Informe d'Anàlisi Metabolòmica

Anna Pérez Bosque – PAC1

Taula de Continguts

1. Abstract

L'objectiu d'aquest estudi és analitzar un conjunt de dades metabolòmiques relacionades amb la pèrdua muscular en pacients amb caquèxia, una condició caracteritzada per la pèrdua de massa muscular i de pes, sovint associada a malalties cròniques. Utilitzant tècniques estadístiques descriptives i una anàlisi de components principals (PCA), s'ha intentat comprendre les diferències entre els grups de pacients amb caquèxia i controls, així com identificar els metabòlits més rellevants que es podrien associar a la presència de caquèxia. Els resultats indiquen una clara diferenciació entre els grups, especialment en funció de determinats metabòlits. Aquest estudi contribueix a la identificació de possibles biomarcadors per a la caquèxia, amb potencial per orientar futures investigacions clíniques.

2. Objectius

L'objectiu principal d'aquest estudi és investigar la relació entre els nivells de metabòlits en pacients amb i sense caquèxia, ja que la identificació de biomarcadors específics podria millorar la diagnosi i el tractament d'aquesta condició debilitant. Els objectius específics són:

- Comparar els nivells de metabòlits entre pacients caquèctics i controls per identificar diferències significatives.
- Utilitzar tècniques d'anàlisi multivariant, com la PCA, per explorar possibles patrons en els nivells de metabòlits.
- Proposar metabòlits que puguin ser indicadors potencials de la caquèxia.

3. Materials i mètodes

3.1. Conjunt de dades

En aquest estudi s'ha utilitzat el conjunt de dades "human_cachexia.csv", que conté nivells de diversos metabòlits en 77 pacients, alguns dels quals pateixen caquèxia. La variable 'Muscle.loss' indica la pèrdua muscular, classificant els pacients en grups 'cachexic' i 'control'. Aquesta variable es considera essencial per entendre les diferències en el metabolisme dels pacients.

3.2. Preparació de les dades

Les dades s'han transformat perquè cada columna representi una mostra i cada fila un metabòlit, format necessari per a l'ús amb *SummarizedExperiment*. Aquesta estructura facilita l'anàlisi multivariant i la comparació entre grups.

3.3. Anàlisi estadística

S'han aplicat les següents tècniques d'anàlisi:

- Estadística descriptiva: Per oferir una visió inicial de les distribucions de cada metabòlit.
- Anàlisi de Components Principals (PCA): S'utilitza per reduir la dimensionalitat i examinar la variabilitat dels metabòlits entre grups, una tècnica adequada per identificar patrons en dades metabòliques complexes.
- Anàlisi de Clústers: Tant k-means com clúster jeràrquic s'han aplicat per explorar possibles agrupacions de pacients segons els seus perfils metabòlics.

4. Resultats

4.1. Carregar dades i comprovació del dataset

Per començar l'anàlisi, es va carregar el dataset "human_cachexia.csv" a R i es van generar metadades a partir de les columnes de pacient i pèrdua muscular. L'objectiu era crear un contenidor SummarizedExperiment que organitzés les dades metabolòmiques i les metadades, assegurant així una estructura de dades adequada per a les anàlisis posteriors.

A continuació es mostra el codi utilitzat per carregar les dades i verificar la consistència de les dimensions entre les dades metabolòmiques i les metadades. La verificació de dimensions és essencial per garantir que cada mostra tingui metadades associades, la qual cosa assegura la fiabilitat del procés d'anàlisi.

```
library(SummarizedExperiment)

# Carregar el dataset
dataset <- read.csv("human_cachexia.csv", header = TRUE)

# Crear el DataFrame de metadades
metadades <- DataFrame(
  PatientID = dataset$Patient.ID,
  MuscleLoss = dataset$Muscle.loss
)

# Convertir les dades dels metabòlits en una matriu
data <- as.matrix(dataset[, -c(1, 2)])  # Exclou les columnes 1 i 2
data <- t(data)  # Transposa la matriu per a que les mostres siguin columnes

dim(metadades)  # Ha de coincidir amb el nombre de columnes de data

## [1] 77  2

dim(data)       # Comprova el nombre de files i columnes

## [1] 63 77

# Crea l'objecte SummarizedExperiment
se <- SummarizedExperiment(assays = list(counts = data), colData = metadades)
```

4.2. Exploració descriptiva de les dades

Per a cada grup (cachexic i control), es va realitzar una estadística descriptiva, incloent la mitjana i la desviació estàndard dels metabòlits, com es mostra a la Taula 1. Aquesta anàlisi descriptiva inicial permet observar la distribució dels valors dels metabòlits i detectar possibles patrons o anomalies.

```r
library(dplyr)
library(tidyr)

# Obtenir els noms dels metabòlits
metabolites <- rownames(assay(se))

# Inicialitzar un dataframe per guardar els resultats de la mitjana i la desviació estàndard
summary_results <- data.frame(metabolite = character(), group = character(), mean = numeric(), sd = numeric(), stringsAsFactors = FALSE)

# Obtenir els valors per a cada metabòlit i calcular les mitjanes i desviacions estàndards per a cada grup (cachexic vs control)
for (metabolite in metabolites) {
  values <- assay(se)[metabolite, ]
  group <- colData(se)$MuscleLoss

  # Calcular la mitjana i desviació estàndard per a cada grup
  mean_cachexic <- round(mean(values[group == "cachexic"], na.rm = TRUE),2)
  sd_cachexic <- round(sd(values[group == "cachexic"], na.rm = TRUE),2)
  mean_control <- round(mean(values[group == "control"], na.rm = TRUE),2)
  sd_control <- round(sd(values[group == "control"], na.rm = TRUE),2)

  # Emmagatzemar els resultats al dataframe
  summary_results <- rbind(summary_results,
                           data.frame(metabolite = metabolite, group = "cachexic", mean = mean_cachexic, sd = sd_cachexic),
                           data.frame(metabolite = metabolite, group = "control", mean = mean_control, sd = sd_control))
}

# Convertir el dataframe a un format ampli per mostrar la mitjana i la desviació estàndard per cada grup en columnes separades
summary_table <- summary_results %>%
  pivot_wider(names_from = group, values_from = c(mean, sd))

# Mostrar la taula resum
print(summary_table)
```

```
## # A tibble: 63 × 5
##    metabolite                  mean_cachexic mean_control sd_cachexic
sd_control
##    <chr>                              <dbl>        <dbl>       <dbl>
<dbl>
##  1 X1.6.Anhydro.beta.D.glucose        129.         69.5        142.       99.6
##  2 X1.Methylnicotinamide               70.6        73.2         85.3     187.
##  3 X2.Aminobutyrate                    23.7         9.53        33.9       7.31
##  4 X2.Hydroxyisobutyrate               43.2        27.9         23.4      22.0
##  5 X2.Oxoglutarate                    183.         85.5        412.      180.
##  6 X3.Aminoisobutyrate                100.         39.9        238.       57.8
```

```
##  7 X3.Hydroxybutyrate                  29.3         9.9         30.7       7.99
##  8 X3.Hydroxyisovalerate               27.6        12.3         27.4      17
##  9 X3.Indoxylsulfate                  265.        146.         212.      147.
## 10 X4.Hydroxyphenylacetate            120.         99.8         94.1     155.
## # i 53 more rows
```

Taula 1. Resum estadístic dels nivells mitjans i desviacions estàndard per a cadascun dels metabòlits en els grups 'cachexic' i 'control'. Destaca que els metabòlits com la creatinina i el citrat tenen mitjanes significativament diferents entre grups, la qual cosa suggereix que podrien estar relacionats amb la pèrdua muscular observada en la caquèxia. Aquesta observació aporta indicis preliminars sobre el seu potencial com a biomarcadors.

4.3. Anàlisi de Components Principals (PCA)

L'anàlisi de components principals (PCA) es va utilitzar per reduir la dimensionalitat de les dades i explorar la variabilitat global entre els pacients. La **Figura 1** mostra un gràfic de PCA amb els dos primers components principals, que expliquen un percentatge significatiu de la variabilitat total (PC1: 40.4%, PC2: 8.2%). Tot i que encara queda prop d'un 50% sense explicar-se.

La visualització del PCA mostra una separació parcial entre els grups caquèxia i control, amb els pacients amb caquèxia tendint a agrupar-se en una regió específica del gràfic. Aquesta distribució suggereix que les variacions en alguns metabòlits són rellevants per diferenciar entre pacients amb i sense caquèxia.

La **Figura 2** presenta els metabòlits amb les contribucions més altes en els components PC1 i PC2, mostrant quins d'ells tenen una influència significativa en la separació observada. Metabòlits com l'acetat, l'iydroxiisovalerat, l'acetona, i el fumarat tenen les contribucions més altes a PC1 i/o PC2, indicant que podrien ser rellevants per a la diferenciació entre pacients. Els metabòlits que més influeixen en la variabilitat de PC1 o PC2 estan detallats en la **Taula 2**. Aquests metabòlits inclouen aminobutirat, hidroxisovalerat, acetat, acetona, fumarat, metilguanidina, acetilcarnitina, i tartrat. La seva contribució indica que poden estar implicats en la variabilitat associada a la caquèxia, mostrant canvis metabòlics significatius entre pacients amb i sense caquèxia.

```r
library(factoextra)
library(ggplot2)

# Aplicar l'anàlisi de PCA
pca <- prcomp(t(assay(se)), center = TRUE, scale. = TRUE)

# Crear un vector de colors segons el grup (cachexic vs control)
group_labels <- colData(se)$MuscleLoss
group_colors <- ifelse(group_labels == "cachexic", "#c60000", "#08ee47")
```

```r
# Visualitzar el PCA amb factoextra (components principals 1 i 2)
fviz_pca_ind(pca, geom = "point", col.ind = group_labels,
             palette = c("#c60000", "#08ee47"),
             addEllipses = TRUE, ellipse.level = 0.95,
             legend.title = "Grup",
             title = "Anàlisi de Components Principals (PCA)") +
  labs(x = "PC1 (40.4%)", y = "PC2 (8.2%)") +
  theme_minimal()
```

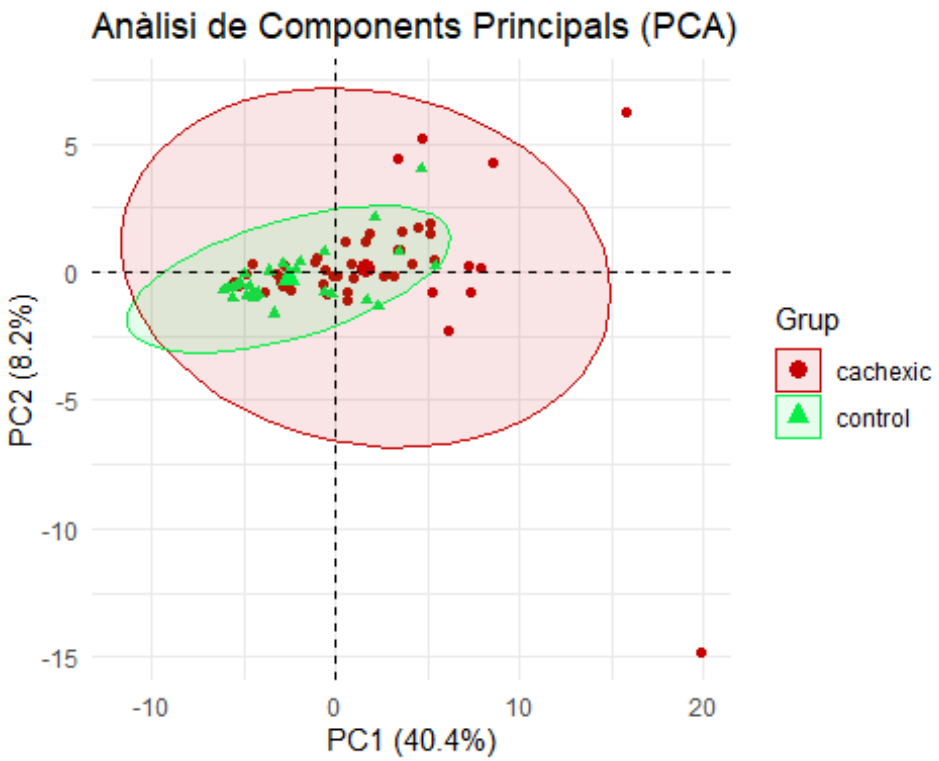

Figura 1. Anàlisi de PCA amb els dos primers components principals. Aquests 2 components principals expliquen prop d'un 50% de la variabilitat total (PC1: 40.4%, PC2: 8.2%). En vermell estan representats els pacients amb caquèxia i en verd el pacients control.

```r
# Calcular la contribució de cada metabòlit a les components principals
contrib_pca <- as.data.frame(pca$rotation)
contrib_pca$metabolite <- rownames(contrib_pca)

# Filtrar només els metabòlits amb contribució significativa a PC1 o PC2 (superior a 0.2)
threshold <- 0.2
significant_metabolites <- contrib_pca %>%
  select(metabolite, PC1, PC2) %>%  # Seleccionem només les columnes de PC1 i PC2
  filter(abs(PC1) > threshold | abs(PC2) > threshold) %>%  # Filtre per contribució significativa
  mutate(across(starts_with("PC"), ~ round(.x, 3)))  # Arrodonir a 3 decimals
```

```
# Visualitzar el PCA de contribució dels metabòlits amb només els més
significatius
fviz_pca_var(pca, col.var = "contrib",
             select.var = list(name = significant_metabolites$metabolite),
             gradient.cols = c("#2697f5", "#ffd733", "#ff6b33"),
             repel = TRUE, title = "Contribució dels Metabòlits") +
  labs(x = "PC1 (40.4%)", y = "PC2 (8.2%)") +
  theme_minimal()
```

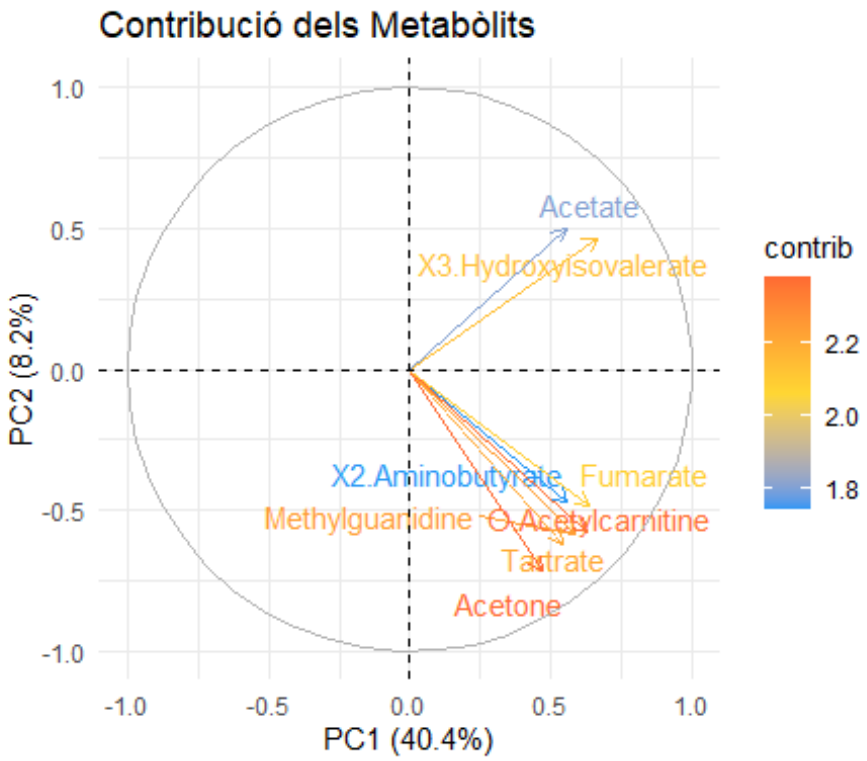

Figura 2. Gràfic de contribució dels metabòlits a les components principals (PC1 i PC2) en l'anàlisi de PCA. La direcció i llargada dels vectors indiquen la importància i el sentit de cada metabòlit en la distribució de les dades. Els compostos amb coloració més taronja són els que presenten una contribució major.

```
# Visualitzar la taula amb els metabòlits significatius (contribució > 0.2 en PC1
o PC2)
print(significant_metabolites)
```

```
##                                  metabolite   PC1    PC2
## X2.Aminobutyrate           X2.Aminobutyrate 0.111 -0.208
## X3.Hydroxyisovalerate X3.Hydroxyisovalerate 0.131  0.205
## Acetate                             Acetate 0.110  0.219
## Acetone                             Acetone 0.092 -0.315
## Fumarate                           Fumarate 0.126 -0.214
## Methylguanidine             Methylguanidine 0.116 -0.258
```

```
## O.Acetylcarnitine         O.Acetylcarnitine 0.124 -0.253
## Tartrate                           Tartrate 0.108 -0.272
```

Taula 2. Metabòlits amb contribucions significatives a PC1 o PC2. Els valors indiquen el pes de cada metabòlit en cadascuna de les components, ajudant a identificar quins metabòlits tenen un paper més destacat en la variabilitat entre pacients.

4.4. Anàlisi de Clústers

Per tal de confirmar la separació observada en l'anàlisi PCA i explorar possibles agrupacions naturals dels pacients, es va aplicar una anàlisi de clúster, tant jeràrquica com amb k-means.

El dendrograma, obtingut mitjançant l'anàlisi de clúster jeràrquic, mostra les relacions de similitud entre pacients (**Figura 3**). Els pacients caquèctics es troben agrupats en branques específiques, fet que suggereix un perfil metabolòmic distintiu en aquest grup. Les diferents distàncies entre els pacients indiquen variabilitat dins de cada grup, però amb una tendència clara cap a una separació. També es pot observar que si vé en general els pacients Control presenten una branca força separada dels pacients amb caquèxia, hi ha un cert grup de pacients Control que es localitzen entremig, i no tant allunyats dels que presenten caquèxia.

El gràfic de clúster k-means (amb dos grups) complementa el dendrograma i ofereix una visualització més específica dels grups 'cachexic' i 'control' (**Figura 4**). En aquest cas, els pacients han estat assignats automàticament a un dels dos grups en funció de les seves similituds metabolòmiques. Tot i que es produeix una certa superposició entre els grups, els pacients amb caquèxia tendeixen a agrupar-se de manera coherent, cosa que reafirma la diferenciació identificada en l'anàlisi de PCA.

```r
library(ggplot2)
library(dendextend)

# Matriu de distàncies euclidianes
dist_matrix <- dist(t(assay(se)))

# Clúster jeràrquic amb mètode d'enllaç complet
hc <- hclust(dist_matrix, method = "complete")
dend <- as.dendrogram(hc)

colors <- ifelse(colData(se)$MuscleLoss == "cachexic", "#c60000", "#08ee47")

# Assignar els colors a les etiquetes del dendrograma
labels_colors(dend) <- colors

# Visualitzar el dendrograma amb els colors aplicats
plot(dend, main = "Dendrograma de les mostres",
     ylab = "Distància", xlab = "Mostres")
```

```
legend("topright", legend = c("Cachexic", "Control"),
       fill = c("#c60000", "#08ee47"), bty = "n", title = "Grups")
```

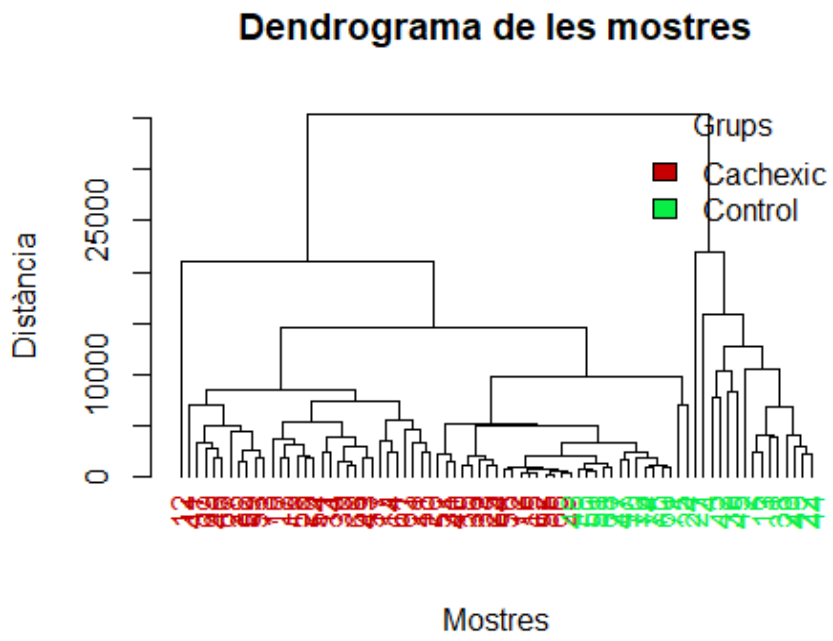

Figura 3. Dendrograma de les mostres. El dendrograma mostra les relacions de similitud entre pacients. Els pacients que presenten caquèxia estan representats en vermell i els pacients Control en verd.

```
# Aplicar k-means amb 2 grups (cachexic i control)
library(factoextra)
library(ggplot2)

# Aplicar k-means amb 2 grups (cachexic i control)
set.seed(123)
kmeans_result <- kmeans(t(assay(se)), centers = 2)

# Crear un vector amb la informació dels grups coneguts (cachexic vs control)
group_labels <- colData(se)$MuscleLoss

# Visualitzar els resultats del k-means amb fviz_cluster
fviz_plot <- fviz_cluster(kmeans_result, data = t(assay(se)), geom = "point",
ellipse.type = "convex",
                          main = "Clúster k-means de les mostres",
                          palette = c("#ff6b33", "#2697f5"),
                          ggtheme = theme_minimal(), show.clust.cent = FALSE)

# Convertir el plot a un objecte ggplot per a una personalització posterior
```

```
fviz_data <- fviz_plot$data
fviz_data$group <- group_labels

# Modificar el plot per afegir formes plenes o buides segons el grup conegut
final_plot <- ggplot() +
  # Dibuixar les el·lipses (convex hulls) només amb el contorn
  stat_ellipse(data = fviz_data, aes(x = x, y = y, group = cluster, color = factor(cluster)),
               geom = "polygon", alpha = 1, fill = NA) +  # fill = NA per deixar sense farciment
  # Afegir els punts amb formes segons el grup conegut
  geom_point(data = fviz_data, aes(x = x, y = y, color = factor(cluster), shape = group), size = 3) +
  scale_color_manual(values = c("1" = "#ff6b33", "2" = "#2697f5")) +
  scale_shape_manual(values = c("cachexic" = 16, "control" = 1)) +  # Ple per cachexic, buit per control
  labs(title = "Clúster k-means de les mostres",
       x = "PC1", y = "PC2", color = "Clúster", shape = "Grup") +
  theme_minimal() +
  theme(legend.position = "right")

# Mostrar el gràfic final
print(final_plot)
```

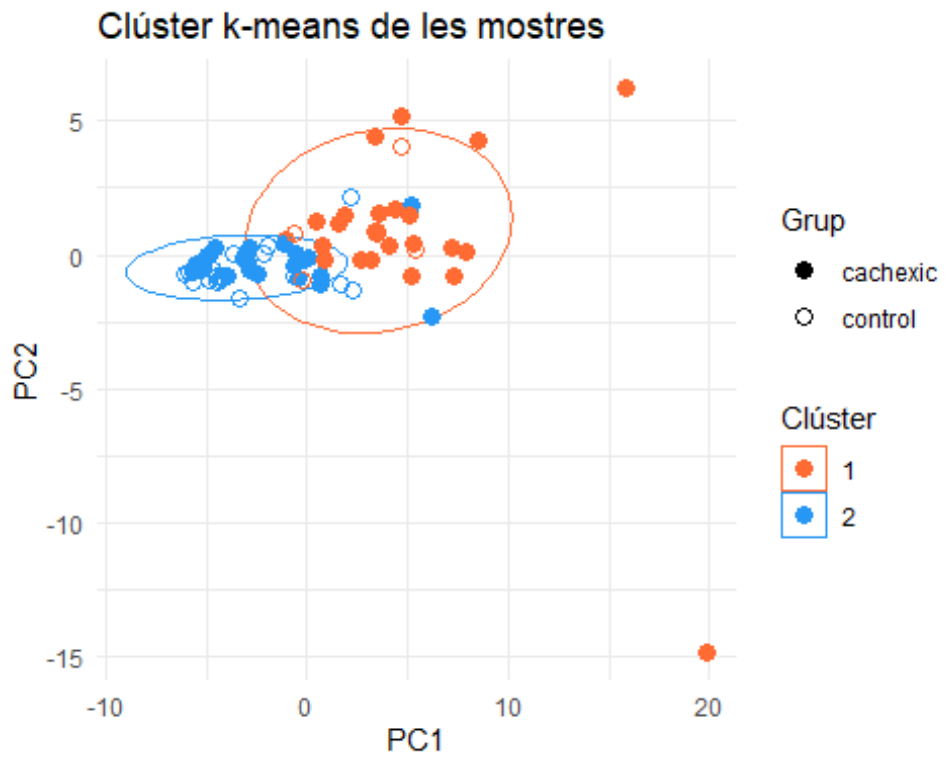

Figura 4. Clúster k-means de les mostres. Els pacients s'han agrupat en dos clusters (taronja i blau) en funció del seu metaboloma. Els punts sòlids són els pacients que presenten caquèxia(cachexic) i els punts buits són els pacients control.

4.5 Anàlisi estadística de diferències entre grups

Per verificar estadísticament les diferències en els nivells de metabòlits entre els grups, es va realitzar el test de Mann-Whitney U per a cada metabòlit. Els resultats es mostren a la Taula 2, on es llisten els metabòlits amb valors de p significatius.

```
# Inicialitzar un dataframe per guardar els resultats del Mann-Whitney U Test
mw_test_results <- data.frame(metabolite = character(), p_value = numeric(),
stringsAsFactors = FALSE)

# Realitzar el Mann-Whitney U Test per a cada metabòlit
for (metabolite in metabolites) {
  # Obtenir els valors del metabòlit per a cada grup (cachexic vs control)
  values <- assay(se)[metabolite, ]
  group <- colData(se)$MuscleLoss

  # Realitzar el Mann-Whitney U Test amb exact = FALSE per evitar l'avís
  mw_test <- wilcox.test(values ~ group, exact = FALSE)

  # Emmagatzemar el nom del metabòlit i el p-valor arrodonit a 3 decimals en el
dataframe
  mw_test_results <- rbind(mw_test_results,
                           data.frame(metabolite = metabolite, p_value =
round(mw_test$p.value, 3)))
}

# Visualitzar els resultats
print(mw_test_results)
```

```
##                     metabolite p_value
## 1  X1.6.Anhydro.beta.D.glucose   0.028
## 2        X1.Methylnicotinamide   0.034
## 3             X2.Aminobutyrate   0.004
## 4        X2.Hydroxyisobutyrate   0.004
## 5              X2.Oxoglutarate   0.045
## 6          X3.Aminoisobutyrate   0.170
## 7           X3.Hydroxybutyrate   0.000
## 8        X3.Hydroxyisovalerate   0.000
## 9            X3.Indoxylsulfate   0.001
## 10     X4.Hydroxyphenylacetate   0.014
## 11                     Acetate   0.000
## 12                     Acetone   0.541
## 13                     Adipate   0.000
## 14                     Alanine   0.000
## 15                  Asparagine   0.001
## 16                     Betaine   0.000
## 17                   Carnitine   0.032
## 18                     Citrate   0.007
## 19                    Creatine   0.000
## 20                  Creatinine   0.001
```

```
## 21                Dimethylamine   0.000
## 22                Ethanolamine   0.007
## 23                     Formate   0.000
## 24                      Fucose   0.005
## 25                    Fumarate   0.002
## 26                     Glucose   0.000
## 27                   Glutamine   0.000
## 28                     Glycine   0.026
## 29                   Glycolate   0.028
## 30              Guanidoacetate   0.020
## 31                   Hippurate   0.004
## 32                   Histidine   0.002
## 33                Hypoxanthine   0.122
## 34                  Isoleucine   0.066
## 35                     Lactate   0.002
## 36                     Leucine   0.000
## 37                      Lysine   0.001
## 38                 Methylamine   0.000
## 39             Methylguanidine   0.311
## 40         N.N.Dimethylglycine   0.000
## 41           O.Acetylcarnitine   0.019
## 42                Pantothenate   0.067
## 43              Pyroglutamate   0.000
## 44                   Pyruvate   0.003
## 45                 Quinolinate   0.000
## 46                      Serine   0.001
## 47                   Succinate   0.000
## 48                     Sucrose   0.000
## 49                    Tartrate   0.100
## 50                     Taurine   0.034
## 51                   Threonine   0.001
## 52                Trigonelline   0.008
## 53      Trimethylamine.N.oxide   0.009
## 54                  Tryptophan   0.000
## 55                     Tyrosine   0.001
## 56                      Uracil   0.337
## 57                      Valine   0.000
## 58                      Xylose   0.000
## 59                cis.Aconitate   0.000
## 60                myo.Inositol   0.000
## 61             trans.Aconitate   0.001
## 62          pi.Methylhistidine   0.019
## 63         tau.Methylhistidine   0.008
```

Taula 2. Resultats del test de Mann-Whitney U per a cada metabòlit.Els valors de p inferiors a 0.05 indiquen diferències estadísticament significatives entre els grups ‘cachexic’ i ‘control’ per a aquests metabòlits.

5. Discussió i conclusions

Els resultats d'aquest estudi suggereixen que diversos metabòlits tenen una expressió diferent entre els pacients amb i sense caquèxia. La PCA i l'anàlisi de clústers han revelat patrons que diferencien clarament els dos grups, amb metabòlits com l'acetat, l'acetona, i el fumarat mostrant variacions significatives. Aquests canvis podrien estar relacionats amb la disfunció de vies metabòliques importants en el context de la caquèxia.

Tot i això, hi ha uns pacients Control que no semblen acabar de separar-se dels pacients amb caquèxia, potser si hi ha altres factors que ajudarien a diferenciar-los i explicarien una variabilitat que amb les dades que hi ha a l'estudi no s'acaba d'explicar. Potser dades clíniques o d'hàbits aportarien més informació a la variabilitat dels sobretot dels pacients Control.

6. Limitacions

L'estudi presenta certes limitacions com ara la mida limitada de la mostra i la variabilitat en les dades, que poden dificultar la generalització dels resultats. A més, la manca d'informació clínica detallada dels pacients, com els hàbits alimentaris o l'activitat física, podrien haver influït en els nivells de metabòlits.

En futurs estudis, es podria incorporar informació clínica addicional per entendre millor els factors que contribueixen a la caquèxia. A més, un conjunt de dades més gran permetria una millor robustesa dels resultats i una anàlisi més detallada de les vies metabòliques implicades en la malaltia. Finalment, experiments addicionals per confirmar la validesa dels biomarcadors identificats serien necessaris per tal de transferir aquests resultats a la pràctica clínica.

Enllaç al repositori: https://github.com/aperezbos/Perez-Bosque-Anna-PAC1.git