal · lació de les llibreries necessàries. Una vegada la part més tècnica estigui a punt, es procedeix a realitzar l'anàlisis de microarrays seguint la metodologia proposada a Gonzalo Sanz and Sánchez-Pla (2019), la qual segueix els següents passos:

1. Identificar els grups y quins grups pertanyen a cada mostra.
2. Control de qualitat de les dades crues
3. Normalització
4. Control de qualitat de les dades normalitzades
5. Filtratge no específic
6. Identificació de genes diferencialment expresats
7. Anotació dels resultats
8. Comparació entre diferents comparacions
9. Anàlisis de significació biològica ("Gene Enrichment Analysis")

3.3.Procediment

Abans de començar pròpiament l'anàlisis es preparen els directoris on emmagatzemar les dades i els resultats. És molt important estructurar la informació, desar-la en directoris diferents segons la seva funció. Tenir la informació correctament endreçada facilitat molt tot el procés, treballar cooperativament amb altres investigadors, etc. Aquest punt es pot fer des de R o bé directament creant els directoris des d'un explorador de windows. Per a fer-ho des de R:

```
> # aquest codi només el fem la primera vegada per crear directoris
> setwd(".")
>
> # directori on desar les dades de partida: dades brutes en foramt .CEL i el fitxer "targets.csv"
> dir.create("data")
```

```
>
> # directori on es dirigiran automàticament els resultats
> dir.create("results")
>
> # també es crea el directori "figures" on es desen figures necessàries per a documentar l'informe
> # igualment pot ser molt necessari crear aquest directori per a guardar figures amb el mateix propòsi
> # per a realitzar els informes d'altres dades.
> dir.create("figures")
```

Convé tenir instal · lades les següents llibreries. Com que algunes llibreries són grans i el procés d'instal · lació és llarg es recomana instal · lar-les només si és necessari (quan no es tenen o convé actualitzar-les)

```
> # per una banda convé instalat Bioconductor:
>
> #if (!requireNamespace("BiocManager", quietly = TRUE))
> #     install.packages("BiocManager")
> #BiocManager::install()
>
> #Juntament amb altres paquets específics que es van carregant al llarg del procés i
> #alhora, algun paquet d'anotacions també de Bioconductor i altra informació complementària.
```

3.3.1. Identificar els grups y quins grups pertanyen a cada mostra

A continuació hem preparat el fitxer "target.csv" a partir de la informaicó descrita a l'experiment (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131667). El fitxer csv és de la setÜent manera:

```
> # veiem primer la taula que hem construit, anomenant-la "targets"
> targets <- read.csv2("./data/targets.csv", header = TRUE, sep = ";")
> # té la següent forma
> knitr::kable(
+    targets, booktabs = TRUE,
+    caption = 'Content of the targets file used for the current analysis')
```

Table 1: Content of the targets file used for the current analysis

FileName	Group	STATE	GEN	ShortName
GSM3801054_PP01_175_Clariom_S_Human.CEL	CTRL.NORM	NORM	CTRL	CTRL.NORM.175
GSM3801055_PP02_176_Clariom_S_Human.CEL	CTRL.NORM	NORM	CTRL	CTRL.NORM.176
GSM3801056_PP03_177_Clariom_S_Human.CEL	CTRL.NORM	NORM	CTRL	CTRL.NORM.177
GSM3801057_PP04_187_Clariom_S_Human.CEL	SICD248.NORM	NORM	SICD248	SICD248.NORM.187
GSM3801058_PP05_188_Clariom_S_Human.CEL	SICD248.NORM	NORM	SICD248	SICD248.NORM.188
GSM3801059_PP06_189_Clariom_S_Human.CEL	SICD248.NORM	NORM	SICD248	SICD248.NORM.189
GSM3801060_PP07_202_Clariom_S_Human.CEL	CTRL.HYPO	HYPO	CTRL	CTRL.HYPO.202
GSM3801061_PP08_203_Clariom_S_Human.CEL	CTRL.HYPO	HYPO	CTRL	CTRL.HYPO.203
GSM3801062_PP09_204_Clariom_S_Human.CEL	CTRL.HYPO	HYPO	CTRL	CTRL.HYPO.204
GSM3801063_PP10_208_Clariom_S_Human.CEL	SICD248.HYPO	HYPO	SICD248	SICD248.HYPO.208
GSM3801064_PP11_209_Clariom_S_Human.CEL	SICD248.HYPO	HYPO	SICD248	SICD248.HYPO.209
GSM3801065_PP12_210_Clariom_S_Human.CEL	SICD248.HYPO	HYPO	SICD248	SICD248.HYPO.210

A continaució es procedeix a llegir propiament els fitxers .CEL (dades brutes o raw data). Es crearà la variable "rawData:

```
> # ara llegim la mateixa taula creant una nova variable "my.targets"
> # ho fem amb `read.AnnotatedDataFrame()` així ho associem amb els fitxers .CEL
```

```
> library(oligo)   # carregar aquesta llibreria tarda una miqueta
> celFiles <- list.celfiles("./data", full.names = TRUE)
> library(Biobase)
> my.targets <-read.AnnotatedDataFrame(file.path("./data","targets.csv"),
+                                      header = TRUE,
+                                      row.names = 1,
+                                      sep=";")
> rawData <- read.celfiles(celFiles, phenoData = my.targets)
```

```
Reading in : ./data/GSM3801054_PP01_175_Clariom_S_Human.CEL
Reading in : ./data/GSM3801055_PP02_176_Clariom_S_Human.CEL
Reading in : ./data/GSM3801056_PP03_177_Clariom_S_Human.CEL
Reading in : ./data/GSM3801057_PP04_187_Clariom_S_Human.CEL
Reading in : ./data/GSM3801058_PP05_188_Clariom_S_Human.CEL
Reading in : ./data/GSM3801059_PP06_189_Clariom_S_Human.CEL
Reading in : ./data/GSM3801060_PP07_202_Clariom_S_Human.CEL
Reading in : ./data/GSM3801061_PP08_203_Clariom_S_Human.CEL
Reading in : ./data/GSM3801062_PP09_204_Clariom_S_Human.CEL
Reading in : ./data/GSM3801063_PP10_208_Clariom_S_Human.CEL
Reading in : ./data/GSM3801064_PP11_209_Clariom_S_Human.CEL
Reading in : ./data/GSM3801065_PP12_210_Clariom_S_Human.CEL
```

```
> # amb aquesta funció s'ha baixat el paquet d'anotacions requerit ("pl.clarioms.human")
> print(pData(rawData)) #una manera de veure el fitxer targets.csv amb la informació bruta
```

```
                                          Group STATE     GEN
GSM3801054_PP01_175_Clariom_S_Human.CEL    CTRL.NORM  NORM    CTRL
GSM3801055_PP02_176_Clariom_S_Human.CEL    CTRL.NORM  NORM    CTRL
GSM3801056_PP03_177_Clariom_S_Human.CEL    CTRL.NORM  NORM    CTRL
GSM3801057_PP04_187_Clariom_S_Human.CEL SICD248.NORM  NORM SICD248
GSM3801058_PP05_188_Clariom_S_Human.CEL SICD248.NORM  NORM SICD248
GSM3801059_PP06_189_Clariom_S_Human.CEL SICD248.NORM  NORM SICD248
GSM3801060_PP07_202_Clariom_S_Human.CEL    CTRL.HYPO  HYPO    CTRL
GSM3801061_PP08_203_Clariom_S_Human.CEL    CTRL.HYPO  HYPO    CTRL
GSM3801062_PP09_204_Clariom_S_Human.CEL    CTRL.HYPO  HYPO    CTRL
GSM3801063_PP10_208_Clariom_S_Human.CEL SICD248.HYPO  HYPO SICD248
GSM3801064_PP11_209_Clariom_S_Human.CEL SICD248.HYPO  HYPO SICD248
GSM3801065_PP12_210_Clariom_S_Human.CEL SICD248.HYPO  HYPO SICD248
                                          ShortName
GSM3801054_PP01_175_Clariom_S_Human.CEL    CTRL.NORM.175
GSM3801055_PP02_176_Clariom_S_Human.CEL    CTRL.NORM.176
GSM3801056_PP03_177_Clariom_S_Human.CEL    CTRL.NORM.177
GSM3801057_PP04_187_Clariom_S_Human.CEL SICD248.NORM.187
GSM3801058_PP05_188_Clariom_S_Human.CEL SICD248.NORM.188
GSM3801059_PP06_189_Clariom_S_Human.CEL SICD248.NORM.189
GSM3801060_PP07_202_Clariom_S_Human.CEL    CTRL.HYPO.202
GSM3801061_PP08_203_Clariom_S_Human.CEL    CTRL.HYPO.203
GSM3801062_PP09_204_Clariom_S_Human.CEL    CTRL.HYPO.204
GSM3801063_PP10_208_Clariom_S_Human.CEL SICD248.HYPO.208
GSM3801064_PP11_209_Clariom_S_Human.CEL SICD248.HYPO.209
GSM3801065_PP12_210_Clariom_S_Human.CEL SICD248.HYPO.210
```

```
> my.targets@data$ShortName->rownames(pData(rawData))
> colnames(rawData) <-rownames(pData(rawData))
>
```

```
> # veiem les metadades de la taula
> head(rawData)
```

```
ExpressionFeatureSet (storageMode: lockedEnvironment)
assayData: 1 features, 12 samples
  element names: exprs
protocolData
  rowNames: CTRL.NORM.175 CTRL.NORM.176 ... SICD248.HYPO.210 (12 total)
  varLabels: exprs dates
  varMetadata: labelDescription channel
phenoData
  rowNames: CTRL.NORM.175 CTRL.NORM.176 ... SICD248.HYPO.210 (12 total)
  varLabels: Group STATE GEN ShortName
  varMetadata: labelDescription channel
featureData: none
experimentData: use 'experimentData(object)'
Annotation: pd.clariom.s.human
```

```
> class(rawData)
```

```
[1] "ExpressionFeatureSet"
attr(,"package")
[1] "oligoClasses"
```

Si mirem la classe de les dades brutes s'observa que són `expressionFeatureSet` enlloc de l'esperat `expressionSet`.

3.3.2. Control de qualitat de les dades crues

Es fa un control de la qualitat de les dades. La llibreria utilitzada `arrayQualityMetrics()` realitza diferents testos de qualitat. Aquest procés és una miqueta lent anivell de processament, especialment si el nombre de mostres és elevat.

```
> library(arrayQualityMetrics)
> #a banda d'executar la funció de control de qualitat
> # també se li indica que creei un directori dins de la carpeta de resultats:"rawData_quality"
> # en aquest directori es desarà tota la informació referent al control de qualitat
> arrayQualityMetrics(rawData, outdir = "./results/rawData_quality", force = T)
```

El resultat de tot l'anàlisis de qualitat es resum en el fitxer "index.html" emmagatzemat dins de la carpeta de "results". En la taula següent, les columnes 1,2 i 3 mostren tres criteris diferents per a mesurar la qualitat de les dades. En els casos que es marca una creu convé que es tingui un especial interès en aquesta informació. En l'exemple següent 1 mostra s'ha marcat en dues columnes i una altra mostra en una sola columna, això pot indicar que els problemes potencials de les dades poden ser pocs, per tant, es podrien mantenir totes les dades.

NOTA: la figura següent s'ha generat a partir d'una captura de pantalla del fitxer "index.htm" i s'ha desat com a .png en el directori de figures

A més, es poden aplicar altres anàlisis de qualitat: **A) Anàlisis de components principals.** A partir d'una funció definida per Gonzalo Sanz and Sánchez-Pla (2019) es crea una gràfica amb les dues components principals:

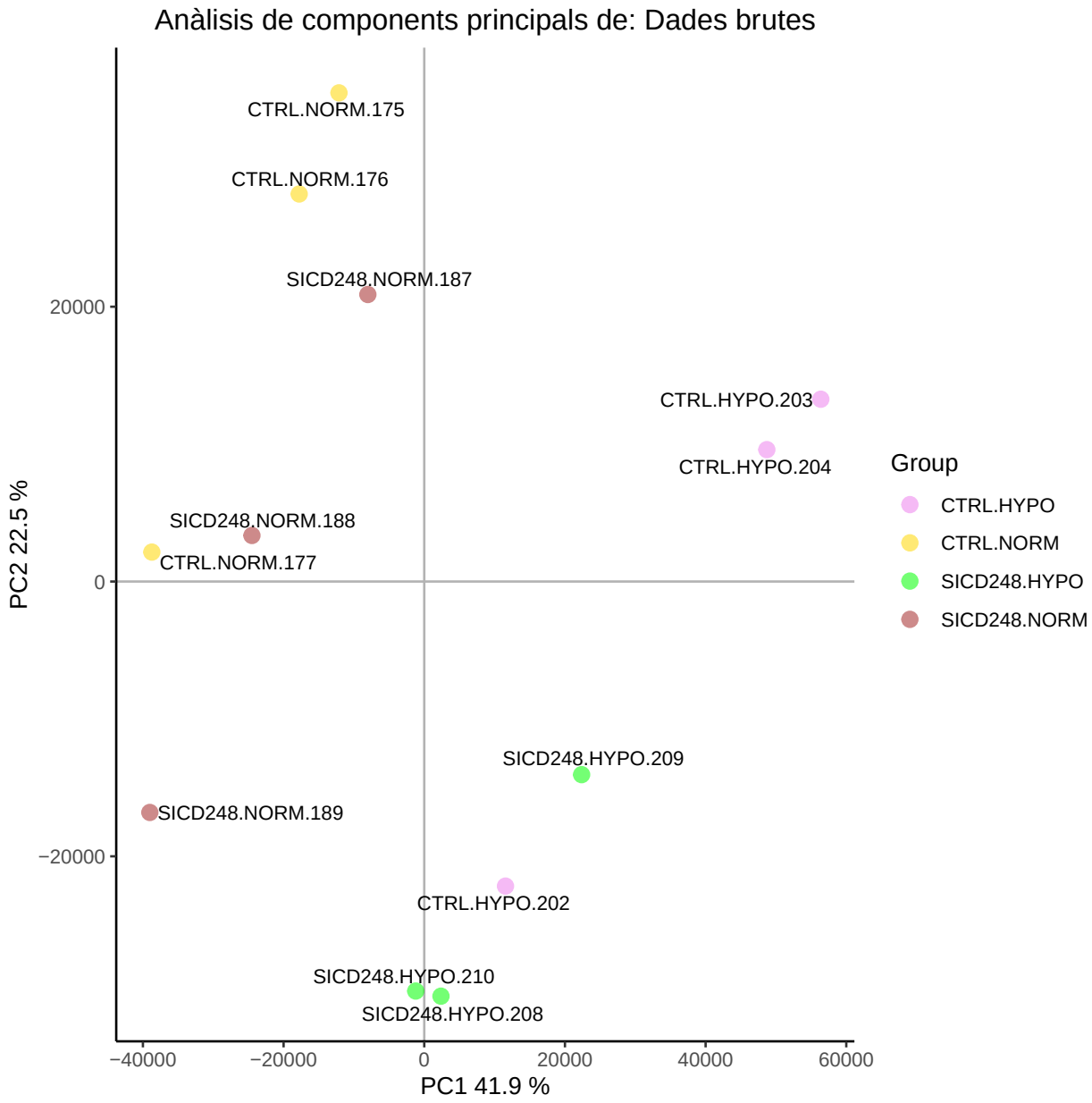
```
> library(ggplot2)
> library(ggrepel)
> # ´la següent funció prepara i edita una PCA
> plotPCA3 <- function (datos, labels, factor, title, scale,colors, size = 1.5, glineas = 0.25) {
+   data <- prcomp(t(datos),scale=scale)
```

	array	sampleNames	*1	*2	*3	Group	STATE	GEN	ShortName
☐	1	CTRL.NORM.175				CTRL.NORM	NORM	CTRL	CTRL.NORM.175
☐	2	CTRL.NORM.176				CTRL.NORM	NORM	CTRL	CTRL.NORM.176
☐	3	CTRL.NORM.177				CTRL.NORM	NORM	CTRL	CTRL.NORM.177
☐	4	SICD248.NORM.187				SICD248.NORM	NORM	SICD248	SICD248.NORM.187
☐	5	SICD248.NORM.188				SICD248.NORM	NORM	SICD248	SICD248.NORM.188
☐	6	SICD248.NORM.189				SICD248.NORM	NORM	SICD248	SICD248.NORM.189
☐	7	CTRL.HYPO.202				CTRL.HYPO	HYPO	CTRL	CTRL.HYPO.202
☐	8	CTRL.HYPO.203	x		x	CTRL.HYPO	HYPO	CTRL	CTRL.HYPO.203
☐	9	CTRL.HYPO.204			x	CTRL.HYPO	HYPO	CTRL	CTRL.HYPO.204
☐	10	SICD248.HYPO.208				SICD248.HYPO	HYPO	SICD248	SICD248.HYPO.208
☐	11	SICD248.HYPO.209				SICD248.HYPO	HYPO	SICD248	SICD248.HYPO.209
☐	12	SICD248.HYPO.210				SICD248.HYPO	HYPO	SICD248	SICD248.HYPO.210

Figure 1: Figura 1. Taula resum de la qualitat de les dades, extret del fitxer index.html produït per la arrayQualityMetrics() sobre les dades brutes

```
+   # plot adjustments
+   dataDf <- data.frame(data$x)
+   Group <- factor
+   loads <- round(data$sdev^2/sum(data$sdev^2)*100,1)
+   # main plot
+   p1 <- ggplot(dataDf,aes(x=PC1, y=PC2)) +
+     theme_classic() +
+     geom_hline(yintercept = 0, color = "gray70") +
+     geom_vline(xintercept = 0, color = "gray70") +
+     geom_point(aes(color = Group), alpha = 0.55, size = 3) +
+     coord_cartesian(xlim = c(min(data$x[,1])-5,max(data$x[,1])+5)) +
+     scale_fill_discrete(name = "Group")
+   # avoiding labels superposition
+   p1 + geom_text_repel(aes(y = PC2 + 0.25, label = labels),segment.size = 0.25, size = size) +
+     labs(x = c(paste("PC1",loads[1],"%")),y=c(paste("PC2",loads[2],"%"))) +
+     ggtitle(paste("Anàlisis de components principals de:",title,sep=" "))+
+     theme(plot.title = element_text(hjust = 0.5)) +
+     scale_color_manual(values=colors)
+ }
```

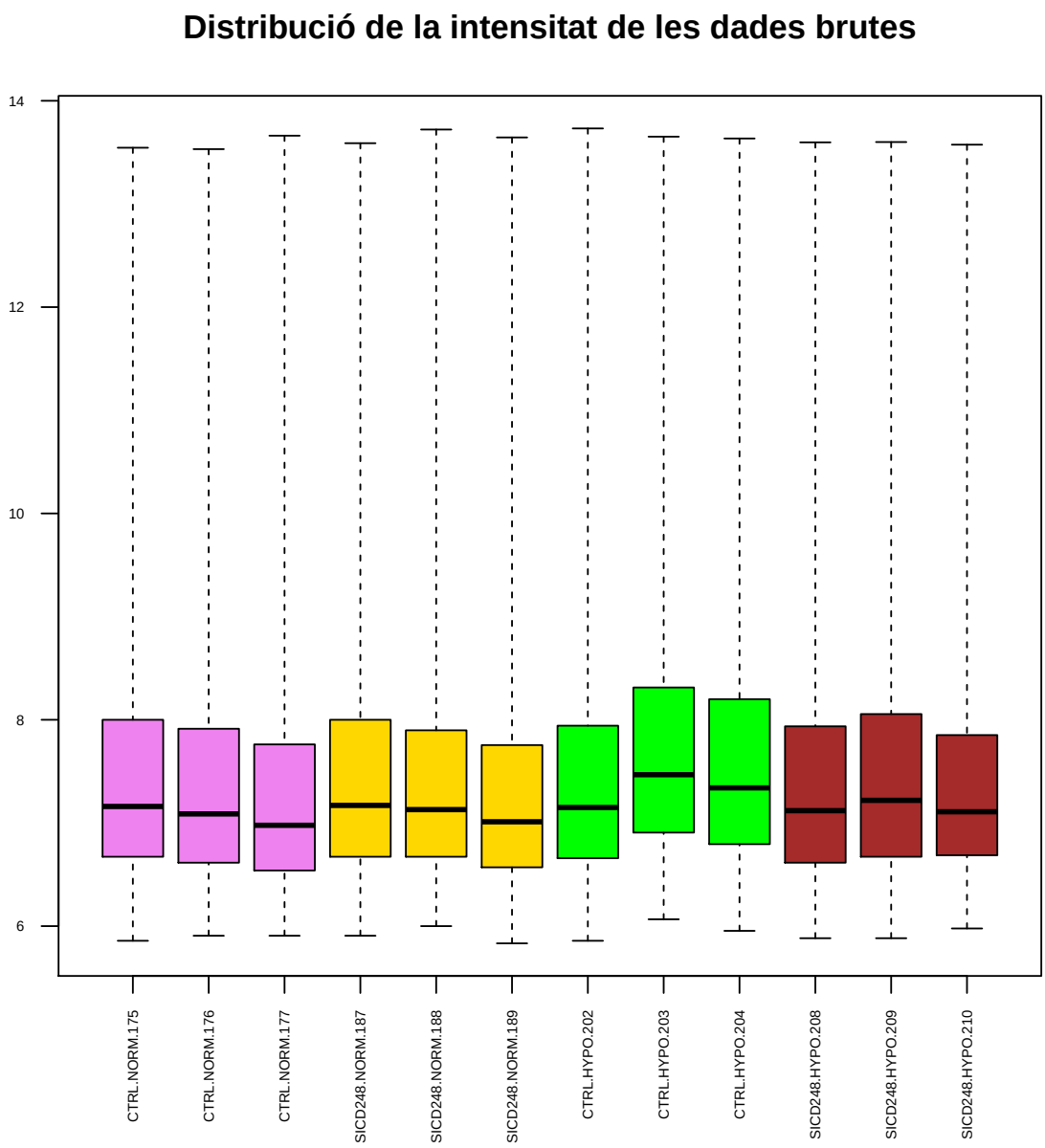
```
> plotPCA3(exprs(rawData), labels = targets$ShortName, factor = targets$Group,
+          title = "Dades brutes",scale = FALSE, size = 3,
+          colors = c("violet", "gold","green","brown"))
```



El gràfic mostra com les dues primers components principals expliquen un 64% de la variabilitat però no es veuen patrons molt clars que difenciïn els diferents grups de dades. Hi ha una tendència a que els grups amb condicions normals d'oxigen estan a la meitat esquerra de l'eix de les X (Component Principal 1) i els grups amb hipòxia a la banda dreta, així com també s'observa una certa tendència a que els grups control tenen valors més elevats de la Component principal 2. Però hi ha grups que no segueixen clarament aquest patró.

B) Distribució de les intensitats També es pot realitzar una figura de distribució de les intensitats:

```
> boxplot(rawData, cex.axis=0.5, las=2,
+         which="all",
+         col = c(rep("violet", 3), rep("gold", 3),rep("green", 3),rep("brown", 3)),
+         main="Distribució de la intensitat de les dades brutes",
+         ylim = c(0,20))
```



La figura no mostra grans discrepàncies entre totes les mostres, tal i com s'esperaria en les dades brutes.

3.3.3.Normalització

Per tal de disminuïr les diferències que hi puguin haver entre les mostres degut a les tècniques emprades i no a diferències biològiques, és necessari normalitzar les dades, així es poden comparar i realitzar realment una comparació diferencial de l'expressió genètica. El mètode de normalització utilitzat és un dels més comuns: *Robust Multichip Analysis* i es descriu a Irizarry et al. (2003)

```
> eset_rma <- oligo::rma(rawData)

Background correcting
Normalizing
Calculating Expression
```

```
> eset_rma
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 27189 features, 12 samples
  element names: exprs
protocolData
  rowNames: CTRL.NORM.175 CTRL.NORM.176 ... SICD248.HYPO.210 (12 total)
  varLabels: exprs dates
  varMetadata: labelDescription channel
phenoData
  rowNames: CTRL.NORM.175 CTRL.NORM.176 ... SICD248.HYPO.210 (12 total)
  varLabels: Group STATE GEN ShortName
  varMetadata: labelDescription channel
featureData: none
experimentData: use 'experimentData(object)'
Annotation: pd.clariom.s.human
```

```
> class(eset_rma)
```

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

```
> #Es desen les dades normalitzades en un fitxer .csv
> write.csv2(exprs(eset_rma), file="./results/normalized.Data.csv")
```

Al fer la normalització veiem que el tipus de dades ja són un `ExpreessionSet`.

A continuació s'incorporen les anotacions perquè es necessitaran per les funcions de més endavant. En aquest cas les anotacions estan en el fitxer *pd.clariom.s.human* i la millor manera per a afegir les anotacions en aquest cass és amb el paquet *"affycoretools"*. Sovint hi ha diferents maneres de fer aquest procediment, cercant el nom del paquet d'anotacions i com afegir-les al fitxer de dades es pot buscar a la web de Biococnductor o altres fonts, la millor manera de realitzar el procés:

```
> #BiocManager::install("affycoretools")
> library(affycoretools)
> all.eset <- annotateEset(eset_rma, annotation(eset_rma))
> all.eset
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 27189 features, 12 samples
  element names: exprs
protocolData
  rowNames: CTRL.NORM.175 CTRL.NORM.176 ... SICD248.HYPO.210 (12 total)
  varLabels: exprs dates
  varMetadata: labelDescription channel
phenoData
  rowNames: CTRL.NORM.175 CTRL.NORM.176 ... SICD248.HYPO.210 (12 total)
  varLabels: Group STATE GEN ShortName
  varMetadata: labelDescription channel
featureData
  featureNames: 23064070 23064071 ... TSUnmapped00000823.hg.1 (27189
    total)
  fvarLabels: PROBEID ID SYMBOL GENENAME
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation: pd.clariom.s.human
```

```
> # es pot dessar també la inforamció normalitzada amb les anotacions
> write.csv(exprs(all.eset), file="./results/normalized_annotated.Data.csv")
```

Veiem que en les anotacions s'ha incorporat informació dels següents camps: *"PROBEID"*, *"ID"*, *"SYMBOL"* i *"GENENAME"*, tal i com es descriu `fvarLabels` al veure "all.eset". Això és rellevant alhora d'escriure el codi per a les següents funcions doncs molt codi està preparat per paquets d'anotacions que contenen columnes etiquetades com "ENTREZID", etc.

3.3.4. Control de qualitat de les dades normalitzades

Tornem a realitzar els processos de control de qualitat descrits en el punt 2 però sobre les dades normalitzades, que hem anomenat "eset_rma"

```
> arrayQualityMetrics(eset_rma, outdir = file.path("./results", "QCDir.Norm"), force=TRUE)
```

La taula resum del fitxer "index.html":

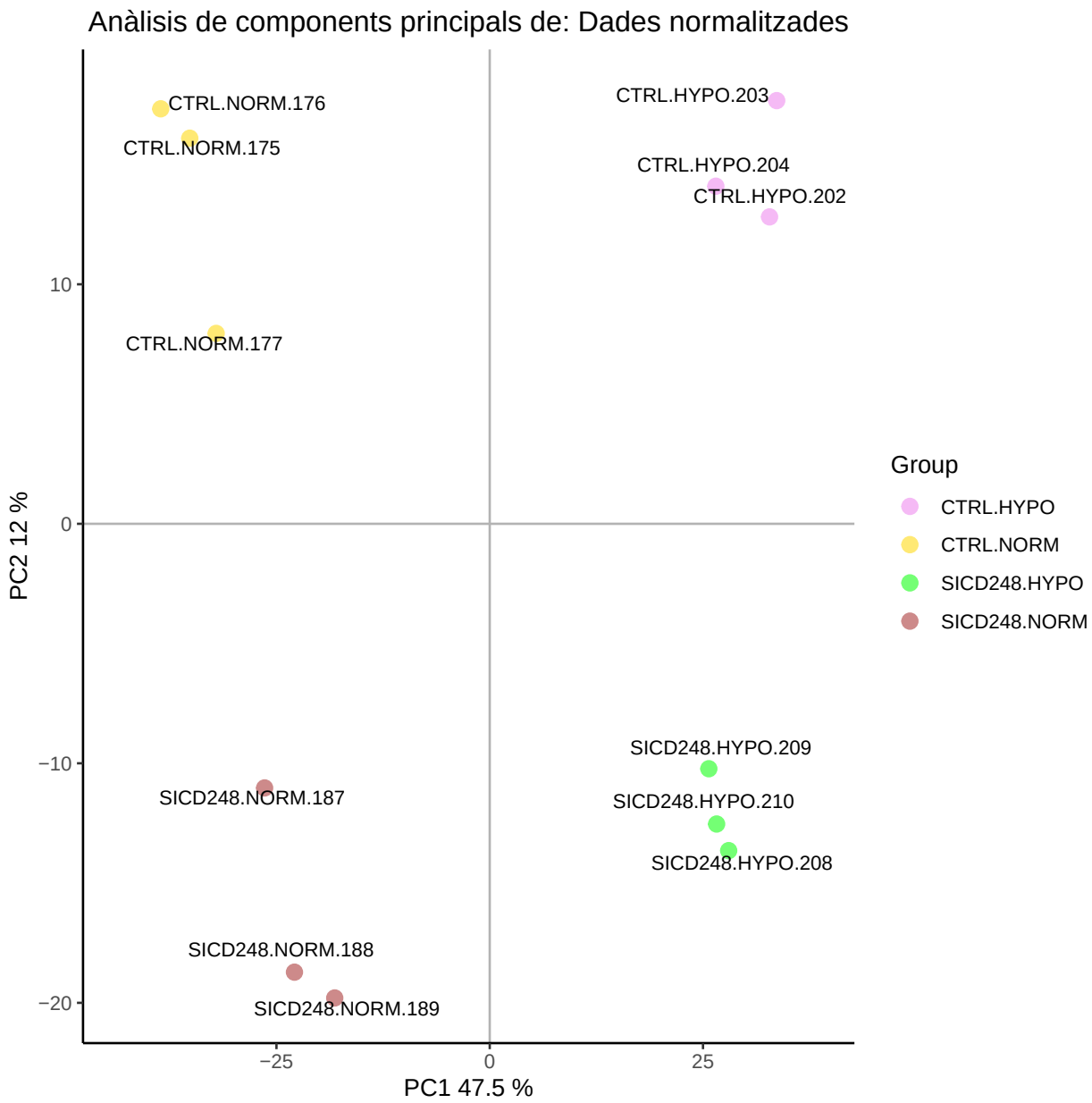
	array	sampleNames	*1	*2	*3	Group	STATE	GEN	ShortName
☐	1	CTRL.NORM.175				CTRL.NORM	NORM	CTRL	CTRL.NORM.175
☐	2	CTRL.NORM.176				CTRL.NORM	NORM	CTRL	CTRL.NORM.176
☐	3	CTRL.NORM.177				CTRL.NORM	NORM	CTRL	CTRL.NORM.177
☐	4	SICD248.NORM.187				SICD248.NORM	NORM	SICD248	SICD248.NORM.187
☐	5	SICD248.NORM.188				SICD248.NORM	NORM	SICD248	SICD248.NORM.188
☐	6	SICD248.NORM.189				SICD248.NORM	NORM	SICD248	SICD248.NORM.189
☐	7	CTRL.HYPO.202				CTRL.HYPO	HYPO	CTRL	CTRL.HYPO.202
☐	8	CTRL.HYPO.203				CTRL.HYPO	HYPO	CTRL	CTRL.HYPO.203
☐	9	CTRL.HYPO.204				CTRL.HYPO	HYPO	CTRL	CTRL.HYPO.204
☐	10	SICD248.HYPO.208				SICD248.HYPO	HYPO	SICD248	SICD248.HYPO.208
☐	11	SICD248.HYPO.209				SICD248.HYPO	HYPO	SICD248	SICD248.HYPO.209
☐	12	SICD248.HYPO.210				SICD248.HYPO	HYPO	SICD248	SICD248.HYPO.210

Figure 2: Figura 4. Taula resum de la qualitat de les dades, extret del fitxer index.html produït per la arrayQualityMetrics() sobre les dades normalitzades

Els resutlats d'aquest anàlisis de qualitat semblen indicar que la qualitat de les dades hagi augmentat. Ara només destaca dues mostres només en un dels anàlisis com de qualitat menor.

Els anàlisis de components principals:

```
> plotPCA3(exprs(eset_rma), labels = targets$ShortName, factor = targets$Group,
+          title = "Dades normalitzades",scale = FALSE, size = 3,
+          colors = c("violet", "gold","green","brown"))
```



La gràfica amb les dades normallitzades mostra com les dues primeres components principals expliquen gairebé un 60% de la variació, a més, a diferència de les dades sense normalitzar es diferencien 4 grups ben separats corresponents als 4 grups de dades.

La distribució de les dades normalitzades s'ha uniformitzat en comparació a les dades sense normalitzar, tal i ocm es veu a la figura següent:

```
> boxplot(eset_rma, cex.axis=0.5, las=2,
+         which="all",
+         col = c(rep("violet", 3), rep("gold", 3),rep("green", 3),rep("brown", 3)),
+         main="Distribució de la intensitat de les dades normalitzades",
+         ylim = c(0,15))
```

La expressió gènica dels microarrays pot se deguda a diferents fonts d'error, una podria ser la data amb que s'ha processat. Per saber si la data del processament realitzem:

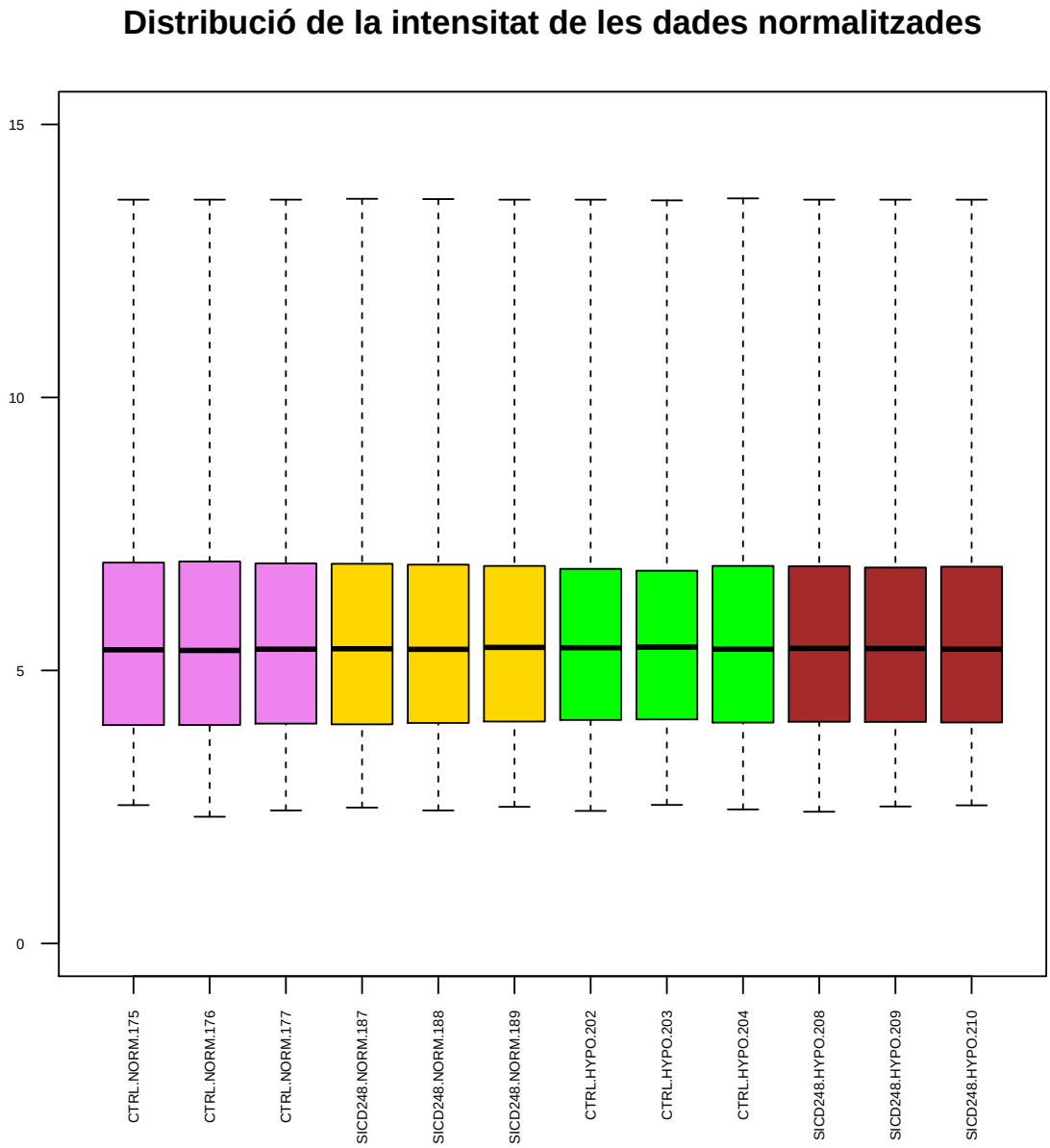


Figure 3: Figura 6.Boxplot de les dades normalitzades

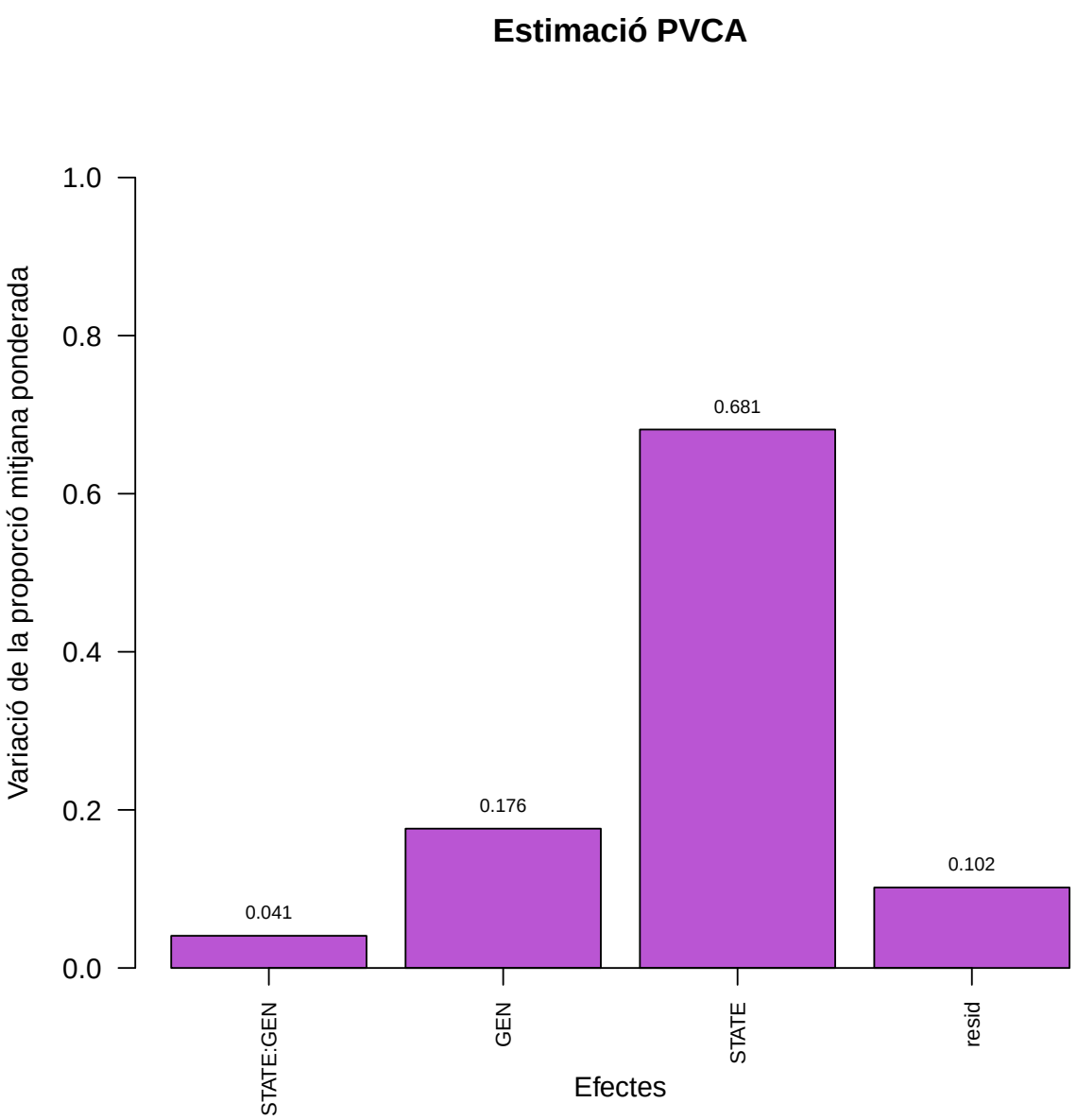
```
> library(affyio)
> get.celfile.dates(celFiles)
```

```
 [1] "2018-08-21" "2018-08-21" "2018-08-21" "2018-08-21" "2018-08-21"
 [6] "2018-08-21" "2018-08-21" "2018-08-21" "2018-08-21" "2018-08-21"
[11] "2018-08-21" "2018-08-21"
```

En aquest cas totes les mostres s'ha processat en la mateixa data: 21/08/2018. Per tant la data no pot ser una font de variació.

Així que es mirarà altres possibles fonts de variació. Es realitza un histograma on cada barra representarà i quantifica la font de variació inclosa en l'anàlisis. En aquest cas la proporció d'oxigen definida en el camp "STATE" i si es tracta de una mostra control o bé la glicoproteina CD248 tractada, descrita en el camp "GEN":

```
> #BiocManager::install("pvca")
> library(pvca)
>
> pData(eset_rma) <- targets
> #select the threshold
> pct_threshold <- 0.6
> #select the factors to analyze
> batch.factors <- c("STATE", "GEN")
> #run the analysis
> pvcaObj <- pvcaBatchAssess(eset_rma, batch.factors, pct_threshold)
>
>
> #plot the results
> bp <- barplot(pvcaObj$dat, xlab = "Efectes",
+  ylab = "Variació de la proporció mitjana ponderada",
+  ylim= c(0,1.1),col = c("mediumorchid"), las=2,
+  main="Estimació PVCA")
> axis(1, at = bp, labels = pvcaObj$label, cex.axis = 0.75, las=2)
> values = pvcaObj$dat
> new_values = round(values , 3)
> text(bp,pvcaObj$dat,labels = new_values, pos=3, cex = 0.7)
```



Els resultats mostren com que la principal variació és deguda a la quantitat d'oxigen a la que es sotmeten les mostres. Aquest és un factor experimental incorporat en l'anàlisis, per tant, és esperable que suposi una variació important. Així doncs, queden descartades fonts de variació degudes a aspectes metolològics de l'experiment propiament.

Per altra banda, si un gen s'expressa diferencialment, s'espera que hi hagi una certa diferència entre els grups i, per tant, la variació global del gen serà més gran que la dels que no tenen expressió diferencial. Dibuixar la variabilitat general de tots els gens és útil per decidir quin percentatge de gens mostra una variabilitat que es pot atribuir a altres causes que no pas la variació aleatòria.

Es realitza un anàlisis de la desviació estàndard i es considera que els gens més variables són aquells que tenen una desviació estàndard per sobre del 90-95%.

Es calcula la desviació estàndard per a cada gen i s'ordenen els resultats. Aquests càlculs es realitza sobre els gens filtrats (eset_filtered)

```
> sds <- apply (exprs(eset_rma), 1, sd)
> sds0<- sort(sds)
```

Es representen els resultats en una gràfica, ordenant de menys a més les desviacions estàndard de tots els gens. La línia vertical representa els gens més variables, amb una desviació estàndard per sobre del 90-95% de totes les desviacions estàndard.

```
> plot(1:length(sds0), sds0, sub="Les línies verticals representa els percentils 90% i 95% ",
+      xlab="Índex de gens (de menys a més variabilitat)", ylab="Desviació estàndard")
> abline(v=length(sds)*c(0.9,0.95))
```

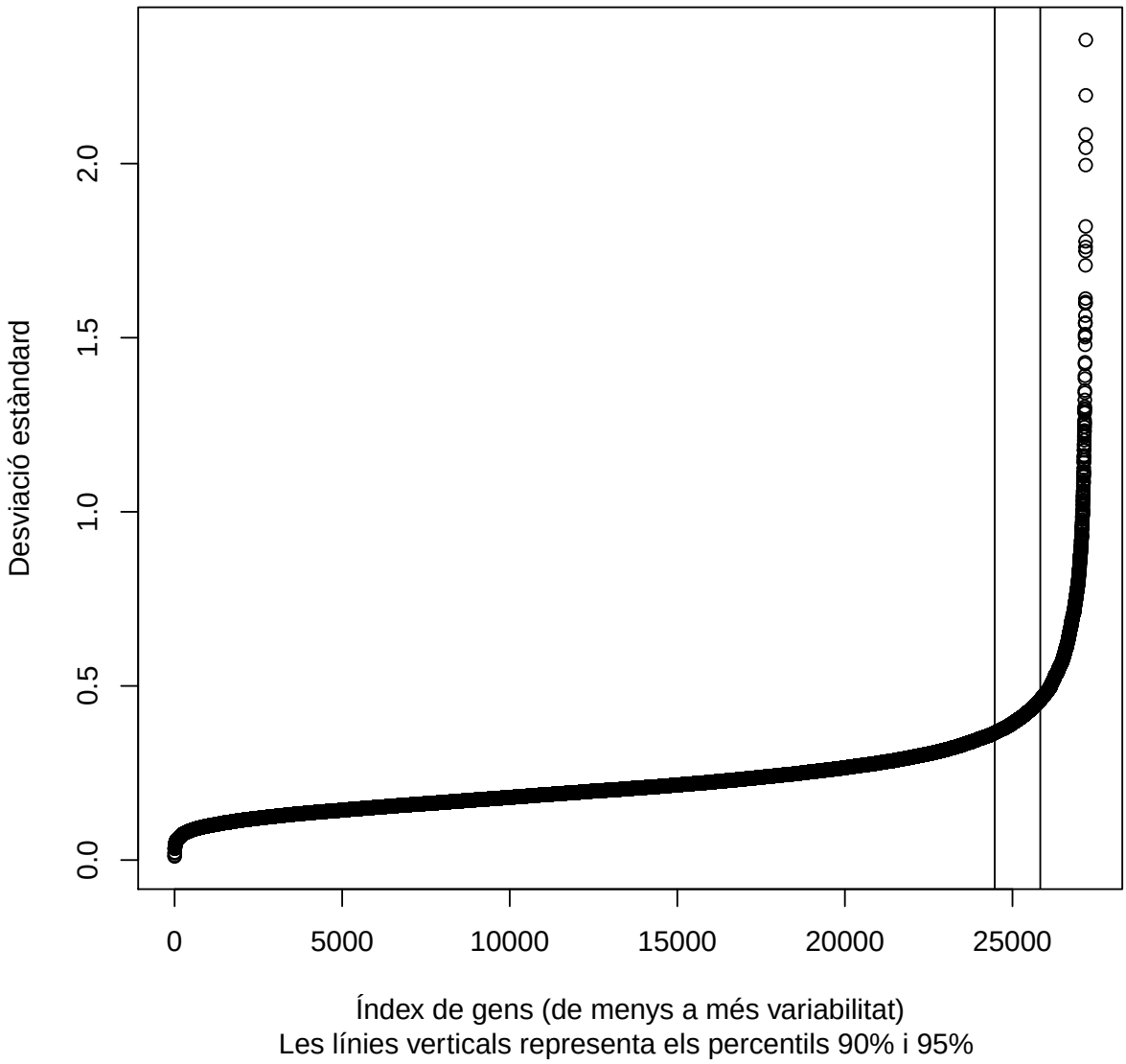


Figure 4: Figura 7.Distribució de la variabilitat de les dades filtrades

3.3.5. Filtrat dels resultats

Hi ha gens pels quins la seva variació pot ser atribuïda a una variabilitat aleatoria. Aquests gens es poden filtrar. Aquest procés es pot realitzr amb la llibreria `genefiltre` i la funció `nsfilter()`, cal però la informació d'un paquet amb anotacions sobre l'estudi en concret que ajudarà a realitzar el filtrat. Aquest paquet d'anotacions es pot trobar a http://www.bioconductor.org/packages/release/data/annotation/. També es descriu el paquet d'anotacions utilitzat quan s'ha executat `head(rawData)`, al final de la informació mostrada.

És possible que per alguns estudis en concret no existeixi aquesta informació, potser cal esperar una nova versió de Bioconductor.

L'estudi seguit en aquest exemple no ha permès instal · lar correctament el paquet d'anotacions, per tant no es podrà utilitzar aquesta informació i es realitzarà un filtrat com el següent:

```
> library(genefilter)
>
> # Com que en les anotacions no hi havia la entrada ENTREZ indico les entrades amb ENTREZ falses
> # es deixen la funció IQR i el punt de tall 0.5 que la funció té per defecte
> filtered <- nsFilter(all.eset,
+                      require.entrez = FALSE, remove.dupEntrez = FALSE,
+                      var.filter=TRUE, var.func=IQR, var.cutoff=0.5,
+                      filterByQuantile=TRUE)
>
> # per a veure els resultats del filtrat
> filtered$eset
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 13560 features, 12 samples
  element names: exprs
protocolData
  rowNames: CTRL.NORM.175 CTRL.NORM.176 ... SICD248.HYPO.210 (12 total)
  varLabels: exprs dates
  varMetadata: labelDescription channel
phenoData
  rowNames: CTRL.NORM.175 CTRL.NORM.176 ... SICD248.HYPO.210 (12 total)
  varLabels: Group STATE GEN ShortName
  varMetadata: labelDescription channel
featureData
  featureNames: 23064072 23064073 ... TSUnmapped00000823.hg.1 (13560
    total)
  fvarLabels: PROBEID ID SYMBOL GENENAME
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation: pd.clariom.s.human
```

```
> filtered$filter.log
```

```
$numLowVar
[1] 13561

$feature.exclude
[1] 68
```

```
> eset_filtered <-filtered$eset
>
> # la informaicó filtrada amb les anotacions es desa com a .csv:
>
```

```
> write.csv(exprs(eset_filtered), file="./results/normalized.Filtered.Data.csv")
```

Segons "filtered$eset" hem passat de *27189* dades a "eset_rma" a *13561* a "eset_filtered".

NOTA: No tinc clar si ha eliminat duplicats al no poder-li dir que remoqui "dupEntrez".

3.3.6. Identificació de genes diferencialment expresats

La identificació de gens expressats diferencialment consisteix bàsicament en comparar l'expressió gènica entre grups. Existeixden diferents metodologies però en aquest procés s'utlitzarà el Lineal Models for Microarrays, implementat al paquet `limma` igual que en el treball de Irizarry et al. (2003).

El primer pas per a l'anàlisi basat en models lineals és crear la matriu de disseny. Bàsicament es tracta d'una taula que descriu l'assignació de cada mostra a un grup o condició experimental. Té tantes files com mostres (en aquest cas 12) i tantes columnes com grups (en aquest cas 4). Cada fila conté una a la columna del grup al qual pertany la mostra i un zero a les altres. La matriu de disseny es pot definir manualment o a partir d'una variable de factor que pot haver estat introduïda al fitxer "targets.csv" amb aquest objectiu creat específicament per a ell. En aquest estudi:

```
> library(limma)
> designMat<- model.matrix(~0+Group, pData(eset_filtered))
> colnames(designMat) <- c("CTRL.HYPO","CTRL.NORM","SICD248.HYPO","SICD248.NORM")
> print(designMat)
```

```
                 CTRL.HYPO CTRL.NORM SICD248.HYPO SICD248.NORM
CTRL.NORM.175            0         1            0            0
CTRL.NORM.176            0         1            0            0
CTRL.NORM.177            0         1            0            0
SICD248.NORM.187         0         0            0            1
SICD248.NORM.188         0         0            0            1
SICD248.NORM.189         0         0            0            1
CTRL.HYPO.202            1         0            0            0
CTRL.HYPO.203            1         0            0            0
CTRL.HYPO.204            1         0            0            0
SICD248.HYPO.208         0         0            1            0
SICD248.HYPO.209         0         0            1            0
SICD248.HYPO.210         0         0            1            0
attr(,"assign")
[1] 1 1 1 1
attr(,"contrasts")
attr(,"contrasts")$Group
[1] "contr.treatment"
```

A continuació es defineix la **Matriu de Contrast** per escriure les comparacions entre grups de dos en dos. Conté tantes columnes com comparacions (en aquest exemple seran 3) i tantes files com grups (en aquestes dades 4 grups). En aquest exemple es realitzen tres comparacions, basicament enfocades a veure diferències entre els control amb els siCD248 en concentració normal d'oxigen (NORM) i el mateix però en hipòxia (HYPO). Es realitzat una tercera comparació entre grups els grups anteriors (INTR).

```
> cont.matrix <- makeContrasts (NORM = CTRL.NORM-SICD248.NORM,
+                               HYPO = CTRL.HYPO-SICD248.NORM,
+                               INTR = (CTRL.NORM-SICD248.NORM) - (CTRL.HYPO-SICD248.NORM),
+                               levels=designMat)
> print(cont.matrix)
```

```
             Contrasts
Levels        NORM HYPO INTR
```

```
  CTRL.HYPO       0    1   -1
  CTRL.NORM       1    0    1
  SICD248.HYPO    0    0    0
  SICD248.NORM   -1   -1    0
```

Una vegada definides la matriu de disseny i els contrastos, podem procedir a l'estimació del model, a estimar els contrastos i a realitzar proves de significació que conduiran a la decisió, per a cada gen i cada comparació, si es poden considerar expressades diferencialment. També s'utilitzarà el paquet `limma`

```
> #library(limma)
> fit<-lmFit(eset_filtered, designMat)
> fit.main<-contrasts.fit(fit, cont.matrix)
> fit.main<-eBayes(fit.main)
> head(fit.main)
```

```
An object of class "MArrayLM"
$coefficients
          Contrasts
                  NORM        HYPO       INTR
  23064072 -0.03670789  0.08977661 -0.1264845
  23064073  0.30209222 -0.25812078  0.5602130
  23064075 -0.41947550 -0.68679418  0.2673187
  23064076  0.16309724  0.64415387 -0.4810566
  23064077  0.14497099 -0.34292081  0.4878918
  23064083  0.46202848  0.58958544 -0.1275570

$rank
[1] 4

$assign
[1] 1 1 1 1

$qr
$qr
                 CTRL.HYPO  CTRL.NORM SICD248.HYPO SICD248.NORM
CTRL.NORM.175    -1.732051  0.0000000     0.000000    0.0000000
CTRL.NORM.176     0.000000 -1.7320508     0.000000    0.0000000
CTRL.NORM.177     0.000000  0.5773503    -1.732051    0.0000000
SICD248.NORM.187  0.000000  0.0000000     0.000000   -1.7320508
SICD248.NORM.188  0.000000  0.0000000     0.000000    0.5773503
7 more rows ...

$qraux
[1] 1.00000 1.57735 1.00000 1.57735

$pivot
[1] 1 2 3 4

$tol
[1] 1e-07

$rank
[1] 4
```

```
$df.residual
[1] 8 8 8 8 8 8

$sigma
 23064072  23064073  23064075  23064076  23064077  23064083
0.1085876 0.2537674 0.1497553 0.1328806 0.1121344 0.2312750

$cov.coefficients
         Contrasts
Contrasts      NORM       HYPO       INTR
     NORM 0.6666667  0.3333333  0.3333333
     HYPO 0.3333333  0.6666667 -0.3333333
     INTR 0.3333333 -0.3333333  0.6666667

$stdev.unscaled
          Contrasts
                NORM      HYPO      INTR
  23064072 0.8164966 0.8164966 0.8164966
  23064073 0.8164966 0.8164966 0.8164966
  23064075 0.8164966 0.8164966 0.8164966
  23064076 0.8164966 0.8164966 0.8164966
  23064077 0.8164966 0.8164966 0.8164966
  23064083 0.8164966 0.8164966 0.8164966

$genes
         PROBEID   ID SYMBOL GENENAME
23064072    <NA> <NA>   <NA>     <NA>
23064073    <NA> <NA>   <NA>     <NA>
23064075    <NA> <NA>   <NA>     <NA>
23064076    <NA> <NA>   <NA>     <NA>
23064077    <NA> <NA>   <NA>     <NA>
23064083    <NA> <NA>   <NA>     <NA>

$Amean
23064072 23064073 23064075 23064076 23064077 23064083
6.842809 5.633629 8.396632 5.085739 7.232466 6.284805

$method
[1] "ls"

$design
                CTRL.HYPO CTRL.NORM SICD248.HYPO SICD248.NORM
CTRL.NORM.175           0         1            0            0
CTRL.NORM.176           0         1            0            0
CTRL.NORM.177           0         1            0            0
SICD248.NORM.187        0         0            0            1
SICD248.NORM.188        0         0            0            1
7 more rows ...

$contrasts
              Contrasts
Levels         NORM HYPO INTR
  CTRL.HYPO       0    1   -1
  CTRL.NORM       1    0    1
```

```
  SICD248.HYPO    0    0    0
  SICD248.NORM   -1   -1    0

$df.prior
[1] 7.140804

$s2.prior
[1] 0.03316687

$var.prior
[1]  28.88996 155.50332 147.14527

$proportion
[1] 0.01

$s2.post
  23064072   23064073   23064075   23064076   23064077   23064083
0.02187257 0.04966852 0.02749202 0.02497200 0.02228621 0.04390409

$t
          Contrasts
                 NORM       HYPO       INTR
  23064072 -0.3039871  0.7434622 -1.0474493
  23064073  1.6601394 -1.4184956  3.0786350
  23064075 -3.0984814 -5.0730471  1.9745658
  23064076  1.2640539  4.9923909 -3.7283370
  23064077  1.1893470 -2.8133342  4.0026813
  23064083  2.7006101  3.4461953 -0.7455852

$df.total
[1] 15.1408 15.1408 15.1408 15.1408 15.1408 15.1408

$p.value
          Contrasts
                 NORM         HYPO        INTR
  23064072 0.76527549 0.4685824620 0.311326361
  23064073 0.11744992 0.1763014458 0.007575948
  23064075 0.00727397 0.0001336418 0.066845447
  23064076 0.22532527 0.0001563428 0.001989702
  23064077 0.25262110 0.0130108948 0.001133570
  23064083 0.01633586 0.0035574536 0.467335958

$lods
          Contrasts
                NORM       HYPO      INTR
  23064072 -6.443011 -7.0352117 -6.733871
  23064073 -5.174442 -6.3209106 -3.395417
  23064075 -2.642366  0.6335656 -5.455990
  23064076 -5.700521  0.4735023 -2.072712
  23064077 -5.787179 -3.9475678 -1.508915
  23064083 -3.405027 -2.6769318 -7.006163

$F
[1]  0.5807648  4.7487286 13.0784346 13.4740985  8.4502844  6.5751516
```

```
$F.p.value
[1] 0.5714572557 0.0250700009 0.0005023375 0.0004351035 0.0034305476
[6] 0.0088029504
```

```
> class(fit.main)
```

```
[1] "MArrayLM"
attr(,"package")
[1] "limma"
```

```
> results<-decideTests(fit.main)
> summary(results)
```

	NORM	HYPO	INTR
Down	269	2806	2488
NotSig	12738	9139	7230
Up	553	1615	3842

La funció `topTABLE()` proporciona els estadístics habituals de proves: com ara els p-valors moderats o ajustats que s'utilitzen per ordenar els gens que s'expressen de manera de més a menys diferenciats en un contrast, alhora, per si de cas faig que es treguin els registres amb NA (sense dades):

```
> #construïm una taula amb la informació de la comparació
> topTAB_NORM_0 <- topTable (fit.main, number=nrow(fit.main), coef="NORM", adjust="fdr")
> # es mostren només els resultats dels primers gens
> topTAB_NORM <- na.omit(topTAB_NORM_0)
> h_NORM <- head(topTAB_NORM)
> knitr::kable(
+    h_NORM, booktabs = TRUE,
+    caption = 'NORM')
```

Table 3: NORM

	PROBEID	ID	SYMBOL	GENENAME
TC1100011282.hg.1	TC1100011282.hg.1	NM_020404	CD248	CD248 molecule, endosialin
TC1100012126.hg.1	TC1100012126.hg.1	NM_001304441	MMP8	matrix metallopeptidase 8
TC0100013162.hg.1	TC0100013162.hg.1	NM_000300	PLA2G2A	phospholipase A2, group IIA (platelets, synovia
TC1900009443.hg.1	TC1900009443.hg.1	NM_000064	C3	complement component 3
TC0100010584.hg.1	TC0100010584.hg.1	NM_001282692	FMO1	flavin containing monooxygenase 1
TC1100009048.hg.1	TC1100009048.hg.1	NM_001931	DLAT	dihydrolipoamide S-acetyltransferase

```
> topTAB_HYPO_0 <- topTable (fit.main, number=nrow(fit.main), coef="HYPO", adjust="fdr")
> topTAB_HYPO <- na.omit(topTAB_HYPO_0)
> h_HYPO <- head(topTAB_HYPO)
> knitr::kable(
+    h_HYPO, booktabs = TRUE,
+    caption = 'HYPO')
```

Table 4: HYPO

	PROBEID	ID	SYMBOL	GENENAME
TC0X00008794.hg.1	TC0X00008794.hg.1	NM_001142805	SLC6A8	solute carrier family 6 (neurotransmitter transp

	PROBEID	ID	SYMBOL	GENENAME
TC1700009318.hg.1	TC1700009318.hg.1	NM_182705	FAM101B	family with sequence similarity 101, member B
TC0600008109.hg.1	TC0600008109.hg.1	NM_001025366	VEGFA	vascular endothelial growth factor A
TC0200009049.hg.1	TC0200009049.hg.1	NM_016133	INSIG2	insulin induced gene 2
TC1200011327.hg.1	TC1200011327.hg.1	NM_001300965	CSRP2	cysteine and glycine-rich protein 2
TC0100012787.hg.1	TC0100012787.hg.1	NM_001135585	SLC2A5	solute carrier family 2 (facilitated glucose/fruct

```
> topTAB_INTR_0 <- topTable (fit.main, number=nrow(fit.main), coef="INTR", adjust="fdr")
> topTAB_INTR <- na.omit(topTAB_INTR_0)
> h_INTR <- head(topTAB_INTR)
> knitr::kable(
+    h_INTR, booktabs = TRUE,
+    caption = 'INTR')
```

Table 5: INTR

	PROBEID	ID	SYMBOL	GENENAME
TC1700009318.hg.1	TC1700009318.hg.1	NM_182705	FAM101B	family with sequence similarity 101, member B
TC0X00008794.hg.1	TC0X00008794.hg.1	NM_001142805	SLC6A8	solute carrier family 6 (neurotransmitter transp
TC0600008109.hg.1	TC0600008109.hg.1	NM_001025366	VEGFA	vascular endothelial growth factor A
TC0800011881.hg.1	TC0800011881.hg.1	NM_001135242	NDRG1	N-myc downstream regulated 1
TC0700006890.hg.1	TC0700006890.hg.1	NM_000600	IL6	interleukin 6
TC0300013970.hg.1	TC0300013970.hg.1	NM_004567	PFKFB4	6-phosphofructo-2-kinase/fructose-2,6-biphosph

També es pòden mostrar gràficament els resultats. Una de les manera de fer-ho és mitjançant gràfiques de "volcans". Aquestes figures mostren si hi ha molts o pocs gens amb un gran canvi i expressats significativament o si aquest nombre és baix. Aquests gràfics representen en l'eix X els canvis d'expressió a escala logarítmica ("efecte biològic") i en l'eix Y el "minus logaritme" del `p-value`. Les figures següent mostra la gràfica en forma de "volcans" per a la comparació entre els difernts grups i es resalten anotant el nom en blau dels gens més diferents en cada cas:

```
> volcanoplot(fit.main, coef=1,highlight = 4, names = topTAB_INTR$SYMBOL,
+             main=paste("Gens expressats de manera diferent", colnames(cont.matrix)[3],sep="\n"))
> abline(v=c(-1,1))
```

3.3.7. Anotació dels resultats

A partir de la taula (o taules) anteriors es pot afegir informació complementària, extreta dels paquets d'anotacions descrits en el punt 3.3.5 per a fer més entenedors els resultats.

En aquest exemple concret, hem afegit la informaicó just després de normalitzar les dades, amb la llibreria `affycoretools` i la funció `annotateEset`, tal i com es mostra cometnat en el codi següent següent. A partir d'aquesta funció, ja hem incorporat informació descriptiva en els camps PROBEID, ID, SYMBOL i GENENAME, tal i com s'ha explicat anteriorment i es pot veure en el fragment de taula que es mostra a continuació.

```
> #library(affycoretools)
> #all.eset <- annotateEset(eset_rma, annotation(eset_rma))
> #all.eset
>
> exemple <- (topTAB_INTR)[1:5,1:4]
> knitr::kable(exemple, booktabs = TRUE,
+              caption = 'Exemple dels camps documentats per el paquet amb anotacions')
```

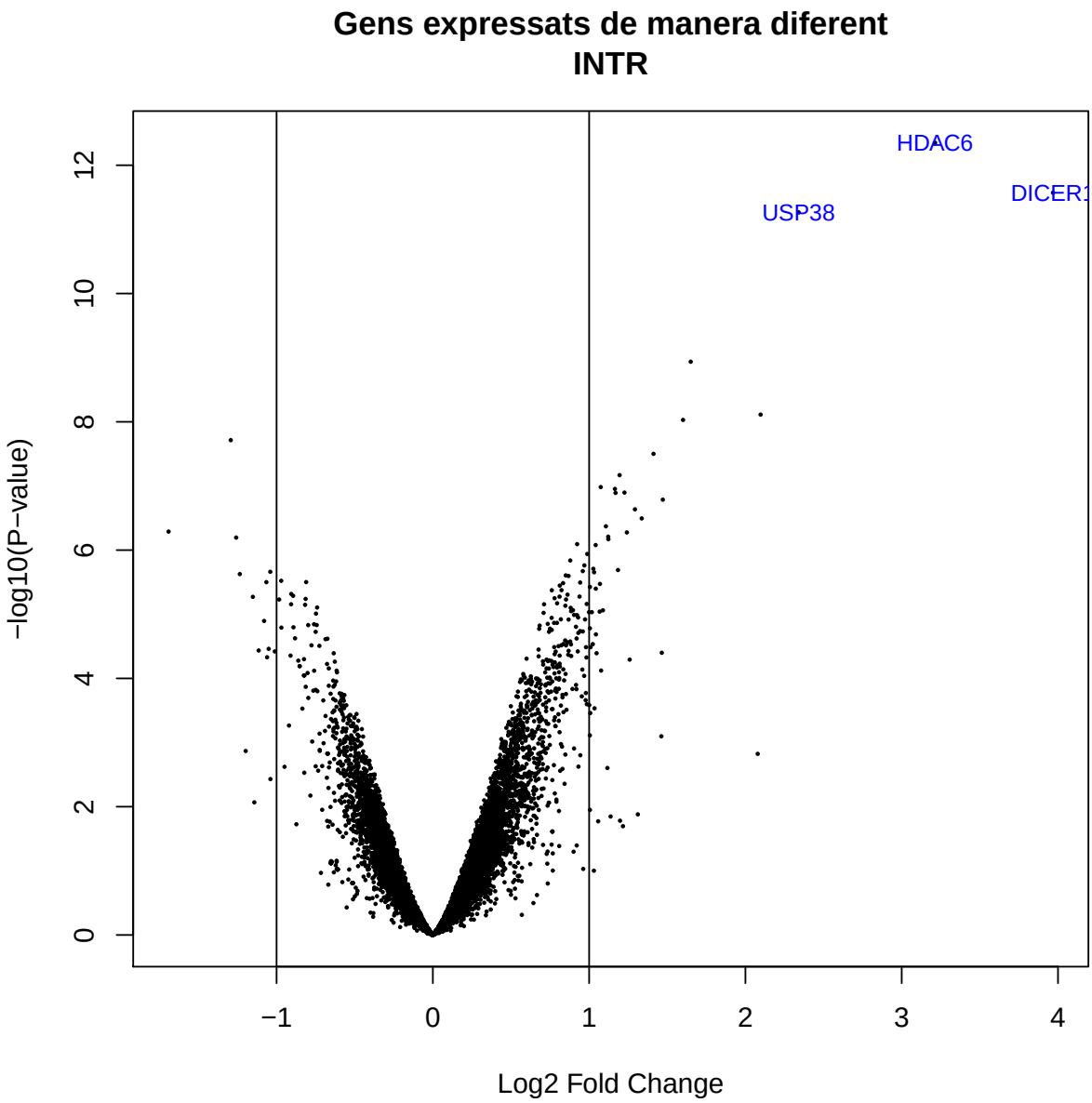


Figure 5: Figura 8.Comparació de l'expressió dels gens entre els dos grups de conrol i els dos grups de siCD248. Destaquen també els 4 gens més extrems

Table 6: Exemple dels camps documentats per el paquet amb anotacions

	PROBEID	ID	SYMBOL	GENENAME
TC1700009318.hg.1	TC1700009318.hg.1	NM_182705	FAM101B	family with sequence similarity 101, member B
TC0X00008794.hg.1	TC0X00008794.hg.1	NM_001142805	SLC6A8	solute carrier family 6 (neurotransmitter transp
TC0600008109.hg.1	TC0600008109.hg.1	NM_001025366	VEGFA	vascular endothelial growth factor A
TC0800011881.hg.1	TC0800011881.hg.1	NM_001135242	NDRG1	N-myc downstream regulated 1
TC0700006890.hg.1	TC0700006890.hg.1	NM_000600	IL6	interleukin 6

3.3.8. Comparación entre diferents comparacions

Per a comparar els resultats dels diferents grups es pot obteni runa matriu de comparacions a patir del model:

```
> library(limma)
> res<-decideTests(fit.main, method="separate", adjust.method="fdr", p.value=0.1, lfc=1)
>
> sum.res.rows<-apply(abs(res),1,sum)
> res.selected<-res[sum.res.rows!=0,]
> print(summary(res))
```

```
        NORM  HYPO  INTR
Down      15   310   277
NotSig 13500 12933 12800
Up        45   317   483
```

O també un diagrama de Venn:

```
> vennDiagram (res.selected[,1:3], cex=0.9)
> title("Gens en comú entre les tres comparacions\n seleccionats amb FDR < 0.1 i logFC > 1")
```

3.3.9. Anàlisis de significació biològica ("Gene Enrichment Analysis")

A continuació es pretén interpretar bé els resultats dotant-los de un significat biològic. Per a la correcta interpretació és necessari conèixer bé el problema biològic però alhora també hi ha aproximacions estadístiques que poden ajudar a realitzar aquesta interpretació.

Hi ha diferents variants d'aquesst tipus d'anàlisis. A continaució s'utilitzarà l'anàlisi d'enriquiment bàsic implementat al paquet *CulsterProfiles* https://yulab-smu.github.io/clusterProfiler-book/

```
> library(AnnotationDbi)
> listOfTables <- list(NORM = topTAB_NORM,
+                      HYPO  = topTAB_HYPO,
+                      INTR = topTAB_INTR)
>
> listOfSelected <- list()
> for (i in 1:length(listOfTables)){
+  # select the toptable
+  topTab <- listOfTables[[i]]
+  # select the genes to be included in the analysis
+  whichGenes<-dplyr::filter(topTab,adj.P.Val<0.15)
+  IDs <- dplyr::select(whichGenes, ID)
+  IDs <- IDs$ID
+  listOfSelected[[i]] <- IDs
+  names(listOfSelected)[i] <- names(listOfTables)[i]
```

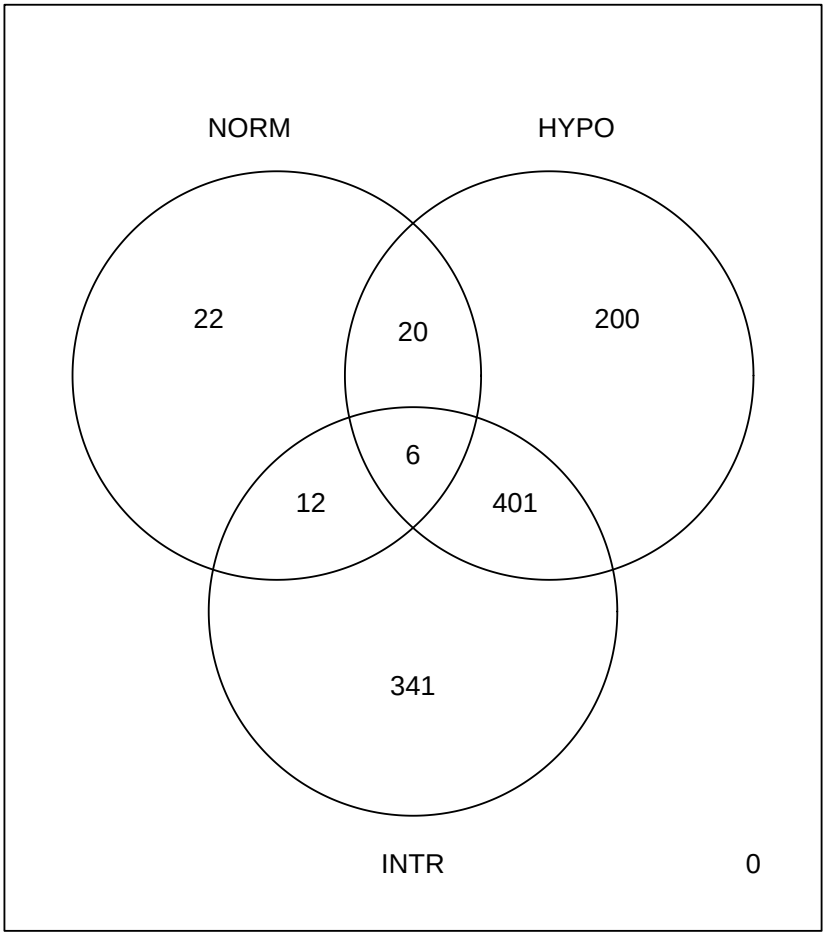


Figure 6: Figura 9.Diagrama de Venn amb la comparació dels tres grups.

```
+ }
> sapply(listOfSelected, length)
```

```
NORM HYPO INTR
1851 5237 6893
```

Finalment s'han seleccionat 1851, 5237 i 6893 per a cada grup.

A continuació es compararia la llista de gens sel · leccionats amb una base de dades de gens, o amb tots els gens. Una de les opcions del clusterProfiles és compara la nostra llista de gens amb la de una base de de dades de lliure accés (PAthway) on existeix informació de diferents espècies, en aquest cas es descarrega la base corresponent als homes.

```
> # treballem amb les dades que agafem del csv que hem creat amb les dades filtrades i normalitzades
> d <- read.csv(file="./results/normalized.Filtered.Data.csv")
> geneList <- d[,2]
> names(geneList) <- as.character(d[,1])
> geneList <- sort(geneList, decreasing = TRUE)
>
> data(geneList, package="DOSE")
> head(geneList)
```

```
    4312     8318    10874    55143    55388      991
4.572613 4.514594 4.418218 4.144075 3.876258 3.677857
```

```
> gene <- names(geneList)[abs(geneList) > 2]
> head(gene)
```

```
[1] "4312"  "8318"  "10874" "55143" "55388" "991"
```

```
> #BiocManager::install("qusage")
> #BiocManager::install("rWikiPathways")
> #BiocManager::install("clusterProfiler")
> #BiocManager::install("magrittr")
> library(rWikiPathways)
> library(qusage)
> library(magrittr)
> library(clusterProfiler)
>
> # base de dades Pathway per a Homo sapiens en format GTM
> wpgmtfile <- downloadPathwayArchive(organism = "Homo sapiens", format = "gmt")
> wpgmtfile
```

```
[1] "wikipathways-20200410-gmt-Homo_sapiens.gmt"
```

```
> wp2gene <- read.gmt(wpgmtfile)
> wp2gene <- wp2gene %>% tidyr::separate(ont, c("name","version","wpid","org"), "%")
> wpid2gene <- wp2gene %>% dplyr::select(wpid, gene) #TERM2GENE
> wpid2name <- wp2gene %>% dplyr::select(wpid, name) #TERM2NAME
>
> head(wpid2name)
```

wpid	name
WP4400	FABP4 in ovarian cancer
WP4400	FABP4 in ovarian cancer
WP23	B Cell Receptor Signaling Pathway
WP23	B Cell Receptor Signaling Pathway

wpid	name
WP23	B Cell Receptor Signaling Pathway
WP23	B Cell Receptor Signaling Pathway

```
> # procés d'enriquimetn de la informació
> ewp <- enricher(gene, TERM2GENE = wpid2gene, TERM2NAME = wpid2name)
>
> # per a veure la taula resultant de l'enriquiment. Veiem que té 6 registres.
> ewp[1:6,]
```

	ID	Description	GeneRatio	BgRa
WP2446	WP2446	Retinoblastoma Gene in Cancer	11/103	88/71
WP2361	WP2361	Gastric Cancer Network 1	6/103	29/71
WP179	WP179	Cell Cycle	10/103	122/7
WP3942	WP3942	PPAR signaling pathway	7/103	67/71
WP4240	WP4240	Regulation of sister chromatid separation at the metaphase-anaphase transition	4/103	16/71
WP2328	WP2328	Allograft Rejection	7/103	90/71
La taula	anterior	mostra les relacions i les estadístiques associades.		

La informació descrita en la taula anterior, guardada com a objecte `ewp` es pot mostrar gràficament en un plot de xarxa amb la funció `cnetplot()`:

```
> cnetplot(ewp, categorySize = "geneNum", schowCategory = 10,
+          vertex.label.cex = 0.50)
```

4.Resultats

Com a resultat tenim els fitxers que s'han anat emmagatzemant a la carpeta corresponent de resultats:

1- Controls de qualitat de les dades brutes

2- Controls de qualitat de les les dades normallitzades

3- csv amb les dades normalitzades

4- csv amb les dades normalitzades i filtrades

5- dades normalitzades amb anotacions

També s'han generat les següents explicatives:

1.Taula resum de la qualitat de les dades, extret del fitxer "index.html" produït per la arrayQualityMetrics() sobre les dades brutes

2.Components principals de les dades brutes

3.Boxplot de les dades brutes

4.Taula resum de la qualitat de les dades, extret del fitxer "index.html" produït per la arrayQualityMetrics() sobre les dades normalitzades

5.Components principals de les dades normalitzades

6.Boxplot de les dades normalitzades

7.Distribució de la variabilitat de les dades filtrades

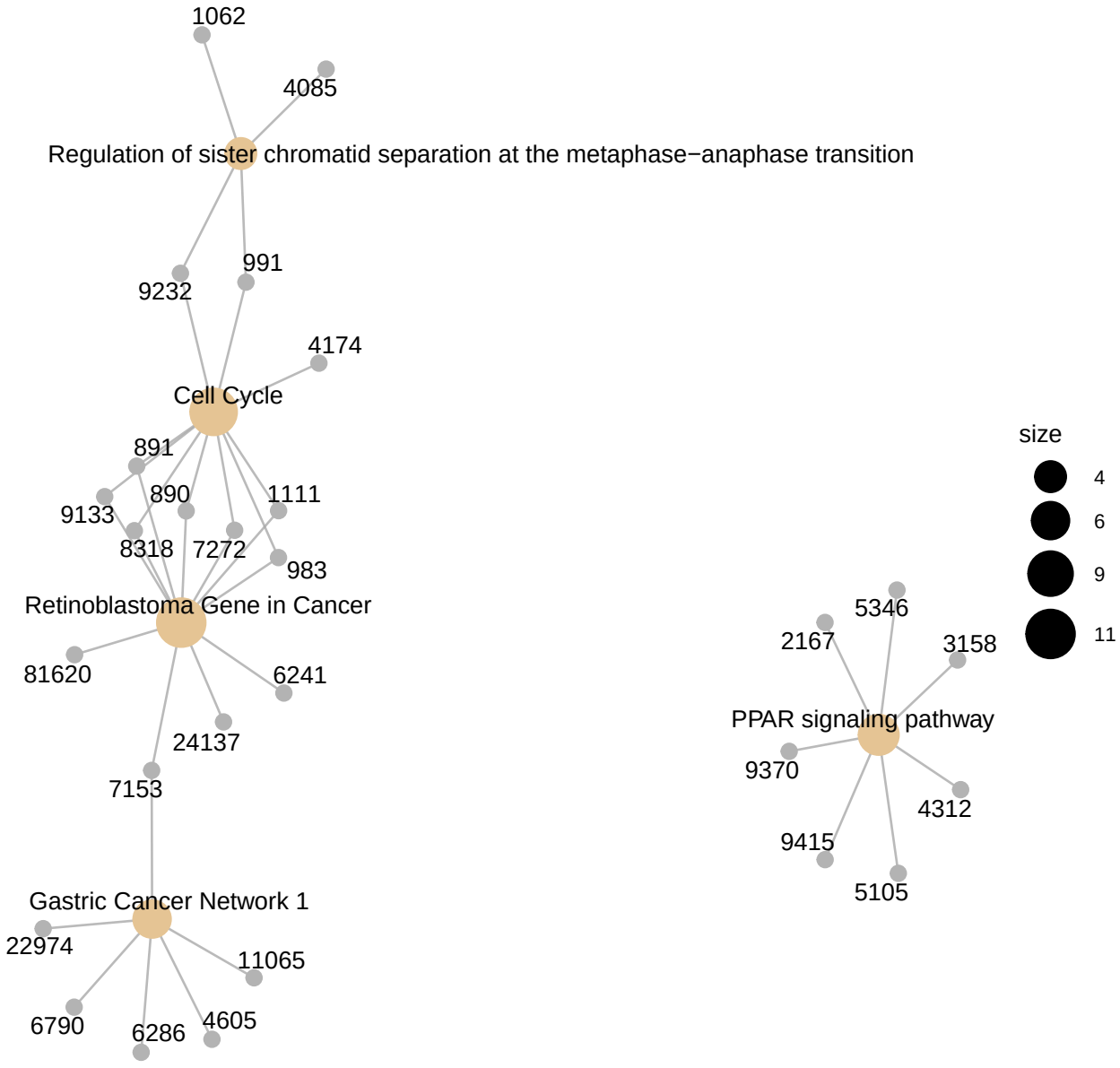


Figure 7: Figura 10.Xarxa obtinguda amb l'enriquiment dels gens del treball amb clusterProfiler

8.Comparació de l'expressió dels gens entre els dos grups de conrol i els dos grups de siCD248. Destaquen també els 4 gens més extrems

9.Diagrama de Venn amb la comparació dels tres grups

10.Xarxa obtinguda amb l'enriquiment dels gens del treball amb "clusterProfiler

5.Discusió

El procés de realització amb Bioconductor planteja algunes dificultats alhora d'instal · lar alguns paquetes. Si hi ha mala connexió de internet, es potden produïr talls en la descàrrega, convé repetir la instal · lació. En alguna paquets convé instal · lar paquets addicionals. Alguns són força voluminosos i el procées d'instal · lació pot acabar sent ferragós.

No obstant, és una eina molt potent per a l'anàlisis de grans volumsn de dades òmiques.

El procés preparat és molt interessant, s'obté informació molt valuosa però per la qual es necessita un coneixament expert sobre la informació tractada, més enllà dels anàlisis estadístics, per a treure conclusions rellevants.

6.Referències

Gonzalo Sanz, Ricardo, and Alex Sánchez-Pla. 2019. "Statistical Analysis of Microarray Data." In *Microarray Bioinformatics*, edited by Verónica Bolón-Canedo and Amparo Alonso-Betanzos, 87–121. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-9442-7_5.

Irizarry, Rafael A., Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. 2003. "Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data." *Biostatistics* 4 (2): 249–64. https://doi.org/10.1093/biostatistics/4.2.249.

Petrus, Paul, Tara L. Fernandez, Michelle M. Kwon, Jenny L. Huang, Victor Lei, Nooshin Seyed Safikhan, Subashini Karunakaran, et al. 2019. "Specific loss of adipocyte CD248 improves metabolic health via reduced white adipose tissue hypoxia, fibrosis and inflammation." *EBioMedicine* 44: 489–501. https://doi.org/10.1016/j.ebiom.2019.05.057.