# PAC1

Xavier David Lluesma

2024-11-04

# PAC 1: XAVIER DAVID LLUESMA

## Selecció de dades

S'ha triat una base de dades corresponent a un estudi d'anàlisi no dirigit del lipidoma obtingudes al repositori *Metabolomics workbench*. En aquest estudi [1] es compara el lipidoma en plasma de controls amb el de pacients d'esquizofrènia i trastorn bipolar. Com que els pacients no han estat tractats de forma prèvia a l'extracció de sang, les potencials diferències entre els perfils poden servir com a marcadors de la malaltia. El nombre de pacients és de 30 per cadascun dels grups, fent una N total de 90 mostres. A aquestes mostres s'afegeixen mostres emprades únicament per a establir la qualitat de l'experiment (QC).

## Càrrega de les dades i generació d'un SummarizedExperiment

Es descarreguen les dades des del web del repositori, i es lligen des d'un arxiu de tipus txt. S'extrau la informació rellevant en tal de sintetitzar-les a un objecte tipus *SummarizedExperiment*.

En aquest experiment cada mostra es troba codificada amb unes sigles (BD, SZ, CT, QC) en funció del grup a què pertanyen. Les dades obtingudes són senyals d'intensitat reflectint els diferents tipus de compostos.

```
# Lectura del .txt

dades<-read.delim(file.path(getwd(), "ST002554_AN004205_Results.txt"), check.names = TRUE)

# Obtinc la informació dels senyals.

comptatge<-as.matrix(dades[ , -1])

# Extrac la informació dels metabolits, per desar-la com a nom de les fileres.

metabolit<-data.frame(row.names = dades$mz_rt)

# Extrac les metadades, codificant una nova variable per grups.

metadades<-data.frame(sampleID = colnames(comptatge), row.names = colnames(comptatge))
metadades$grup<-ifelse(grepl("^CT", metadades$sampleID), "control",
                  ifelse(grepl("^QC", metadades$sampleID), "qualitat",
                  ifelse(grepl("^BD", metadades$sampleID), "bipolar", "esquizofrenia")))

# Genere l'objecte.
```

```r
se<-SummarizedExperiment(assays = list(counts = comptatge),
                         rowData = metabolit,
                         colData = metadades)
se
```

```
## class: SummarizedExperiment
## dim: 867 103
## metadata(0):
## assays(1): counts
## rownames(867): 100.072822683704_851.29554 100.932463481368_1185.06405
##   ... 1307.56856639327_94.097913 1308.57667145311_94.065372
## rowData names(0):
## colnames(103): BD01 BD02 ... QC_5_3 QC_5_4
## colData names(2): sampleID grup
```

```r
# Dese les dades.

save(se, file = "se.Rda")
```

## Normalització i filtratge

En tal de garantir l'estabilitat de les mesures, les mostres QC varen ser mesurades repetidament al llarg de l'experiment. Comprove la variabilitat dins de les mostres de qualitat, de forma semblant a com es fa en l'article associat a les dades. Aquells compostos que presenten una desviació estàndard relativa a les mostres QC per damunt del 20% són eliminats.

```r
# Faig un subset de les mostres de qualitat.

se_qc<-se[, se$grup == "qualitat"]

# Calcule la desviació estàndard relativa.

rsd<-function(x) {
  sd(x) / mean(x) * 100
}

# Applique la funció per fileres.
# Cal cridar la matriu amb els valors d'intensitat amb assay.

valors_dsr<-apply(assay(se_qc), 1, rsd)

# Identifique els compostos que tenen una desviació estàndard relativa major del 20 %.

compostos_variables<-names(valors_dsr[valors_dsr > 20])

# Verifiquem si aquests noms estan en la matriu original i els eliminem

se_filtrat<-se[!rownames(se)%in%compostos_variables,]
```

Donat que les mostres de qualitat han estat testades repetidament, seria interessant agafar-les com a referència i realitzar una normalització tipus BRDG [2]. Tanmateix, tot i que s'indica a l'article, no queda reflectit a les dades penjades quines de les mostres problema han estat mesurades al mateix moment que quines

mostres QC. Així doncs, tot i que aquesta opció no sembla ser la més recomanable, es farà una normalització tipus TIC (*Total Ion Current*).

```
# Genere una còpia de les dades.

se_normalitzat<-se_filtrat
dades_normalitzades<-assay(se_filtrat)

# Apply the TIC normalization function within the loop

for (i in 1:ncol(dades_normalitzades)){
  dades_normalitzades[,i]<-dades_normalitzades[,i]/sum(dades_normalitzades[,i])
}

assay(se_normalitzat)<-dades_normalitzades
```

Una vegada emprades, elimine les dades corresponents al grup de qualitat.

```
se_problema<-se_normalitzat[,se_normalitzat$grup!= "qualitat"]

# Dese aquestes dades també.

save(se_problema, file = "Dades processades.Rda")
```

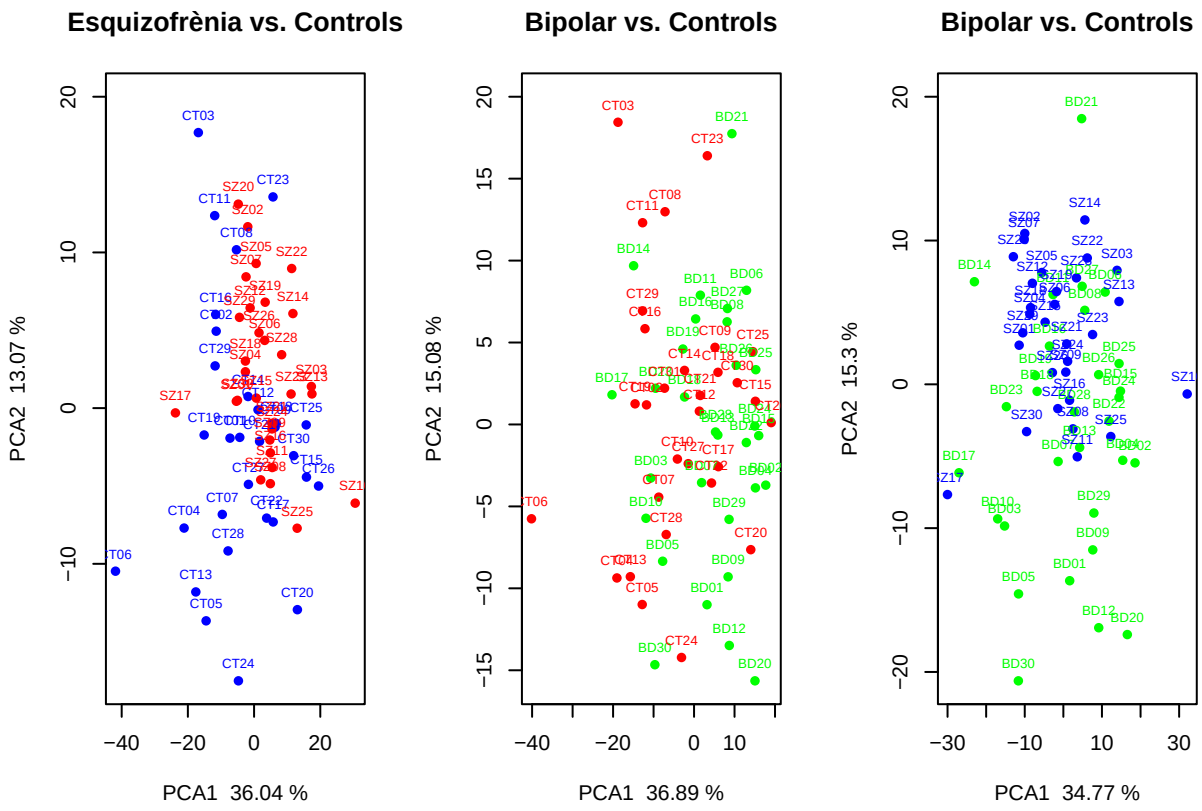
## Anàlisi Exploratòria: PCA

Faig una anàlisi per PCA, per comparar els pacients amb esquizofrènia amb els controls, i també comparant bipolar amb controls.

```
# Prepare els setting per mostrar ambdues gràfiques una al costat de l'altra.

par(mfrow=c(1,3))

# Aïlle les mostres de controls i pacients amb esquizofrènia.

se_esquizo<-se_problema[, se_problema$grup == "control" |
                          se_problema$grup == "esquizofrenia"]
int_escalada<-scale(t(assay(se_esquizo)), center = TRUE, scale = TRUE)
S<-cov(int_escalada)

# Calcule els eigenvalues.

EIG<-eigen(S)

# Calcule el percentatge de la variància explicada per cadascuna de les components.

pes<-EIG$values / sum(EIG$values)
eigenvecs<-EIG$vectors
PCAS<-int_escalada %*% eigenvecs

# Recupere el nom de les mostres.
```

```r
sample_labels<-colnames(se_esquizo)

# Definisc els colors per grup.

grup_colors<-as.numeric(factor(se_esquizo$grup))
palette_colors<-c("blue", "red")

# Prepare les etiquetes. i codifique la grâfcia.

xlabel<-paste("PCA1 ", round(pes[1] * 100, 2), "%")
ylabel<-paste("PCA2 ", round(pes[2] * 100, 2), "%")
plot(PCAS[, 1], PCAS[, 2], main = "Esquizofrènia vs. Controls",
     xlab = xlabel, ylab = ylabel,
     pch = 16, col = palette_colors[grup_colors],
     ylim = c(min(PCAS[, 2]), 20))

# Afegisc les etiquetes de les mostres.

text(PCAS[, 1], PCAS[, 2], labels = sample_labels,
     pos = 3, cex = 0.6, col = palette_colors[grup_colors])

# Repetisc el procés amb les mostres de bipolar.

se_bipolar<-se_problema[, se_problema$grup == "control" | se_problema$grup == "bipolar"]

int_escalada<-scale(t(assay(se_bipolar)), center = TRUE, scale = TRUE)

S<-cov(int_escalada)

EIG<-eigen(S)

pes<-EIG$values / sum(EIG$values)
eigenvecs<-EIG$vectors
PCAS<-int_escalada %*% eigenvecs

sample_labels<-colnames(se_bipolar)

grup_colors<-as.numeric(factor(se_bipolar$grup))
palette_colors<-c("green", "red")

xlabel<-paste("PCA1 ", round(pes[1] * 100, 2), "%")
ylabel<-paste("PCA2 ", round(pes[2] * 100, 2), "%")

plot(PCAS[, 1], PCAS[, 2], main = "Bipolar vs. Controls",
     xlab = xlabel, ylab = ylabel,
     pch = 16, col = palette_colors[grup_colors],
     ylim = c(min(PCAS[, 2]), 20))
text(PCAS[, 1], PCAS[, 2], labels = sample_labels,
     pos = 3, cex = 0.6, col = palette_colors[grup_colors])

# I per a esquizofrènia front a bipolar.

bip_esqui<-se_problema[, se_problema$grup == "esquizofrenia" | se_problema$grup == "bipolar"]
```

```
int_escalada<-scale(t(assay(bip_esqui)), center = TRUE, scale = TRUE)

S<-cov(int_escalada)

EIG<-eigen(S)

pes<-EIG$values / sum(EIG$values)
eigenvecs<-EIG$vectors
PCAS<-int_escalada %*% eigenvecs

sample_labels<-colnames(bip_esqui)

grup_colors<-as.numeric(factor(bip_esqui$grup))
palette_colors<-c("green", "blue")

xlabel<-paste("PCA1 ", round(pes[1] * 100, 2), "%")
ylabel<-paste("PCA2 ", round(pes[2] * 100, 2), "%")

plot(PCAS[, 1], PCAS[, 2], main = "Bipolar vs. Controls",
     xlab = xlabel, ylab = ylabel,
     pch = 16, col = palette_colors[grup_colors],
     ylim = c(min(PCAS[, 2]), 20))
text(PCAS[, 1], PCAS[, 2], labels = sample_labels,
     pos = 3, cex = 0.6, col = palette_colors[grup_colors])
```



Sembla haver-hi un patró diferent de dispersió entre les mostres de pacients amb esquizofrènia i les mostres dels controls, agrupant-se les primeres cap a valors alts d'ambdues components. En el cas de controls i pacients de trastorn bipolar, el patró de dispersió sembla ser el mateix, no hi ha tendències subjacents a les dades. Aquesta semblança entre pacients de trastorn bipolar s'observa amb la mateixa diferència entre pacients amb esquizofrènia i trastorn bipolar que entre controls i pacients amb esquizofrènia. Note's que, tot i que l'anàlisi no ha estat el mateix, el resultat obtingut és molt semblant al que presenten els autors de l'article, si bé els percentatges de la variància explicada per les components no són equivalents, el que possiblement reflectisc les diferències en el processament previ de les dades.

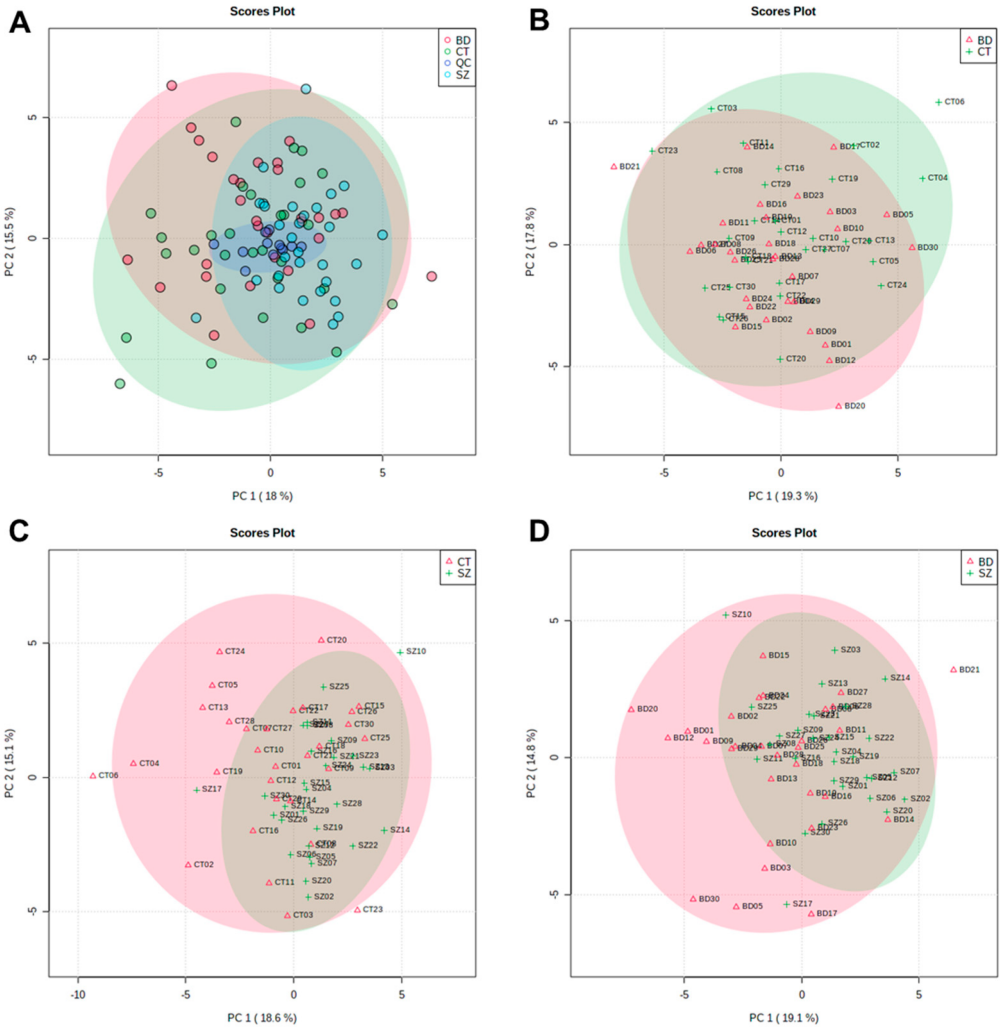



Figure 1: PCA de l'article original.

## Selecció de gens amb diferències significatives

Una vegada hem observat que semblen existir diferències latents entre el lipidoma dels pacients amb esquizofrènia i tant controls com pacients amb trastorn bipolar, passem a identificar els metabòlits que presenten diferències significatives. En primer lloc, i en tal de mantindre un error tipus I al 0.05, ajuste el p-valor al total de comparacions que seran realitzades.

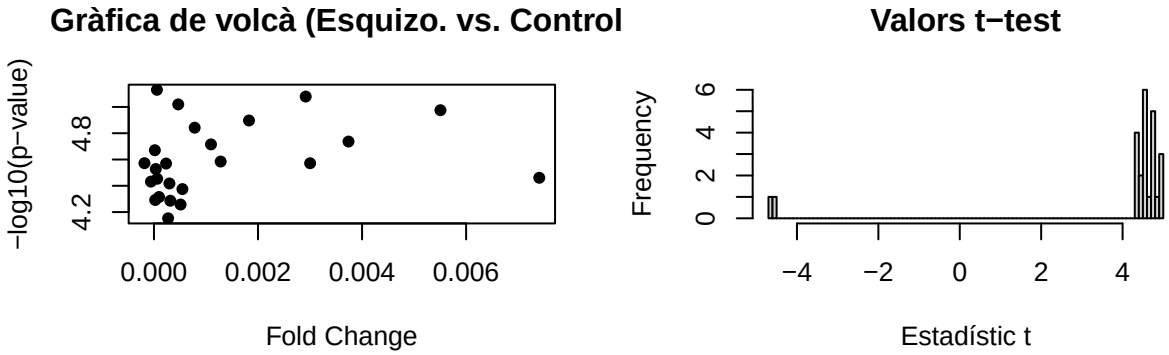
```
nombre_metabolotits<-nrow(assay(se_problema))
pvalor<-0.05/nombre_metabolotits
pvalor
```

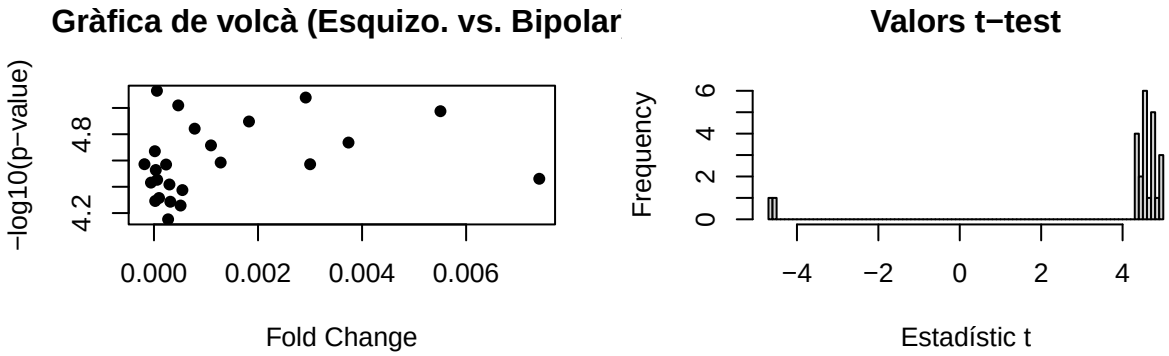
```
## [1] 0.0001256281
```

Per triar els gens de major interés, es fa un t-test amb el llindar de significativitat a un p-valor ajustat. Es pot observar com per a la majoria de les mostres la presència dels metabòlits estudiats amb diferències significatives és major als controls que als pacients amb esquizofrènia. Així mateix, mentre no trobem cap diferència significativa entre trastorn bipolar i controls, entre trastorn bipolar i esquizofrènia es repliquen aquestes diferències.

```
# Prepare els setting per mostrar ambdues gràfiques una al costat de l'altra.

par(mfrow=c(2,2))

# Extact la mtraiu amb els valors.

x<-assay(se_esquizo)

# Definisc els grups

group<-factor(colData(se_esquizo)$grup)

# Declare una funció per calcular els valors del t-test.

ttest<-function(expression_values){
  controls<-expression_values[group == "control"]
  pacients<-expression_values[group == "esquizofrenia"]
  tt<-t.test(controls, pacients)
  fc<-tt$estimate[1] - tt$estimate[2]
  return(c(t_statistic = tt$statistic,
           p_value = tt$p.value,
           fold_change = fc))
}

# Aplique la funció i filtre en funció del p-valor

ans<-apply(x, 1, ttest)
ans<-data.frame(t(ans))
ans_filtrat <- ans[ans$p_value < pvalor, ]

# Realitze el volcano plot.

plot(ans_filtrat$fold_change.mean.of.x, -log10(ans_filtrat$p_value), pch = 16,
     xlab = "Fold Change", ylab = "-log10(p-value)",
     main = "Gràfica de volcà (Esquizo. vs. Control)")

# I la distribució de valors de t-test.

hist(ans_filtrat$t_statistic.t, breaks = 100,
     main = "Valors t-test", xlab = "Estadístic t")
```

```r
# Faig el mateix amb les altres dues comparacions.

x<-assay(se_bipolar)
group<-factor(colData(se_bipolar)$grup)

ttest<-function(expression_values){
  controls<-expression_values[group == "control"]
  pacients<-expression_values[group == "bipolar"]
  tt<-t.test(controls, pacients)
  fc<-tt$estimate[1] - tt$estimate[2]
  return(c(t_statistic = tt$statistic,
           p_value = tt$p.value,
           fold_change = fc))
}

ans2<-apply(x, 1, ttest)
ans2<-data.frame(t(ans))
ans_filtrat2 <- ans[ans$p_value < pvalor, ]

# Al cas del trastorn bipolar i els controls,
# cap dels metabolits compleix els requisits del filtrat pel p-valor.

x<-assay(bip_esqui)
group<-factor(colData(bip_esqui)$grup)

ttest<-function(expression_values){
  controls<-expression_values[group == "bipolar"]
  pacients<-expression_values[group == "esquizofrenia"]
  tt<-t.test(controls, pacients)
  fc<-tt$estimate[1] - tt$estimate[2]
  return(c(t_statistic = tt$statistic,
           p_value = tt$p.value,
           fold_change = fc))
}

ans3<-apply(x, 1, ttest)
ans3<-data.frame(t(ans))
ans_filtrat3 <- ans[ans$p_value < pvalor, ]

plot(ans_filtrat3$fold_change.mean.of.x, -log10(ans_filtrat$p_value), pch = 16,
     xlab = "Fold Change", ylab = "-log10(p-value)",
     main = "Gràfica de volcà (Esquizo. vs. Bipolar)")


hist(ans_filtrat3$t_statistic.t, breaks = 100, main = "Valors t-test", xlab = "Estadístic t")
```



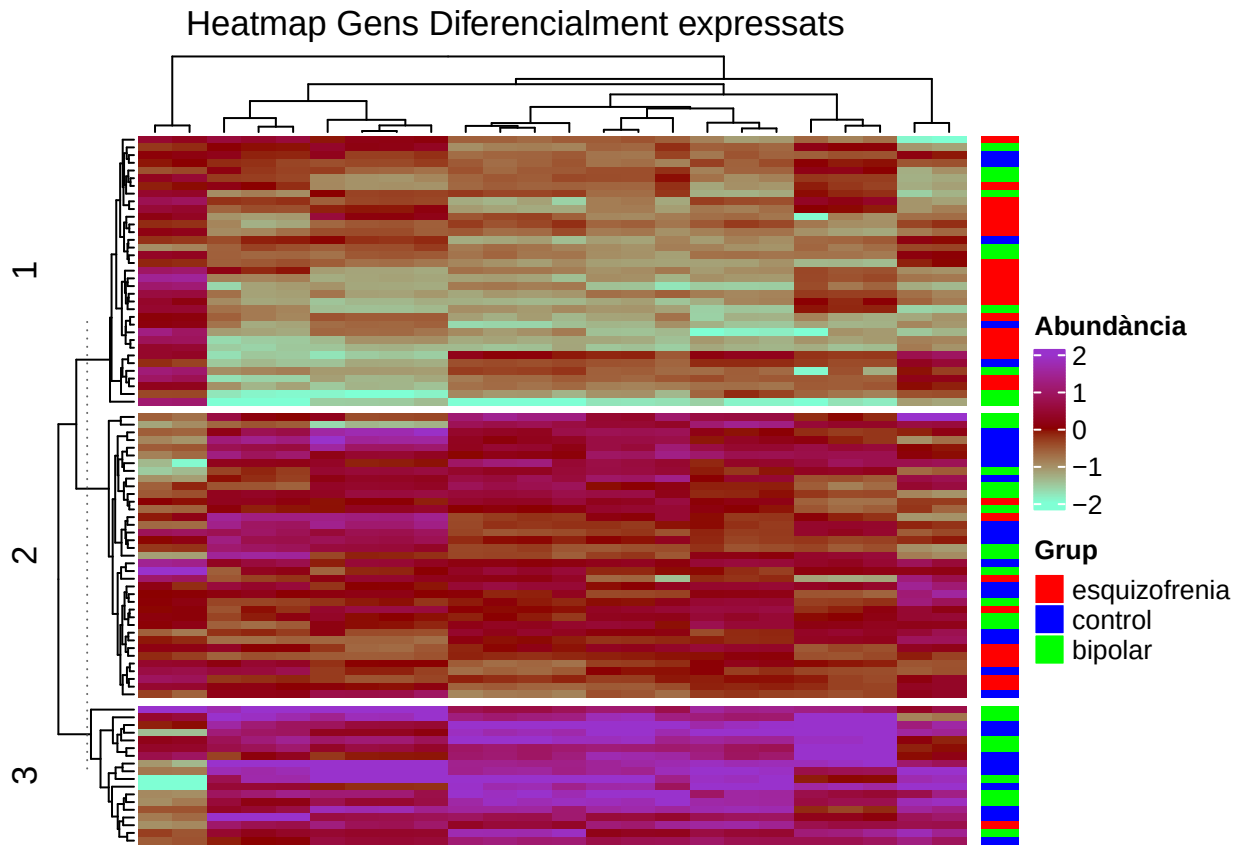




### *Clustering*

Una vegada triats els compostos amb diferències signitivatives i per acabar, obtinc un heatmap per vore si les mostres són agrupades de forma semblant als grups als quals pertanyen.

```
# Obtinc les dades dels gens expressats diferencialment.

se_diferent <- se_problema[rownames(se_problema) %in% c(rownames(ans_filtrat), rownames(ans_filtrat3)),
diferents <- scale(t(assay(se_diferent)))

# Extrac els grups i definisc els colors per a cadascun d'ells.

grups <- colData(se_diferent)$grup
grup_colors <- c("esquizofrenia" = "red", "control" = "blue", "bipolar" = "green")

# Genere una primera matriu limitant el nombre de clusters dels dendograma a 3,
# en tal d'observar si els grups inicials es reconstitueixen.

ht_expression <- Heatmap(diferents,
                         name = "Abundància",
                         col = colorRamp2(c(-2, 0, 2), c("aquamarine", "red4", "darkorchid")),
                         cluster_rows = TRUE,
                         row_km = 3,
                         cluster_columns = TRUE,
                         show_row_names = FALSE,
```

```
                          show_column_names = FALSE,
                          row_names_gp = gpar(fontsize = 8),
                          column_names_gp = gpar(fontsize = 8))

# Create a separate heatmap for the group annotations
ht_group <- Heatmap(as.matrix(grups),
                    name = "Grup",
                    col = grup_colors,
                    width = unit(5, "mm"),
                    show_row_names = FALSE,
                    show_column_names = FALSE,
                    cluster_rows = FALSE,
                    cluster_columns = FALSE)

# Concatenate the heatmaps

draw(ht_expression + ht_group, heatmap_legend_side = "right",
     annotation_legend_side = "right",
     column_title = "Heatmap Gens Diferencialment expressats")
```




Resulta evident que els perfils definits pels lípids per als quals s'han trobat diferències significatives són encavalcants entre esquizofrènics, bipolars, i controls. Així i tot, no sembla destrellatat apuntar que dos dels grups semblen enriquits en mostres de controls-bipolars (a l'inferior a la gràfica), mentre que el romanent ho està en mostres de pacients amb esquizofrènia (grup superior).

## Conclusions i apunts finals

Pareix clar que el lipidoma dels pacients amb esquizofrènia en plasma difereix del lipidoma dels controls i els pacients amb trastorn bipolar. Amb el procés de filtratge i anàlisi elaborada, s'ha arribat a una llista de gens que semblen tindre diferències significatives entre els grups considerats.

A partir d'aquests resultats, seria interessant fer una anàlisi sobre les vies metabòliques implicades a aquests compostos, i així intentar entendre cap a quines vies metabòliques apunten les diferències trobades. També seria interessant intentar generar un model predictiu a partir d'aquests gens en tal d'intentar classificar entre esquizofrènics-no esquizofrènics a partir del perfil lipidòmic. Aquests plantejaments queden fora de l'abast de l'anàlisi present, i es deixaran al calaix d'idees a testar.

## Fonts:


1. Costa, A. C. *et al.* Application of Lipidomics in Psychiatry: Plasma-Based Potential Biomarkers in Schizophrenia and Bipolar Disorder. *Metabolites* **13**, 600 (2023).
2. Wulff, J. E., & Mitchell, M. W. A comparison of various normalization methods for LC/MS metabolomics data. *Advances in Bioscience and Biotechnology* **9**, e98022 (2018).
3. Chapter 4. *A List of Heatmaps. ComplexHeatmap Complete Reference.* Available at: https://jokergoo.github.io/ComplexHeatmap-reference/book/a-list-of-heatmaps.html. Accessed: 2024-11-06.