# PAC Joan MArtinez Sancho

**02/04/25**

# Contents

# 1. Introducció

Per a la resolució d'aquesta pràctica s'ha seleccionat el dataset a partir de la pàgina web **metabolomicsWorkbench**. El dataset seleccionat correspon a l'experiment amb codi **ST000688**, en aquest cas en concret treballarem amb les dades que corresponen a **MS: C18 NEGATIVE ION MODE**.

L'experiment en qüestió es fonamenta sobre la idea de que el metabolisme està implicat en la neurodegeneració, d'aquesta manera l'anàlisis de metabolòmica permetrà definir la relació entre l'expresió gènica, els metabòlits i el dany axonal en l'esclerosis múltiple progressiva.

He seleccionat aquest experiment perquè hi ha un grup de l'institut de recerca d'on treballo (IDIBGI) que estudia l'esclerosis múltiple i últimament he treballat amb ells.

## 1.1. Objectius

L'objectiu principal de la pràctica és seleccionar un dataset de *metabolòmica*, crear un objecte de classe **SummarizedExperiment** i realitzar un anàlisis exploratori d'aquest dataset.

L'objectiu secundari és intentar replicar alguns dels resultats de l'anàlisis que es va realitzar en el seu moment, com poden ser el gràfic d'abundància relativa o bé el VolcanoPlot.

# 2. Materials i Mètodes

Les dades amb les que es treballarà es descarregaran de la pàgina web de **metabolomicsWorkbench**, corresponen a l'experiment ST000688 i es tractaran les dades **C18 NEGATIVE ION MODE**. Es llegiràn les dades, les quals provenen en format txt i es crearà un objecte SummarizedExperiment, el qual contindrà les dades de l'experiment, informació sobre les covariables, i altres aspectes de l'experiment.

L'exploració es portarà a terme seguint les activitats que s'han realitzat durant el curs. D'aquesta manera consistirà en, l'anàlisis univariant de les dades, mitjançant boxplots per estudiar la forma general dels mateixos. Anàlisis multivariant mitjançant l'Anàlisis de Components Principals i Agrupament Geràquic, per a determinar si els grups que apareixen semblen relacionar-se amb les fonts de variabilitat de l'estudi. I finalment es realitzarà un boxplot de l'abundància relativa en logaritmes (replicant l'anàlisis de l'experiment) simulant la funció **Rlaplots** (no és operativa actualment, però es pot treure la informació del codi) i es realitzarà un VolcanoPlot per a determinar quins biomarcadors estan diferencialment expressats entre BMS i SPMS.

# 3. Resultats

## 3.1. Obtenció de dades

Podeu trobar el codi de la obtenció de dades en el següent [Github] (https://github.com/jmsancho18/Martinez-Sancho-Joan-PEC1.git). Aqui podeu veure per exemple la matriu de dades de metabòlits (no tota, nomes algunes columnes i files) i les seves metadades. Però també s'han descarregat les dades de les covariables i informació de l'experiment.

```
##         Benign MS01 Benign MS03 Benign MS05     SPMS01     SPMS12     SPMS20
## CL 66:1   100607.06    34187.88    24702.52   79696.26   22956.20   31691.21
## CL 74:3  3738277.20  2231944.50  1121579.16 2511926.36 1709678.52 1423563.96
## CL 76:5    66437.68    23355.53    20890.40   29452.91   26180.93   39259.81
## CL 78:5  1839780.91   484486.82   663194.40 1083133.83  610195.43  786445.19
```

```
##   rownames(metabolite_metadata) moverz_quant   ri ri_type pubchem_id inchi_key
## 1      CL 66:1; [M-2H](2-)@6.26           NA 6.26      NA         NA        NA
## 2      CL 74:3; [M-2H](2-)@6.99           NA 6.99      NA         NA        NA
## 3      CL 76:5; [M-2H](2-)@6.98           NA 6.98      NA         NA        NA
## 4      CL 78:5; [M-2H](2-)@6.51           NA 6.51      NA         NA        NA
##   kegg_id other_id     other_id_type
```

```
## 1      NA CL(66:1) LipidMaps Bulk ID
## 2      NA CL(74:3) LipidMaps Bulk ID
## 3      NA CL(76:5) LipidMaps Bulk ID
## 4      NA CL(78:5) LipidMaps Bulk ID
```

### 3.2. Creació SummarizedExperiment

```
se
```

```
## class: SummarizedExperiment
## dim: 163 33
## metadata(1): general_metadata
## assays(3): counts escalados log_abund
## rownames(163): CL 66:1 CL 74:3 ... plasmenyl-PE 42:5 plasmenyl-PE 42:6
## rowData names(9): rownames(metabolite_metadata) moverz_quant ...
##   other_id other_id_type
## colnames(33): Benign MS01 Benign MS02 ... SPMS19 SPMS20
## colData names(6): Subject Tipus MS ... Gender Race
```

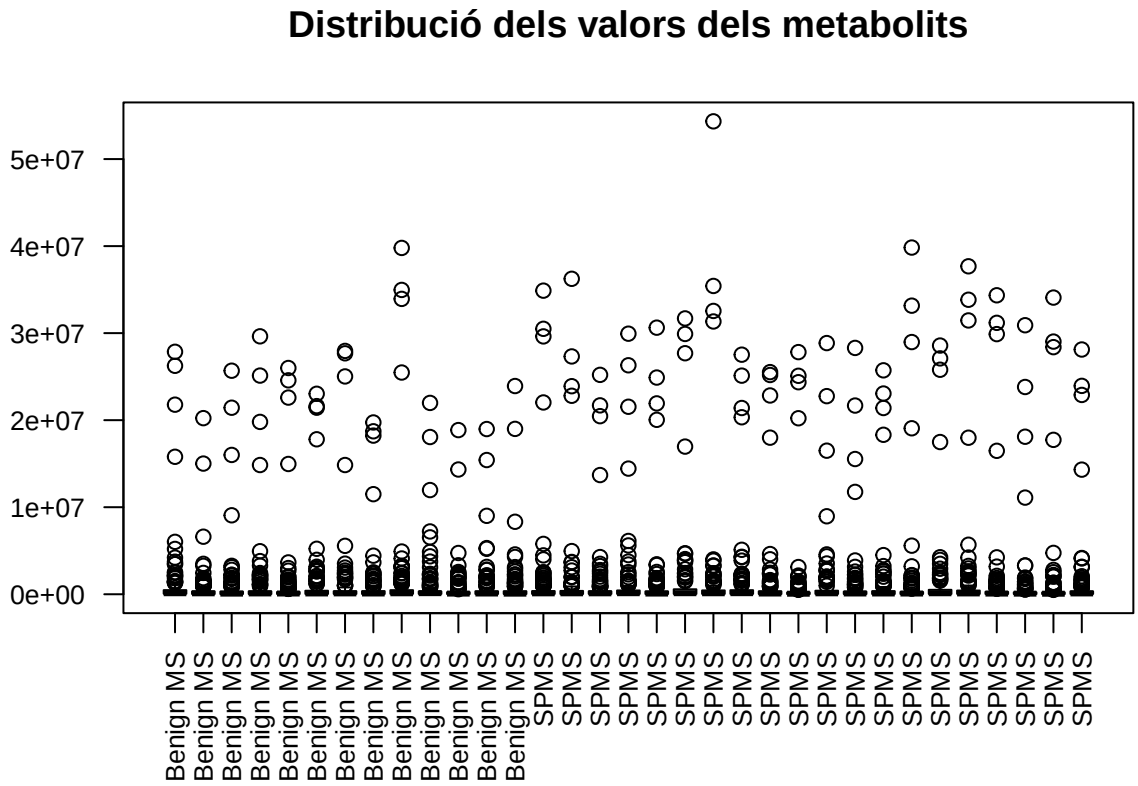
El nostre objecte **SummarizedExperiment** conté un conjunt de metadades (informació de l'experiment), 3 conjunts de dades (counts, escalados i log_abund), rowData és la informació de sobre els metabolits i colData són les covariables dels individus.

Els objectes **SummarizedExperiment** són molt similars en molts d'aspectes als **ExpressionSet**, una de les diferències és que SummarizedExperiment és més flexible en quant a la informació de les files, a més és compatiable amb GRanges cosa que facilita el tractament de dades genòmiques amb coordanades. La principal diferència però, la trobem en que SummarizedExperiment permet guardar múltiples matrius de dades a 'assays()'.

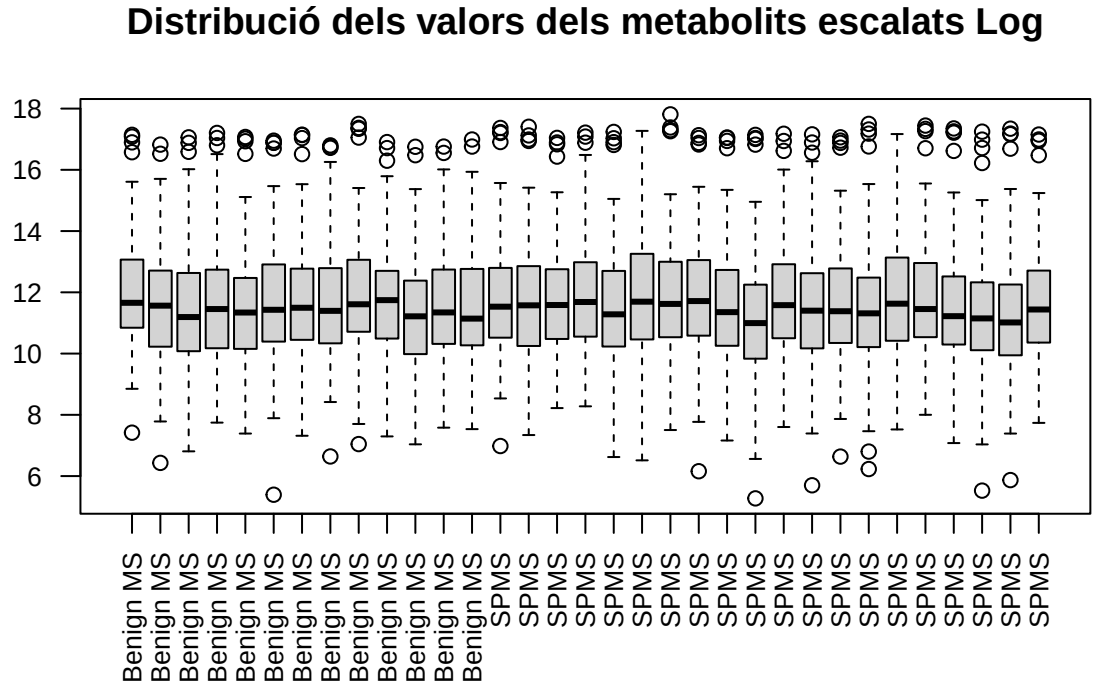
### 3.3. Anàlisis exploratori de dades

Una vegada fet l'extracció de les dades i la informació, podem procedir a realitzar una exploració bàsica.

```
boxplot(assays(se)$counts,
        las = 2,
        names = colData(se)$`Tipus MS`,
        cex.axis=0.8,
        main="Distribució dels valors dels metabolits"

)
```




Veiem que les dades són asimètriques, per tant provarem d'aplicar logaritmes.



Amb els logaritmes aplicats continuem veient una asimetría positiva, però queda bastant millor que sense escalar. Per tant treballaríem amb les dades escalades.

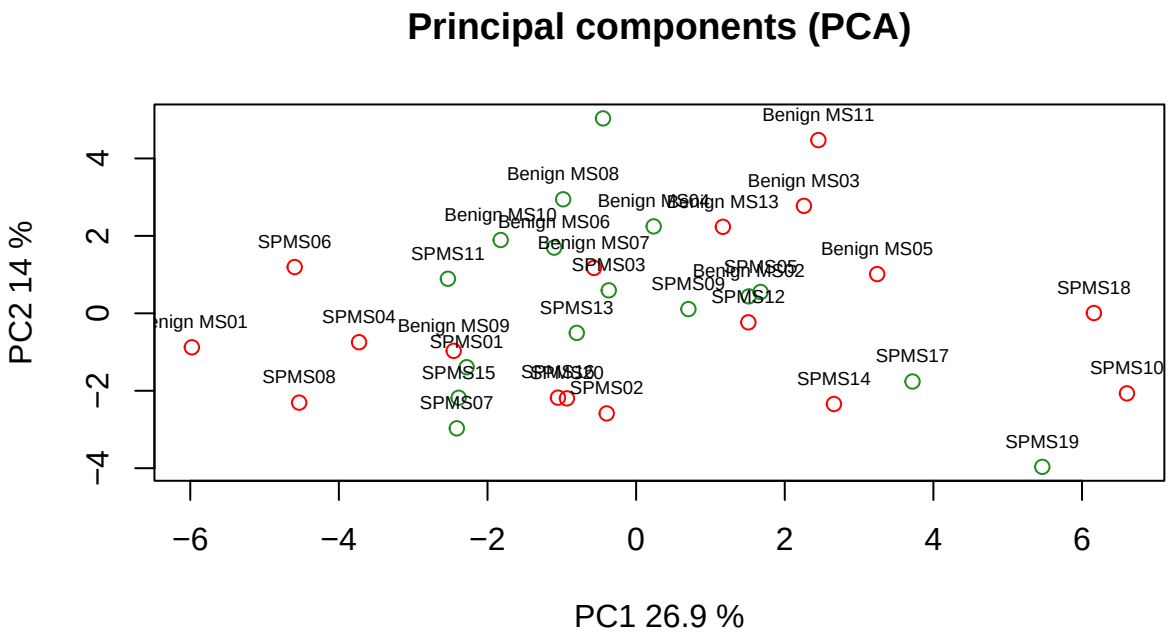
### 3.4. Anàlisis multivariant

En primer lloc realitzarem un anàlisis de components principals que ens faciliti la visualització de les dades per a detectar posibles patrons que no es puguin detectar a simple vista.

```
pcX<-prcomp(t(assays(se)$escalados), scale=FALSE) # Ja s'han escalat les dades
loads<- round(pcX$sdev^2/sum(pcX$sdev^2)*100,1)

summary(pcX)$importance[,1:4]
```

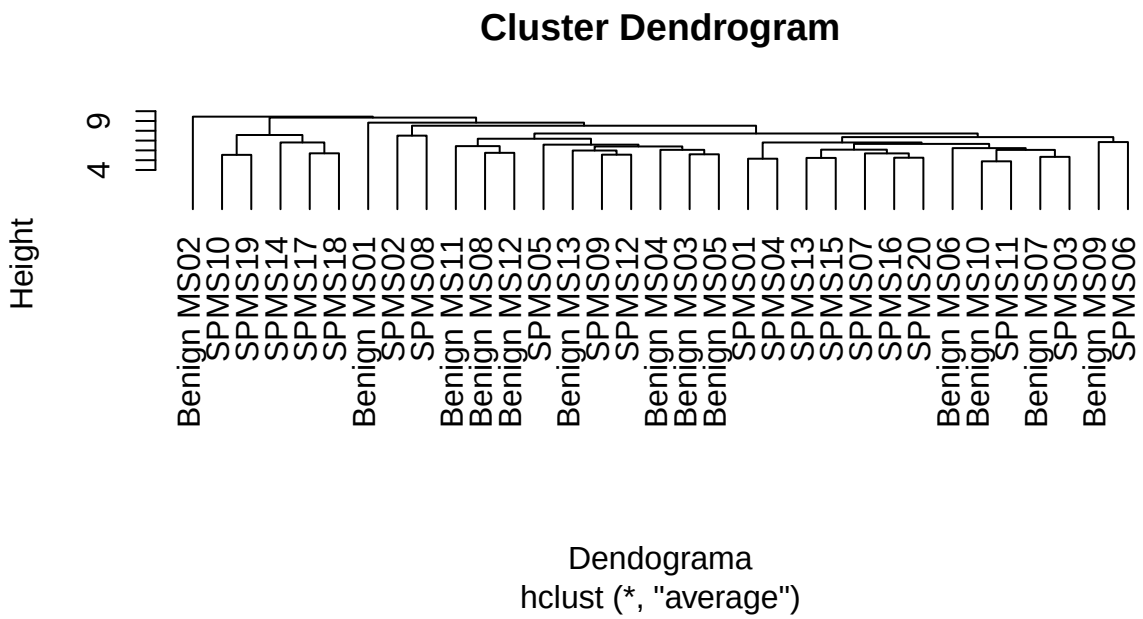
```
##                             PC1      PC2      PC3      PC4
## Standard deviation     3.024312 2.179251 1.738993 1.553039
## Proportion of Variance 0.269030 0.139690 0.088950 0.070940
## Cumulative Proportion  0.269030 0.408720 0.497670 0.568610
```

```
colores <- c("red", "forestgreen")

xlab<-c(paste("PC1",loads[1],"%"))
ylab<-c(paste("PC2",loads[2],"%"))
plot(pcX$x[,1:2],xlab=xlab,ylab=ylab, col=colores,
     main ="Principal components (PCA)")
names2plot<-colnames(se)

text(pcX$x[,1],pcX$x[,2],names2plot, pos=3, cex=.6)
```




Les dues primeres components expliquen el 41% de la variabilitat de les dades, ens inidica que les dues components principals no capturen la major part de l'estructura de les dades, ja que queda un 59% de variabilitat per explicar. Potser existeix un soroll en les dades que dificulta l'explicació de la variabilitat a partir de PCA.

```r
clust.euclid.average <- hclust(dist(t(assays(se)$escalados)),method="average")
plot(clust.euclid.average, hang=-1,,xlab="Dendograma")
```

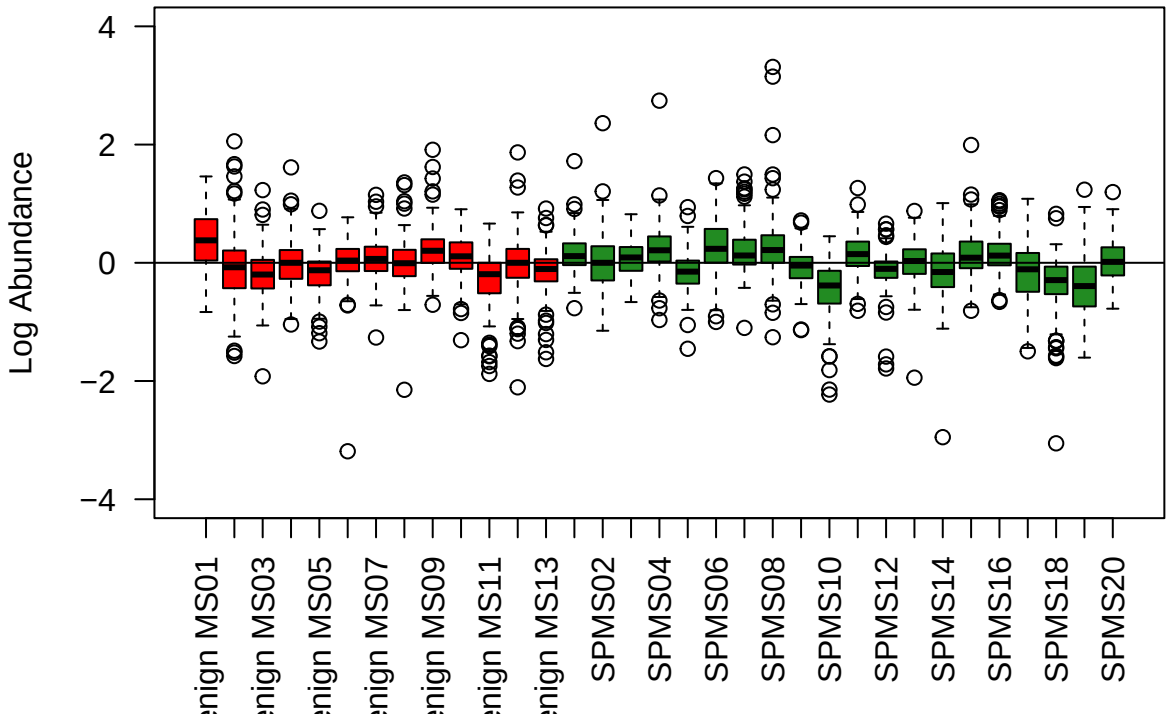



Si realitzem el clúster geràrquic i observem el dendograma, veiem que tot i haver-hi molts de subgrups, almenys els subjectes del mateix tipus si que estan pròxims entre ells.

### 3.5. Relative log abundance plots (across groups)

Generem el gràfic d'abundància logarítmica relativa, **entre grups**. A simple vista no veiem molta diferència en la distribució entre els 2 grups.

```r
library(ggplot2)
library(reshape2)

boxplot(assays(se)$log_abund,
        las = 2,
        col = colores[as.numeric(colData(se)$`Tipus MS`)],
        ylab = "Log Abundance",
        ylim = c(-4, 4)

)

# Añade la línea horizontal en y = 0
abline(h = 0, col = "black", lty = 1, lwd = 1)
```

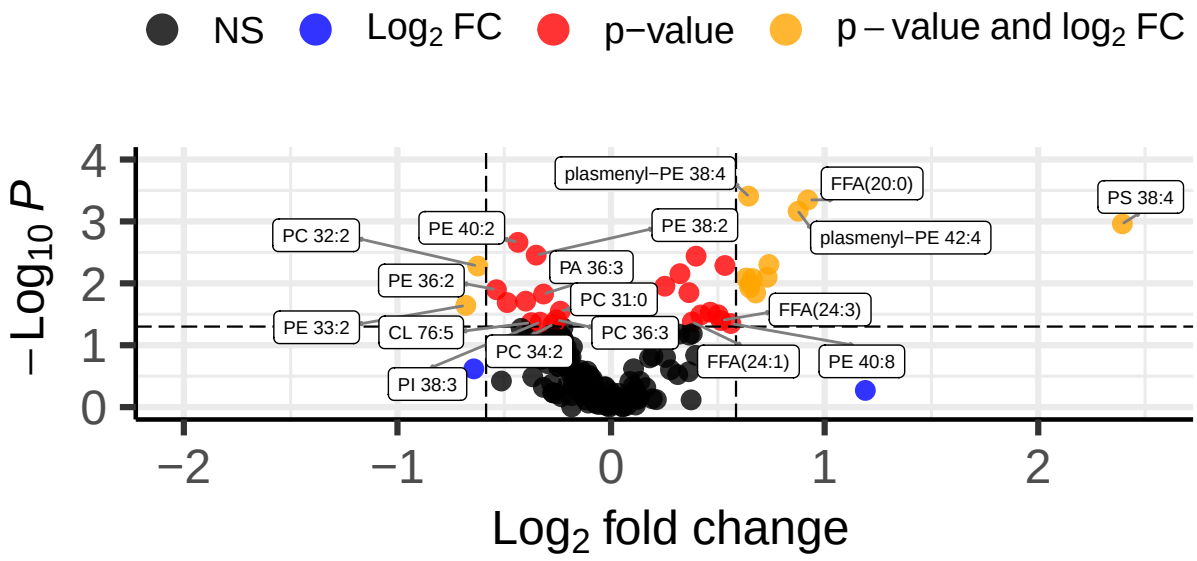



### 3.6. Comparació de metabolics entre grups

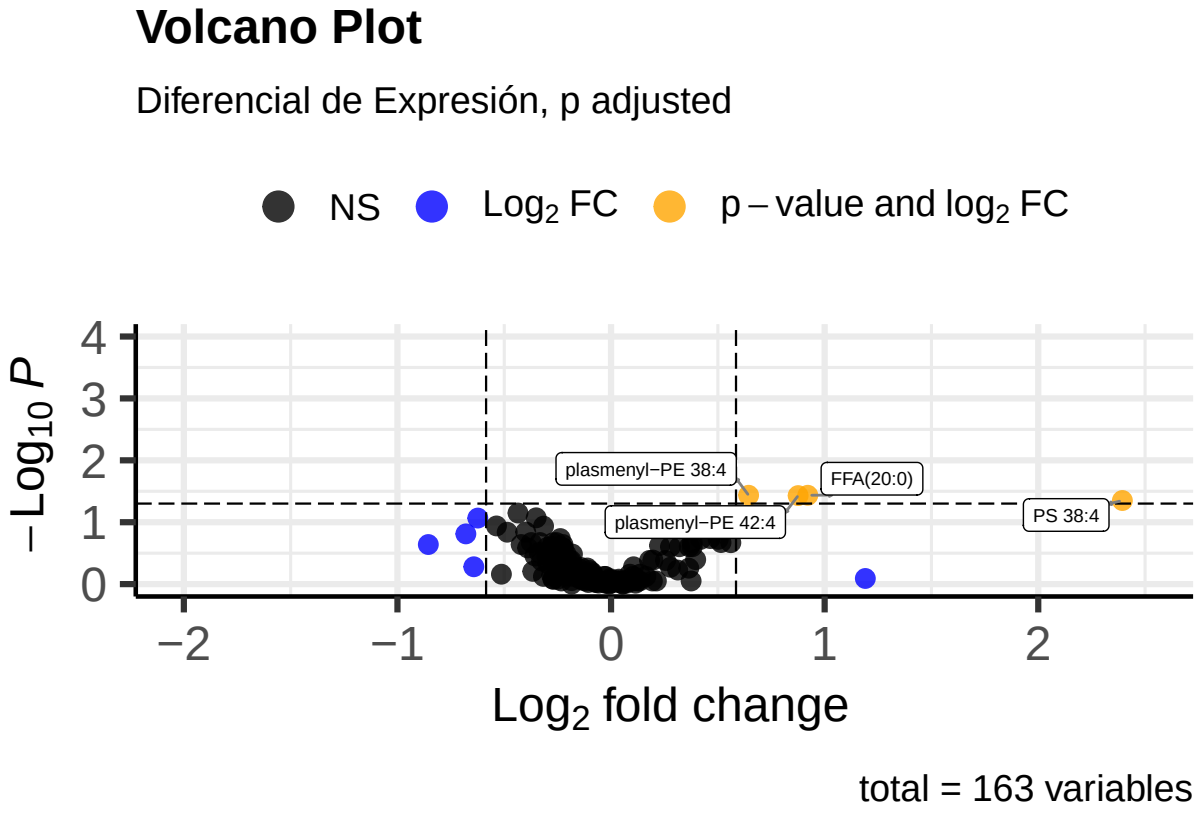
Comparem l'expressió dels diferents metabòlits en funció dels dos grups d'esclerosis múltiple. El primer gràfic surten molts de metabòlits significatius (n=35) però el pvalor no està ajustat per cap correció de comparacions múltiples. Un cop ajustem per bonferroni només trobem 4 metabòlits significatius (n=4).

## Volcano Plot

Diferencial de Expresión




total = 163 variables




## 4. Discussió/Conclussió

L'anàlisis exploratori realitzat ens ha posat de manifest que existeixen diverses fonts de variació. Part d'aquesta l'hem pogut corregir de certa manera aplicant logaritmes a les dades. També hem pogut veure que no hi ha evidents problemes en les dades, ja que les distribucions d'aquestes són similars entre elles i no hi ha valors perduts o zeros. Amb els gràfics i anàlisis multivariants que hem realitzat (Dendograma i PCA), no hem sigut capaços de trobar o distingir en precisió la variabilitat de les mostrar, ni de crear grups que en un primer moment podíem arribar a pensar que es crearíen en base al diagnòstic del subjecte (tipus d'esclerosis)

Donat que l'anàlisis que hem realitzat (volcano plot i abundància relativa) dóna resultats similars als que podem veure a l'anàlisis de l'estudi, podríem suposar que les dades no s'han tractat gaire més.

De cara a següents pasos es podrien estudiar els metabòlits que han sortit significatius i mirar en quins processos biològics estan implicats.

## 5. Referències

Enllaç a Github: https://github.com/jmsancho18/Martinez-Sancho-Joan-PEC1.git