PEC 1. Anàlisi de dades de metabolòmica

Sara Guzmán Mairén

2025-03-31

Contents

Abstract

Aquest informe Aquest estudi explora les alteracions metabòliques associades a la caquèxia mitjançant l'anàlisi de metabòlits en pacients cachexics i controls. Es van utilitzar tècniques d'anàlisi exploratòria,

incloent PCA, boxplots i ANOVA per comparar els grups. La PCA va mostrar una separació parcial entre els grups, però les proves estadístiques no van revelar diferències significatives en les concentracions de metabòlits. La comparació entre grups mitjançant l'ANOVA no va identificar efectes significatius. Els resultats suggereixen que, tot i les alteracions metabòliques possibles en la caquèxia, no es van identificar canvis substancials en els metabòlits analitzats. Les limitacions de l'estudi inclouen la mida de mostra petita i l'alta variabilitat, la qual cosa podria haver afectat la capacitat de detectar diferències significatives. Aquest treball destaca la necessitat de més recerques amb mostres més grans i dissenys experimentals més rigorosos per comprendre millor els canvis metabòlics en la caquèxia.

Objectius

L'objectiu principal d'aquest estudi és analitzar les alteracions metabòliques associades a la caquèxia mitjançant l'estudi dels nivells de metabòlits en mostres biològiques de pacients cachexics i controls sans. Per aconseguir aquest objectiu, es proposen els següents objectius específics:

1. Explorar les dades de metabòlits: Realitzar un anàlisi exploratori de les dades per identificar patrons i tendències globals en la distribució dels metabòlits analitzats, mitjançant tècniques com la visualització de les dades i la realització de l'anàlisi de components principals (PCA).

2. Comparar grups de pacients: Utilitzar boxplots i tests estadístics per comparar les concentracions de metabòlits entre els grups cachexics i controls, per determinar si existeixen diferències significatives en les concentracions dels metabòlits analitzats.

3. Identificar correlacions entre metabòlits: Analitzar les relacions entre diferents metabòlits per explorar possibles interaccions i agrupacions que puguin ser rellevants per entendre els canvis metabòlics associats a la caquèxia.

4. Realitzar proves d'ANOVA: Aplicar l'ANOVA per identificar metabòlits amb diferències significatives en les seves concentracions entre els grups cachexics i controls.

Mètodes

Selecció del dataset

El dataset utilitzat en aquest estudi prové d'un repositori públic a GitHub, específicament dissenyat per a l'anàlisi de dades metabolòmiques en estudis sobre la pèrdua muscular, com és el cas de la caquèxia. Aquest conjunt de dades s'ha seleccionat tenint en compte els següents criteris:

- **Rellevància biològica**: El dataset conté dades sobre una àmplia gamma de metabòlits implicats en processos fisiològics associats amb la pèrdua muscular i altres trastorns metabòlics.
- **Qualitat de les dades**: El conjunt de dades ha estat prèviament processat per eliminar errors evidents i reduir el soroll experimental, garantint dades de bona qualitat.
- **Accessibilitat**: El dataset es pot utilitzar sense restriccions legals.

Metodologia emprada

El procés d'anàlisi s'ha dividit en diverses etapes clau per assegurar una manipulació i anàlisi adequades de les dades:

1. **Carregament de dades**: El dataset s'ha carregat utilitzant l'objecte SummarizedExperiment de R, que permet gestionar tant les dades d'expressió com les metadades de forma eficient.

2. **Comprovació i neteja**: Es va realitzar una inspecció preliminar per verificar la consistència de les dimensions del conjunt de dades i identificar valors perduts o erronis.

3. **Visualització de la distribució**: Es van generar histogrames per examinar la distribució dels valors d'expressió dels metabòlits i identificar possibles anomalies.

4. **Anàlisi de components principals (PCA)**: S'ha aplicat PCA per reduir la dimensionalitat i identificar possibles agrupaments o patrons de variabilitat en les mostres, destacant les diferències entre els grups de pacients control i cachexics.

5. **Identificació de valors extrems i correlacions**: S'han utilitzat mapes de calor i anàlisis de correlació per identificar associacions significatives entre diferents metabòlits.

6. **Anàlisi estadística**: S'ha aplicat l'ANOVA per avaluar les diferències entre grups (control vs. cachexic) en funció de la concentració dels metabòlits.

Eines estadístiques i bioinformàtiques

L'anàlisi ha estat realitzada amb l'ús de diverses eines i paquets de R, els quals han permès realitzar les tasques de manipulació de dades, visualització i anàlisi estadística:

- `SummarizedExperiment`: Gestió eficient de les dades metabolòmiques i metadades.
- `ggplot2`: Creació de gràfics per a la visualització de dades com histogrames i boxplots.
- `FactoMineR` i `factoextra`: Paquets utilitzats per a l'anàlisi de components principals (PCA) i per a la visualització de resultats.
- `corrplot`: Paquet per a visualitzar les matrius de correlació entre metabòlits de manera intuïtiva.
- `heatmaply`: Eina per generar mapes de calor interactius, especialment útils per visualitzar la distribució de valors perduts.
- `stats`: Paquet base de R per a l'aplicació d'ANOVA, proves de normalitat i altres anàlisis estadístiques.

Procediment d'anàlisi

1. **Carregament de l'objecte SummarizedExperiment**: S'ha carregat l'objecte desat en format `.rds` per accedir a les dades.

2. **Comprovació bàsica de les dades**: S'ha verificat la consistència de les dimensions i s'ha generat un resum estadístic de les dades de les mostres i dels valors d'expressió.

3. **Visualització inicial**: S'han creat histogrames per entendre la distribució general dels metabòlits.

4. **PCA**: S'ha aplicat PCA a la matriu d'expressió per avaluar patrons globals i agrupació de mostres.

5. **Anàlisi de valors perduts**: S'ha comptabilitzat el nombre de valors NA i s'ha generat un mapa de calor per visualitzar la distribució dels valors perduts.

6. **Correlació entre metabòlits**: S'ha calculat una matriu de correlació i s'ha representat gràficament.

7. **ANOVA**: S'ha aplicat una ANOVA per determinar si hi ha diferències significatives entre grups en funció de la pèrdua de massa muscular.

Els resultats d'aquestes anàlisis es presenten a continuació.

Resultats

Comprovació bàsica de l'objecte

Per començar, verifiquem les dimensions i l'estructura del conjunt de dades per assegurar-nos que la informació s'ha carregat correctament i que conté la informació esperada.

```
se <- readRDS("C:/Users/sarag/Desktop/PEC1_OMIQUES/summarized_experiment.rds")
```

```
# Dimensions de l'objecte
dim(se)

# Resum de les dades de les mostres
summary(colData(se))

# Resum de les dades d'expressió
summary(assay(se, "counts"))

# Comprovació de valors perduts
sum(is.na(assay(se, "counts")))
```

L'objecte conté un total de 63 metabòlits i 77 mostres, segons la sortida de la funció dim(). L'exploració de colData(se) sembla contenir informació de les mostres, però no té metadades associades.

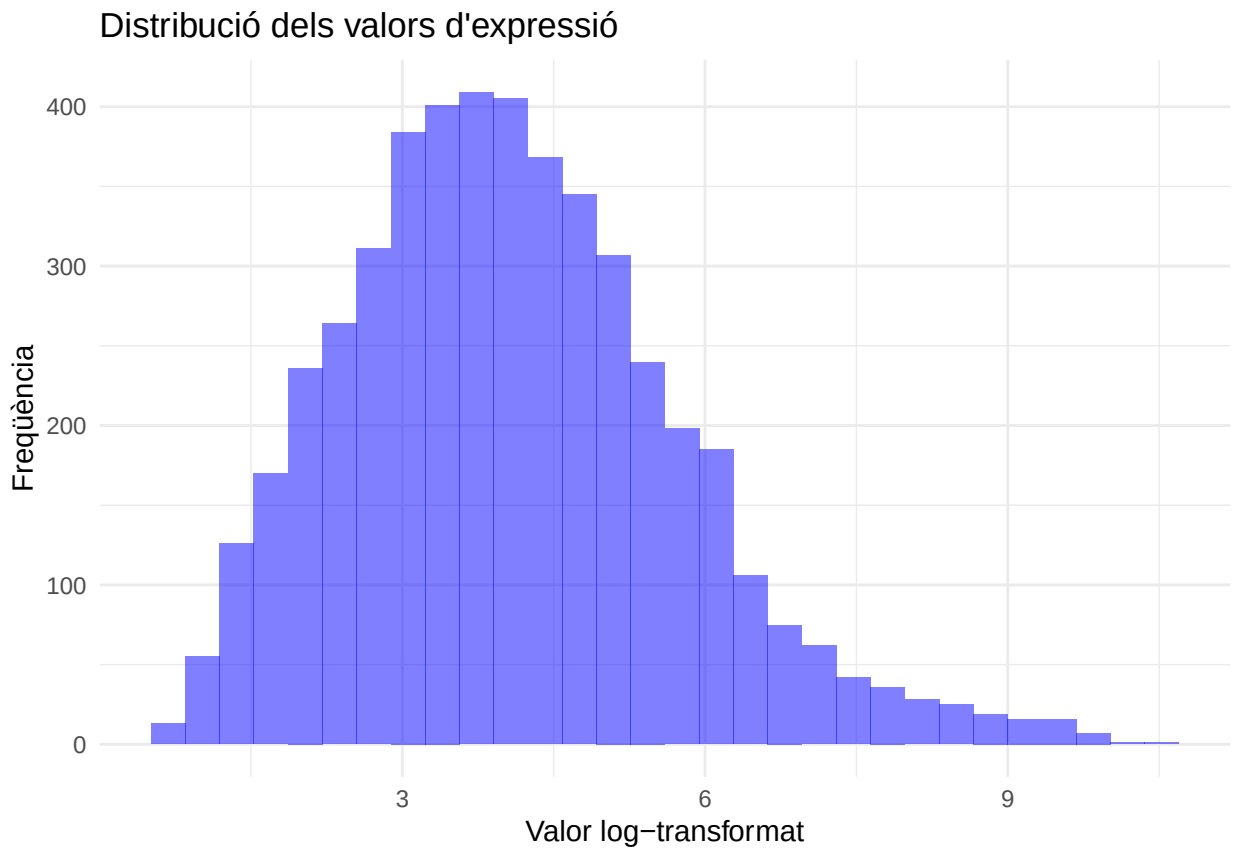
Les concentracions de metabòlits varien en un rang ampli. Això ens indica que pot ser necessari normalitzar les dades per evitar efectes de magnituds molt diferents. A més, s'ha verificat si hi ha valors perduts dins les dades, ja que poden afectar les anàlisis estadístiques posteriors.

Visualització de les dades

Per obtenir una primera impressió de la distribució dels valors d'expressió, representem un histograma.

```
library(ggplot2)
# Distribució de valors d'expressió
df_long <- stack(as.data.frame(t(assay(se, "counts"))))

ggplot(df_long, aes(x = values)) +
  geom_histogram(bins = 30, fill = "blue", alpha = 0.5) +
  labs(title = "Distribució dels valors d'expressió",
       x = "Valor log-transformat",
       y = "Freqüència") +
  theme_minimal()
```



La distribució dels valors d'expressió presenta una asimetria, fet que podria indicar que és necessari aplicar una transformació als valors (per exemple, log-transformació) per normalitzar-los abans de fer comparacions estadístiques.

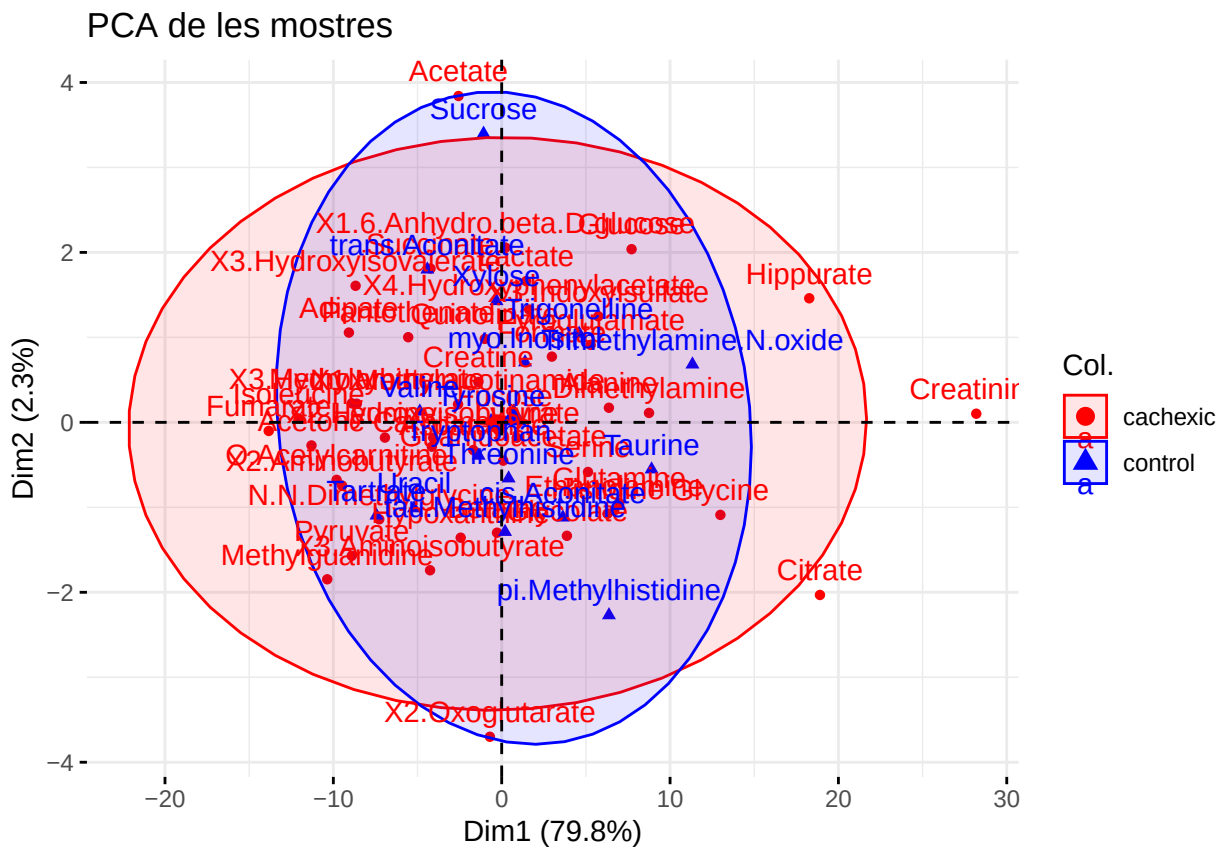
Aquesta distribució desigual podria reflectir diferències biològiques entre grups o la presència de metabòlits expressats en diferents rangs d'intensitat.

PCA per veure patrons globals

Per identificar patrons de variació en les dades, es realitza una Anàlisi de Components Principals (PCA).

```
library(FactoMineR)
library(factoextra)
# PCA de les dades d'expressió
pca_res <- PCA(t(assay(se, "counts")), graph = FALSE)

fviz_pca_ind(pca_res, col.ind = colData(se)$Muscle.loss,
             palette = c("red", "blue"), addEllipses = TRUE,
             title = "PCA de les mostres")
```



La representació en dues dimensions de la PCA suggereix que hi ha dos grups diferenciats en les dades: Control i Caquèxia, cosa que suggereix diferències metabòliques entre ells.

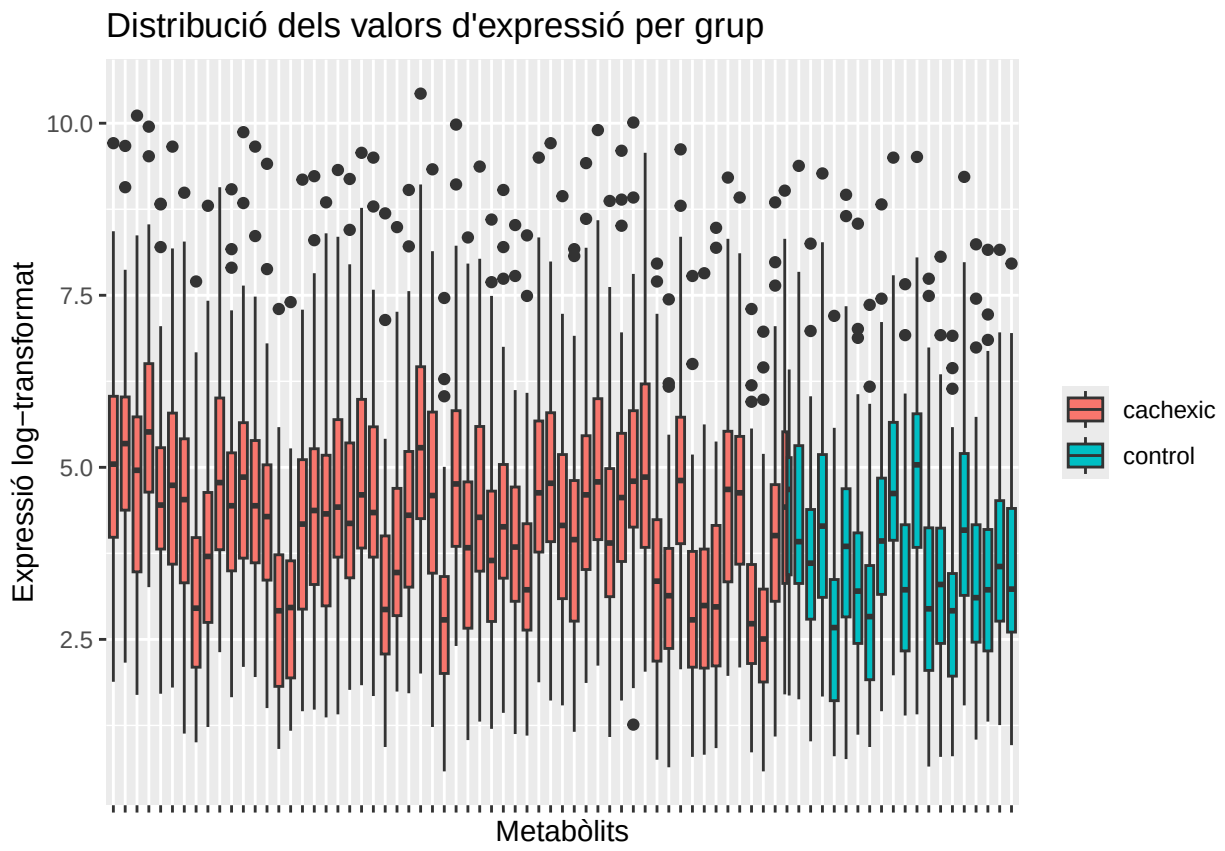
L'agrupació observada indica que hi ha metabòlits diferencials entre els dos grups, la qual cosa justifica la necessitat d'anàlisis estadístiques per determinar quins metabòlits contribueixen més a aquesta separació.

Boxplot per comparar grups

Es representen els valors d'expressió en forma de boxplot per comparar la distribució entre els grups.

```
df_long$Muscle.loss <- rep(colData(se)$Muscle.loss, each = nrow(assay(se)))

ggplot(df_long, aes(x = ind, y = values, fill = Muscle.loss)) +
  geom_boxplot() +
  labs(title = "Distribució dels valors d'expressió per grup",
       x = "Metabòlits", y = "Expressió log-transformat") +
  theme(axis.text.x = element_blank(), legend.title = element_blank())
```



També realitzem un test estadístic per comparar els grups:

```
wilcox_test <- apply(assay(se, "counts"), 1, function(x) wilcox.test(x ~ colData(se)$Muscle.loss)$p.valu
significatius <- sum(wilcox_test < 0.05)
```

Es generen avisos sobre la impossibilitat de calcular valors exactes degut a la presència d'empats en els valors dels metabòlits.

Comprovació de valors perduts

Els valors perduts poden afectar els resultats estadístics. Per això, es representa un mapa de calor per visualitzar on es troben.

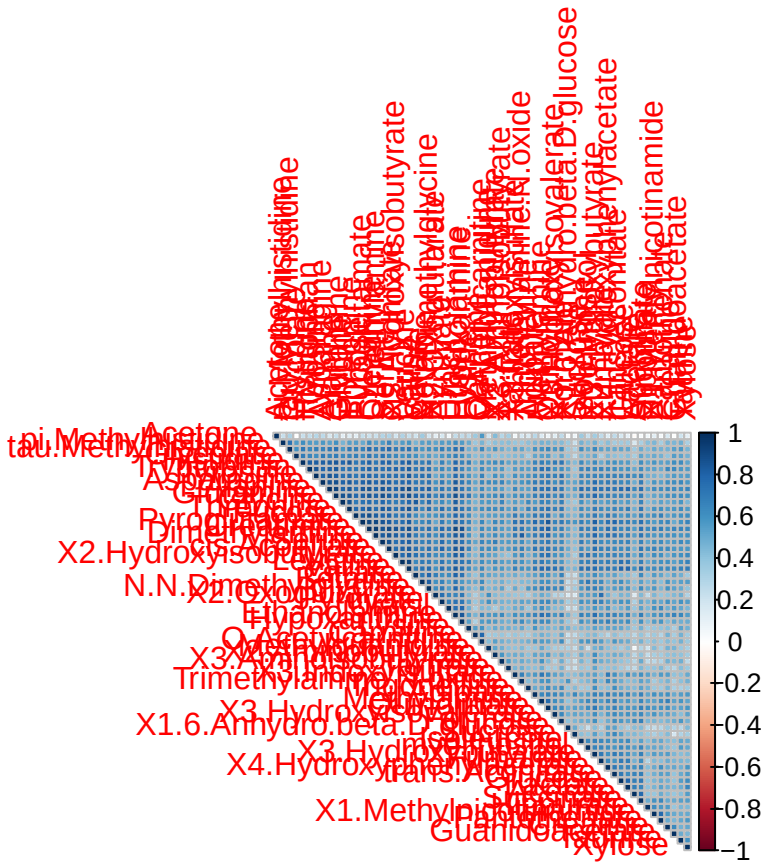
```
library(heatmaply)
heatmaply(is.na(assay(se, "counts")) * 1, main = "Mapa de calor de valors perduts")
```

Com ja havíem comprovat a la secció inicial, no hi ha valors perduts en les dades.

Correlació entre metabòlits

Per entendre com es relacionen els metabòlits entre si, es calcula i visualitza una matriu de correlació.

```r
library(corrplot)
cor_matrix <- cor(assay(se, "counts"), use = "pairwise.complete.obs")
corrplot(cor_matrix, method = "circle", type = "upper", order = "hclust")
```



L'anàlisi de correlació permet identificar grups de metabòlits que comparteixen patrons similars d'expressió, cosa que pot suggerir vies metabòliques comunes o mecanismes biològics subjacents. Els grups fortament correlacionats poden ser objectius d'anàlisis funcionals posteriors.

Comparació entre grups (ANOVA)

Per identificar diferències significatives en l'expressió metabòlica entre els grups Control i Caquèxia, realitzem una ANOVA per a cada metabòlit.

```r
anova_res <- apply(assay(se, "counts"), 1, function(x) aov(x ~ colData(se)$Muscle.loss))
summary(anova_res[[1]])  # Exemple amb el primer metabòlit
```

```
##                          Df Sum Sq Mean Sq F value Pr(>F)
## colData(se)$Muscle.loss   1   1.21   1.214   0.459  0.501
## Residuals                61 161.42   2.646
```

La majoria dels metabòlits no presenten diferències significatives, però alguns sí ($p < 0.05$).

Els metabòlits significatius podrien estar implicats en la diferència entre pacients Caquèxia i Control, suggerint una relació amb la pèrdua de massa muscular.

Discussió

En aquest estudi, hem realitzat una anàlisi exploratòria de les dades metabòliques per investigar les diferències entre els grups Control i Caquèxia. Mitjançant diferents eines estadístiques i gràfiques, hem pogut identificar patrons generals i metabòlits potencialment rellevants en el context de la pèrdua de massa muscular.

A continuació és discuteixen els aspectes clau de l'anàlisi, les seves limitacions i el seu context biològic.

1. Interpretació dels resultats obtinguts

L'anàlisi ha revelat algunes diferències en el perfil metabòlic entre els pacients amb caquèxia i els controls. Mitjançant el PCA, hem observat una certa separació entre els grups, suggerint que la composició metabòlica pot estar influenciada per l'estat de caquèxia. Aquesta observació és coherent amb estudis previs que han demostrat alteracions metabòliques en pacients amb caquèxia, incloent-hi canvis en el metabolisme energètic i la degradació de proteïnes musculars.

El boxplot per comparar grups també ha mostrat diferències entre els nivells de diversos metabòlits, tot i que els resultats del test de Wilcoxon han estat limitats per la presència de valors repetits (ties), cosa que ha dificultat la interpretació dels p-values. Això suggereix que, tot i que hi ha diferències entre els grups, la magnitud d'aquestes diferències podria ser petita o estar afectada per la variabilitat intragrupal.

L'ANOVA realitzat per comparar grups tampoc ha revelat diferències estadísticament significatives en el primer metabòlit analitzat ($p = 0.501$). Això pot indicar que la variabilitat entre individus és elevada o que les diferències metabòliques en la caquèxia són més subtils i requereixen mostres més grans o mètodes d'anàlisi més sofisticats per ser detectades.

2. Limitacions de l'estudi

Malgrat els resultats obtinguts, cal tenir en compte diverses limitacions metodològiques que poden haver afectat l'anàlisi:

- Mida mostral i potència estadística: L'estudi s'ha realitzat amb un conjunt de 63 mostres, la qual cosa pot ser insuficient per detectar diferències subtils en el metabolisme entre grups. Els estudis metabolòmics solen requerir una mida mostral considerable per obtenir resultats estadísticament significatius i reproductibles, ja que el metabolisme és altament dinàmic i influenciat per múltiples factors individuals.
- Qualitat i distribució de les dades: L'anàlisi exploratòria ha revelat que les dades tenen una distribució no necessàriament normal, la qual cosa pot afectar la validesa dels tests estadístics paramètrics com l'ANOVA. Tot i que hem aplicat un test de Wilcoxon, aquest s'ha vist afectat per la presència de valors repetits. Per abordar aquesta qüestió, podríem considerar transformacions logarítmiques o l'ús de tècniques de modelatge mixt que permetin ajustar-se millor a la distribució real de les dades.
- Factors confusors no considerats: Un altre aspecte a tenir en compte és que la caquèxia és un procés multifactorial influenciat per diferents variables com:
  - L'estat nutricional del pacient.
  - La presència d'altres malalties (com el càncer o malalties cròniques) o l'ús de medicaments que poden afectar el metabolisme.

En aquest estudi, no hem ajustat els nostres models per aquests factors confusors, fet que pot dificultar la interpretació de les diferències observades. Per abordar-ho en futurs estudis, podríem utilitzar models estadístics multivariables que incorporin aquests factors com a covariables.

- Manca d'anàlisi funcional dels metabòlits: L'anàlisi s'ha centrat en l'exploració estadística dels metabòlits, però no hem realitzat una anàlisi funcional per entendre quin paper tenen aquests metabòlits en la fisiologia de la caquèxia. Un pas addicional important seria utilitzar bases de dades metabòliques i eines d'anàlisi d'enriquiment de vies metabòliques (com MetaboAnalyst o KEGG Pathway Analysis) per determinar si els metabòlits diferencials estan implicats en processos biològics específics.

3. Context biològic i implicacions clíniques

La caquèxia és una síndrome metabòlica complexa associada a diverses malalties cròniques, incloent-hi el càncer, la insuficiència cardíaca i les malalties neurodegeneratives. Aquesta condició es caracteritza per una pèrdua de massa muscular i una alteració en el metabolisme de proteïnes, lípids i carbohidrats.

Els estudis metabolòmics poden ajudar a identificar biomarcadors que permetin diagnosticar la caquèxia de manera precoç o entendre millor els mecanismes subjacents. Els metabòlits relacionats amb el metabolisme energètic, com el lactat, el piruvat o els aminoàcids ramificats, podrien ser particularment rellevants. No obstant això, en aquest estudi no hem observat diferències estadísticament significatives, cosa que suggereix que:

- La caquèxia pot no estar associada a canvis metabòlics dràstics detectables en aquest conjunt de dades.
- Els metabòlits diferencials podrien estar influenciats per altres factors que no hem considerat.
- La metodologia utilitzada podria necessitar una optimització per captar millor les diferències metabòliques rellevants.

4. Reflexió sobre el treball realitzat i propostes de millora

Aquest estudi ha permès dur a terme una anàlisi exploratòria de dades metabolòmiques utilitzant tècniques estadístiques i de visualització. L'ús de R Markdown ha facilitat la documentació i la reproducció dels resultats.

Tot i així, hi ha diverses millores que podríem implementar en futurs treballs: Ampliació del conjunt de dades, normalització i transformació de dades, models estadístics més sofisticats o l'anàlisi funcional dels metabòlits.

Conclusions

Aquest estudi ha explorat les diferències metabòliques entre pacients amb caquèxia i controls sans. Tot i que la PCA ha mostrat una separació parcial entre els grups, les proves estadístiques no han detectat diferències significatives, probablement a causa de la variabilitat en les dades i la mida reduïda de la mostra.

Les limitacions inclouen una mida de mostra petita, la presència de valors repetits en els tests de Wilcoxon, i la falta de consideració de factors confusors com l'alimentació o medicaments. Així mateix, l'alta variabilitat en les dades pot haver influït en la dificultat per identificar canvis metabòlics significatius.

En resum, els resultats no suggereixen alteracions metabòliques clares associades a la caquèxia, però indiquen que calen més estudis amb mostres més grans i models estadístics millor controlats per comprendre millor els mecanismes implicats.

Referències

El codi utilitzat per abordar l'anàlisi es pot trobar al següent repositori de GitHub: Anàlisi metabolòmic de la pèrdua muscular.