Anàlisis de dades òmiques (ADO) – Informe PAC1

TAULA DE CONTINGUTS

Abstract 2
Hipòtesis i Objectius 2
Hipòtesi 2
Objectius 2
Materials i Mètodes 2
Dades 2
Processament de les dades 2
Pre-processament i preparació de l'entorn de treball 2
Control de qualitat 3
Normalització de les dades i control de qualitat 3
Anàlisis de l'expressió diferencial de metabòlits 3
Resultats 4
Processament i control de qualitat de les dades crues 4
Normalització i control qualitat 5
Anàlisis de l'expressió diferencial de metabòlits 6
Anàlisis de significança biològica 7
Discussió 7
Conclusió 8
Referències 8
Annex. Procediment i codi emprat per la resolució de la PAC1 10
Preparació entorn de treball 10
Creació de l'objecte *SummarizedExperiment* 11
Exploració i control de qualitat de les dades 12
Normalització dades (logarítmica) 14
Control de qualitat després de la normalització (logarítmica) 16
Anàlisis de metabòlits diferencials 19
Anàlisis de significància biològica 23

ABSTRACT

La metabolòmica permet estudiar centenars de metabòlits presents en diferents fluids corporals amb la intenció d'identificar patrons bioquímics que venen per processos fisiològics, malalties o factors ambientals. D'aquesta manera, es pot arribar comprendre el funcionament de diversos mecanismes biològics (1).

Aquest treball descriu el procés d'anàlisi (obtenció de dades, processament i anàlisis) d'un *data set* amb diferents concentracions de metabòlits en orina en grups controls i pacients amb caquèxia, una síndrome metabòlica caracteritzada per la pèrdua de massa muscular (2,3). Els resultats mostren que els perfils metabòlics dels pacients amb caquèxia són clarament diferents en comparació al grup control, indicant possibles biomarcadors (creatinina, fumarat, etc) o vies associades a la malaltia. La identificació d'aquests patrons pot ajudar a entendre el funcionament de la caquèxia, i alhora pot servir d'eina clínica pel diagnòstic d'algunes malalties com ara el càncer, on la caquèxia és un símptoma comú.

HIPÒTESIS I OBJECTIUS

HIPÒTESI

Els pacients amb caquèxia presenten un perfil metabòlic diferent en comparació al grup control. Aquests canvis es caracteritzen per alteracions en la concentració de determinats metabòlits.

OBJECTIUS

L'objectiu principal d'aquest treball és analitzar els perfils metabòlics presents en pacients que pateixen caquèxia en comparació amb els dels individus sans mitjançant l'ús de la metabolòmica. Per poder complir aquest objectiu, s'han plantejat diferents objectius específics:

1. Descarregar, processar i explorar el conjunt de dades *cachexia*.
2. Dur a terme un estudi bioestadístic i bioinformàtic per identificar metabòlits diferencialment expressats en pacients amb caquèxia i controls.
3. Dur a terme un estudi bioestadístic i bioinformàtic per identificar vies metabòliques o processos cel·lulars relacionats amb pacients amb caquèxia.
4. Interpretar els resultats per identificar biomarcadors i vies metabòliques associades a la malaltia.

MATERIALS I MÈTODES

DADES

El conjunt de dades emprat en aquest estudi prové d'un *data set* de metabolòmica de caquèxia que es troba en el repositori GitHub (https://github.com/nutrimetabolomics/metaboData/tree/main/Datasets/2024-Cachexia). Aquest conjunt de dades conté mesures de concentracions de 63 metabòlits en l'orina d'un total de 77 individus, classificats en dues condicions experimentals: 47 pacients amb caquèxia i 30 individus control.

La descripció de l'obtenció de les dades no és molt detallada, però el *data set* es pot trobar com a exemple d'ús a la plataforma MetaboAnalyst (https://www.metaboanalyst.ca/), però probablement s'hagin obtingut amb diferents mètodes de mesura de proteïna com podria ser l'espectrometria de masses.

PROCESSAMENT DE LES DADES

Pre-processament i preparació de l'entorn de treball

Les dades es van descarregar des del repositori de GitHub mitjançant un enllaç directe al fitxer CVS utilitzant la funció 'download.file'. Seguidament, aquestes dades es van carregar en un *dataframe* amb les dades del fitxer original. Seguidament, es va comprovar l'estructura de les dades per assegurar que la informació s'havia carregat correctament i no hi hagués presència de valors nuls.

Seguidament, es va procedir a transforma la taula original en una matriu, eliminant les dues primeres columnes amb *metadata* numèrica (ID pacient i grup pèrdua muscular). Les dades es van transposar per poder tenir les mostres com a

columnes i els metabòlits com a files, fet important per poder treballar de manera més eficient amb dades òmiques i amb alguns objectes com *SummarizedExperiment*. Seguidament, es va definir un conjunt de *metadata* amb la informació eliminada de la matriu. Un cop vam tenir aquests dos elements, es va crear l'objecte *SummarizedExperiment* contingut en el paquet SummarizedExperiment de Bioconductor. Aquest paquet està dissenyat per emmagatzemar múltiples tipus de dades que inclouen les dades numèriques del *assay* (en aquest cas quantitats de metabòlits) en format de matriu, associant-les amb la seva pròpia *metadata*. Aquest format, permet una manipulació i anàlisis de les dades eficient.

SummarizedExperiment és un tipus d'objecte flexible que permet emmagatzemar diferents tipus de dades òmiques (des de transcriptòmica fins a metabolòmica). En canvi *ExpressionSet* del paquet Biobase, és un objecte més orientat a l'estudi d'expressions gèniques, i per aquest motiu, aquest estudi és més adient utilitzar l'objecte *SummarizedExperiment*.

Control de qualitat

L'exploració de les dades i la realització d'un control de qualitat adient són essencials per assegurar-nos que les dades es troben degudament representades i compleixen condicions concretes per poder realitzar els anàlisis posteriors de manera adient. En aquest estudi, es van realitzar diversos tipus d'anàlisis (descrits en l'Annex de l'informe) per avaluar la distribució de les dades i determinar si requerien de transformacions. El control de qualitat ha inclòs:

- Estudi de la presència de valors NA.
- Generació de boxplot per observar la variabilitat de les concentracions dels diferents metabòlits per pacient.
- Generació d'un histograma de les concentracions de metabòlits per determinar la simetria de dades (de menor a major concentració). Es realitza per veure variabilitat i evitar baixos.
- Anàlisis d'agrupació jeràrquica mitjançant la generació d'un dendrograma per veure possibles unions entre grups.
- Estudi de components principals (PCA) per revelar si hi havia diferències entre les mostres (7).

Les observacions obtingudes en aquest estudi, permetran determinar si cal realitzar una transformació o normalització de les dades.

Normalització de les dades i control de qualitat

La normalització de les dades és un pas fonamental per la anàlisis de metabolòmica, especialment quan ens trobem amb dades que presenten una distribució amb gran variabilitat (que pot identificar-se amb els controls de qualitat). En els casos on hi ha una variabilitat elevada i presència d'outliers, pot ser útil fer una transformació logarítmica (8), d'aquesta manera els extrems són més propers a la mitjana i s'eviten biaixos que poden portar a errors a l'hora d'interpretar els resultats.

En metabolòmica i proteòmica hi ha elevada variabilitat, per aquest motiu s'opta en usar la transformació en **log2**, que permet reduir l'efecte dels metabòlits *outliers* i proporciona una millor comparació entre les mostres. Per dur a terme aquesta normalització, es van extreure les dades d'expressió recollides en l'objecte *SummarizedExperiment* i es va aplicar la transformació logarítmica afegint un 1 per evitar problemes amb valors igual a 0.

Un cop normalitzat, es van realitzar els mateixos estudis de control de qualitat descrits anteriorment, esperant una millora en els diferents paràmetres valorats: boxplot, histogrames, agrupació jeràrquica i estudi PCA. S'espera una reducció en la variabilitat, una distribució normalitzada de les diferents concentracions i una agrupació més distribuïda. En el cas de presentar millores, serà aquest objecte transformat el que es farà servir en els anàlisis.

Anàlisis de l'expressió diferencial de metabòlits

Per explorar les diferències en la presència de certs metabòlits entre els pacients que pateixen caquèxia i els del grup control, es va utilitzar un estudis estadístic utilitzant el paquet **Linear Models for Microarray Data** (**limma)** de Bioconductor (https://www.bioconductor.org/packages/release/bioc/html/limma.html). Aquest paquet permet fer una anàlisis diferencial de les concentracions entre els dos grups tenint en compte el *fold-change* i el p-valor ajustat.

El codi utilitzat es pot trobar en l'Annex de l'informe, però breument, el que es va fer va ser crear una matriu per identificar les diferències entre els dos grups d'estudi tenint en compte la variable Muscle los. Seguidament, amb lmFit i eBayes es van ajustar els p valors i calcular els coeficients de canvi per cada metabòlit. Aquests valors son els que ens permetran determinar quins són els metabòlits diferencials entre grups.

Per visualitzar millor aquestes diferències, es van generar certs gràfics que es descriuen a continuació:

- Volcano plot per identificar els metabòlits amb un *fold-change* i un p-valor ajustat significatiu. Es va realitzar amb el paquet de EnhancedVolcano de Bioconductor (9).
- Heatmap per observar perfils d'expressió dels metabòlits diferencials entre els dos grups. Permet visualitzar nivells d'expressió de cada metabòlit de forma relativa per cada pacient i la seva condició (caquèxia o control). Es va realitzar amb el paquet de ComplexHeatmap de Bioconductor (10).
- Boxplot dels metabòlits amb p-valors i *fold-change* més elevats per comprovar l'expressió diferencial entre grups.
- Boxplot general per comprovar si en general, els pacients amb caquèxia presentaven una expressió diferencial més gran o més petita dels metabòlits estudiats en comparació amb el grup control.

Anàlisis de significança biològica

Un cop obtinguda la llista de gens entre dos grups, cal interpretar el context biològic. Per poder realitzar aquest tipus d'estudis, es realitza el que es coneix com metabolòmic enrichment analysis (MSEA). Aquest tipus d'anàlisis ens permet determinar si hi ha funcions, processos biològics o bé vies moleculars representats per l'expressió diferencial.

Degut a problemes d'incompatibilitat amb el paquet de R de MetaboAnalystR, aquest estudi es va realitzar amb la eina present en la web de MetaboAnalyst (https://dev.metaboanalyst.ca/ModuleView.xhtml) i es va realitzar l'estudi de *Enrichment Analysis*. Un cop s'entra en cada un d'aquests apartats, cal introduir els metabòlits sobre-expressats en el grup d'interès (en aquest cas caquèxia) i es selecciona quin tipus d'anàlisis volem que ens faci. En el nostre cas, es va seleccionar l'anàlisi basada en KEGG, una baser de dades de vies metabòliques ben establerta. Per altra banda, també es va realitzar estudi per veure l'associació d'aquests metabòlits amb malalties. El procediment seguit (juntament amb captures de pantalla) es troba descrit en l'Annex d'aquest treball.

Tot i que la plataforma ofereix altres possibilitats d'anàlisi, les visualitzacions obtingudes en aquest estudi son suficients per la interpretació dels resultats.

RESULTATS

PROCESSAMENT I CONTROL DE QUALITAT DE LES DADES CRUES

Les dades es van carregar i processar correctament, però prèviament a procedir amb el seu anàlisis, es van seguir una sèrie de controls de qualitat per veure si calia realitzar algun canvi o transformació per continuar treballant. Com esta descrit als resultats, es va realitzar un estudi de la distribució de les dades per Boxplot, Histograma, agrupacions jeràrquiques i PCAs.

Observem en el boxplot (Fig.1A) com la distribució de la concentració dels diferents metabòlits és bastant variable entre pacients. A més, l'histograma (Fig.1B) ens rebel·la com les dades es troben esbiaixades cap a concentracions baixes, el que podria fer que es traguessin conclusions errònies posteriors als anàlisis. Normalment cal que les dades presentin una distribució normal. A més a més, l'agrupació jeràrquica (Fig.1C) mostra una distribució poc homogènia dels pacients segons expressió i el PCA (Fig.1D) presenta totes les dades esbiaixades cap a l'esquerra del gràfic juntament amb dades de dos pacients que es poden considerar outliers.

Aquests resultats ens confirmen que cal realitzar una normalització de les dades per evitar que la variabilitat entre mostres i els outliers identificats tinguin un efecte en les conclusions que es poden extreure de l'estudi i anàlisis estadístic.

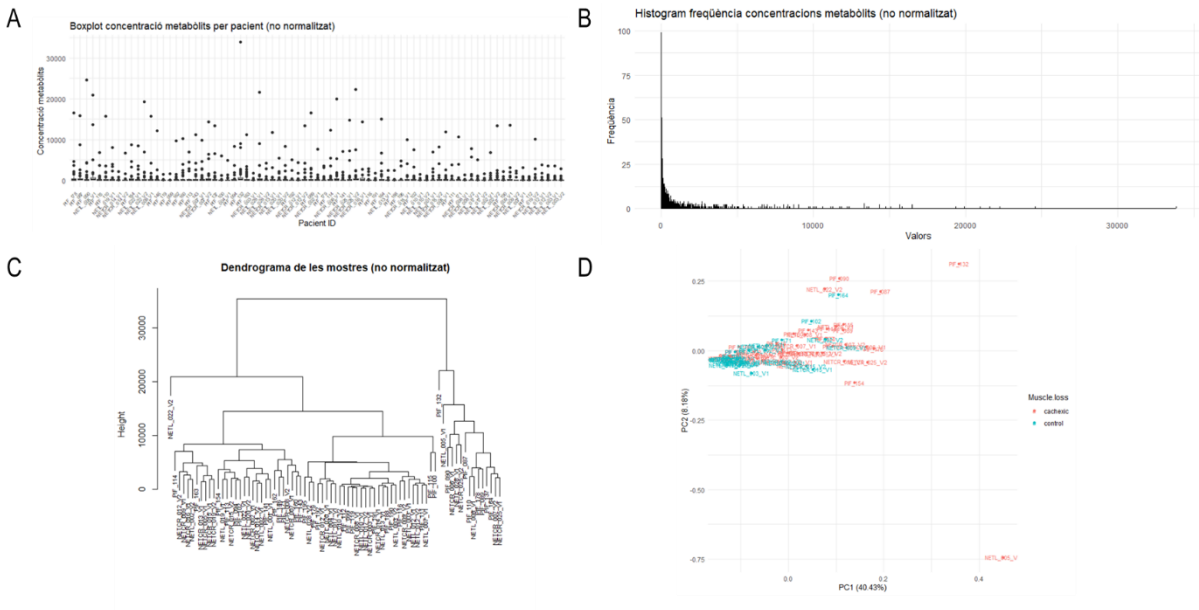


Figura 1. Control qualitat mostres sense normalitzar. *A) Boxplot distribució metabòlits per pacient. B) Histograma distribució concentracions metabòlits. C) Dendograma agrupació jeràrquica pacients. D) PCA pacients segons nivells metabòlits.*

NORMALITZACIÓ I CONTROL QUALITAT

En dades de metabolòmica, una manera comú de normalitzar les dades és mitjançant la transformació logarítmica, per tant procedim a fer-la i per determinar si es poden usar aquestes dades transformades, realitzem el mateix control de qualitat.

En aquest cas observem que en el boxplot (Fig.2A) la distribució dels metabòlits per pacient és molt més homogènia, i per tant comparable. A més a més, aquest canvi, provoca que aparegui una distribució normal a l'histograma (Fig.2B) representat (Fig.2B), el que permetrà que les dades un cop es facin comparacions, no es troben esbiaixades pels metabòlits amb major concentració. Per altra banda, la distribució del dendrograma (Fig.2C) és molt més idònia, i en el PCA (Fig.2D) els pacients es troben distribuïts de manera heterogènia (no esbiaixats a un costat) i no s'ha reduït la diferència entre possibles outliers i la resta de mostres.

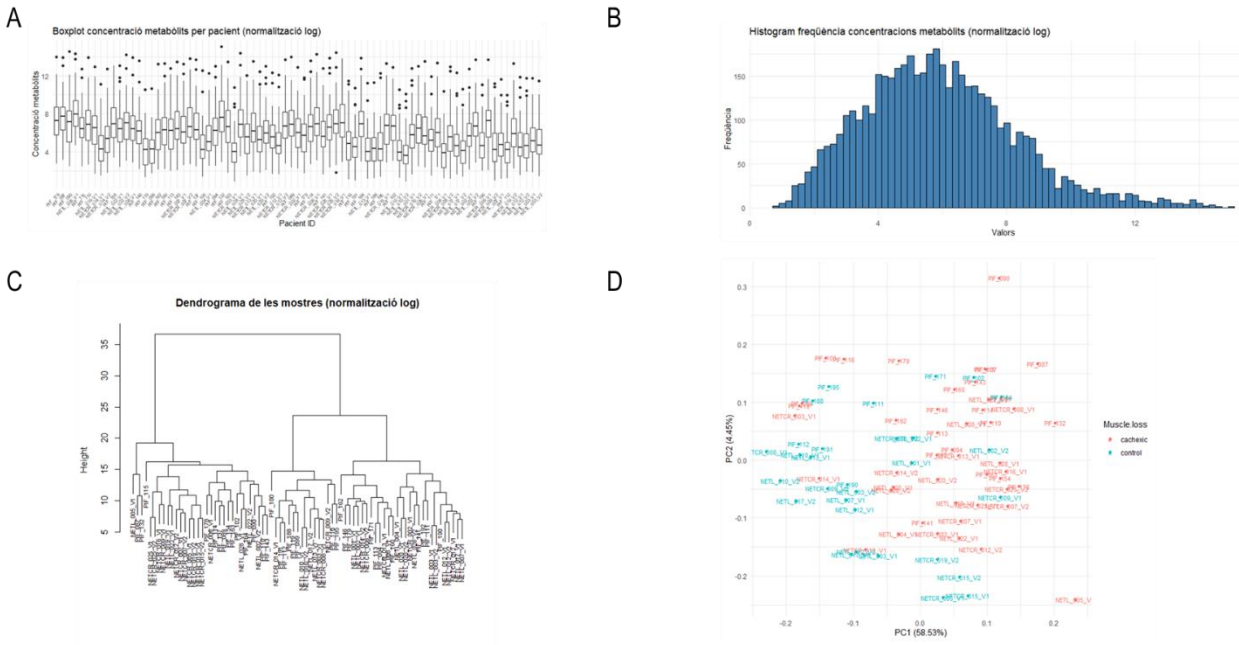


Figura 2. Control qualitat mostres normalitzades logarítmicament. *A) Boxplot distribució metabòlits per pacient. B) Histograma distribució concentracions metabòlits. C) Dendograma agrupació jeràrquica pacients. D) PCA pacients segons nivells metabòlits.*

Aquest estudi rebel·la que la transformació / normalització de dades logarítmiques ha funcionat, i que per tant, es pot procedir amb els diferents anàlisis estadístics i de funció biològica.

ANÀLISIS DE L'EXPRESSIÓ DIFERENCIAL DE METABÒLITS

Per poder realitzar aquest anàlisis, fem servir el paquet de *limma*. Comparant el *fold-change* i el p-valor ajustat entre els diferents grups de treball (control i caquèxia), observem que hi ha diversos metabòlits que es troben en quantitats més elevades en els pacients amb caquèxia (llista en annex de l'informe).

Per poder visualitzar bé els resultats de *limma* es va procedir a realitzar un VolcanoPlot i un Heatmap. Aquests dos tipus de gràfics, ens permeten veure de forma més visual quins metabòlits es troben més expressats en caquèxia comparat amb el grup control. En el volcano plot generat (Fig.3A), sorprenentment observem que els 63 metabòlits analitzats en l'estudi presenten tots concentracions significativament elevades en pacients amb caquèxia en comparació al grup control. Tot i que el heatmap (Fig. 3B) ens hauria de mostrar aquesta mateixa informació, observem que aquesta no és tan clara. Hi ha una tendència del grup caquèxia a tenir més expressió de pràcticament tots els metabòlits, s'observa més vermell, però a ull no és del tot clar.

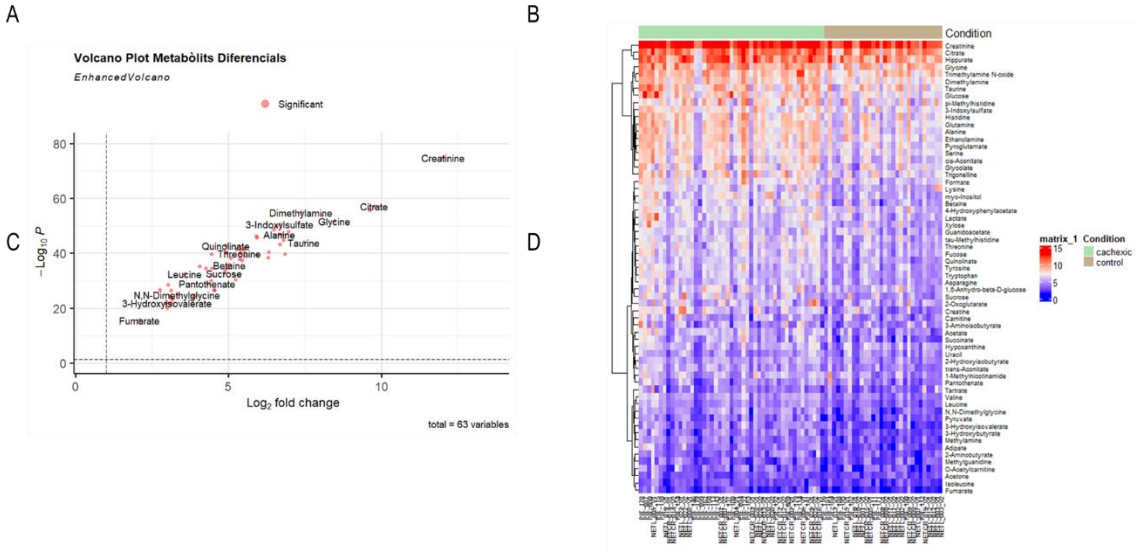


Figura 3. Anàlisi d'expressió diferencial de metabòlits entre pacients amb caquèxia i grup control. *A) Volcano Plot amb expressió diferencial de gens en Caquèxia contra grup control. B) Heatmap amb expressió de metabòlit per pacient i condició.*

Degut a la manca de claredat en el Heatmap i per tenir de forma més visual els resultats del volcano plot, es van realitzar una sèrie de Boxplots per veure la diferent quantitat de certs metabòlits o metabòlits generals entre el grup caquèxia i el grup control. Observem que quan generem els gràfics pels 5 metabòlits amb un p-valor ajustat més elevat (Fig.4A), hi ha una diferència clara entre els pacients i els controls, essent els individus amb caquèxia els que presenten concentracions més elevades. Per altra banda, si ajuntem els nivells de tots els metabòlits en un mateix boxplot (Fig.4B) observem que el nivell de metabòlits generals en orina és molt més elevat en pacients comparat amb controls.

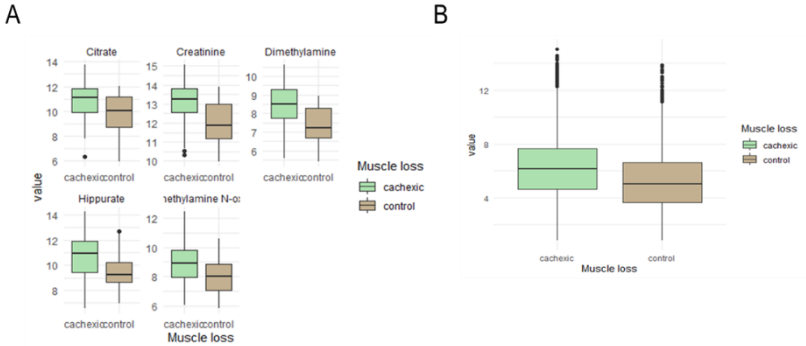


Figura 4. Anàlisi concret d'expressió diferencial de metabòlits entre pacients amb caquèxia i grup control. *A) Boxplot diferència concentracions dels 5 metabòlits amb major p-valor ajustat. B) Boxplot amb diferent concentració de metabòlits general.*

ANÀLISIS DE SIGNIFICANÇA BIOLÒGICA

Finalment realitzem la anàlisi de significança biològica mitjançant l'eina proporcionada per MetaboAnalyst. Per fer-ho, fem estudis d'*Enrichment Analysis* on podem veure quines vies estan sobre-representades i també amb quines malalties es pot trobar associat la presència d'aquests metabòlits.

Quan realitzem l'estudi, podem observar que les vies amb acumulació dels metabòlits que es troben amb pacients amb caquèxia es troben relacionades amb el metabolisme de glicoxilat, dicarboxilat, alanina, aspartat, flutamat, glicina, serina, treonina, el cicle del citrat i la biosíntesi de valina, leucina i isoleucina (Fig.5A). Per altra banda, si mirem amb quines malalties es pot associar, observem que la més representada per aquest perfil metabòlic seria el càncer de pulmó, seguit de esofagitis, càncer colorecta, etc (Fig.5B).

A B

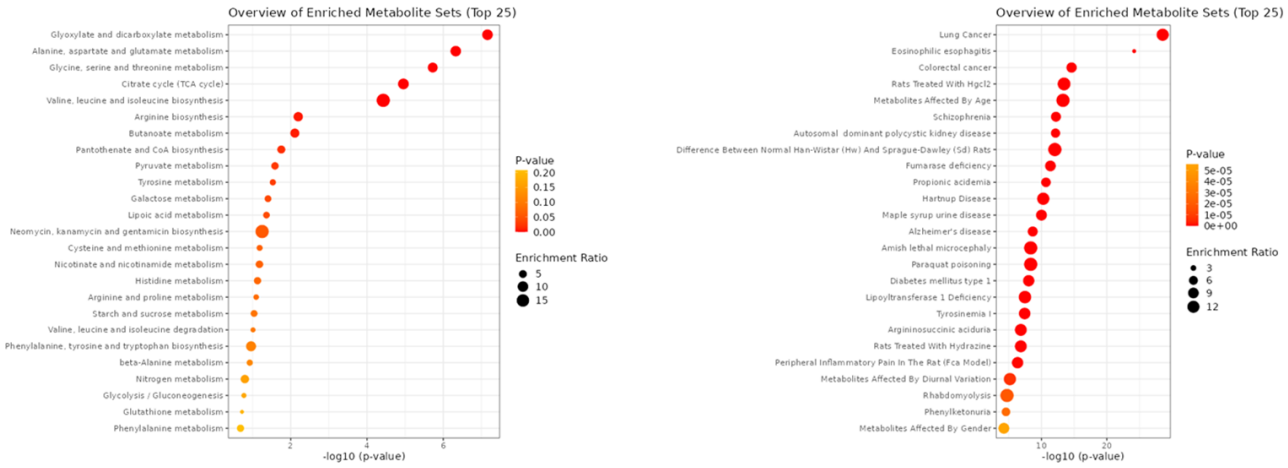


Figura 5. Anàlisi de significança biològica dels metabòlits elevats en caquèxia. *A) Enrichment analysis* en vies metabòliques. *B) Enrichment analysis* en malalties.

DISCUSSIÓ

Els resultats obtinguts, rebel·len que hi ha diferències clares, evidents i significatives en la concentració de metabòlits en orina entre pacients amb caquèxia i el grup control. En primer lloc, els pacients amb caquèxia presenten una quantitat significativament més alta en orina dels 63 metabòlits en estudi en comparació al grup control. Això dona lloc a la identificació d'alteracions de diferents vies metabòliques com ara el metabolisme de glioxilat, dicarboxilat, etc que no es veuen compensades fisiològicament pel pacient.

La caquèxia és una síndrome metabòlica complexa on el ritme metabòlic basal augmenta sense tenir una compensació basal. Aquest fet pot implicar la descomposició de greix i múscul (pèrdua muscular), que pot donar lloc a l'augment de metabòlits en sang i orina degut a aquest augment de metabolisme catabòlic. Aquest fet, explicaria perquè tots els metabòlits en estudi es troben augmentats en comparació als controls, i perquè hi ha certs vies metabòliques més representades (10,11). L'estudi d'aquestes vies metabòliques i metabòlits concrets, podrien servir com a biomarcadors per identificar pacients amb caquèxia, ja que clarament presenten nivells més elevats de metabòlits derivats del catabolisme.

Per altra banda, la presència de caquèxia pot indicar la presència o aparició d'altres malalties cròniques greus com ara el càncer, la insuficiència cardíaca congestiva, la insuficiència renal, malaltia pulmonar obstructiva crònica (MPOC) i infeccions cròniques com ara la tuberculosi o el VIH (11,12). Aquest fet és important, ja que la identificació de la caquèxia mitjançant paràmetres metabòlics, podria ajudar a realitzar un diagnòstic precoç d'aquest tipus de malalties. Per aquest motiu es considera important estudiar la rellevància biològica d'aquesta síndrome.

Tot i que aquest estudi ens ha permès tenir una visió general de com funciona la caquèxia i quin perfil metabòlic tenen els pacients en comparació al grup control, presenta certes limitacions com per exemple la mida de la mostra. El *data set* analitzat presenta 77 pacients i la representació en ambdós grups no és òptima, fet que pot limitar la generalització de

resultats. Per aquest motiu, seria interessant augmenta la mida de la mostra en futurs estudis per poder corroborar i validar els resultats.

A més a més, l'estudi presenta poques dades dels pacients estudiants, simplement sabem el nivell de metabòlits i si son del grup control o bé pateixen caquèxia. Aquest factor és bastant limitant, ja que podem estar obviant altres elements que poden tenir un impacte en el metabolisme dels pacients com per exemple el seu gènere, edat, malalties, etc. La inclusió d'aquesta informació permetria un anàlisis més detallat i precís, ja que ignorar aquests paràmetres pot portar a fer inferències errònies.

Cal afegir que l'estudi de 63 elements en metabolòmica es considera la inclusió d'un nombre moderat de metabòlits. Tot i que podem obtenir informació valuosa, afegint més metabòlits podríem tenir una visió més completa del perfil metabòlic dels pacients. A més, potser es podrien identificar metabòlits amb unes diferències més clares en l'expressió entre grup caquèxia i grup control, ja que com s'observa a PCA i al Heatmap obtinguts durant l'estudi, tot i haver diferències, aquestes no són del tot clares.

En conclusió, malgrat les limitacions mencionades, els resultats d'aquest estudi proporcionen una primera base per entendre els patrons metabòlics associats a la caquèxia, i alhora, per la identificació de biomarcadors pel seu diagnòstic.

CONCLUSIÓ

Després de realitzaré l'estudi del *data set cachexia*, realitzar el processament de les dades i procedir amb el seu anàlisi, podem concloure que la nostra hipòtesis inicial era certa i que el perfil metabòlic dels pacients amb caquèxia difereix àmpliament dels pacients del grup control. Els individus amb caquèxia presenten nivells significativament més elevats de metabòlits, degut a un augment del catabolisme basal no compensat. A més, la presència d'aquests metabòlits i la caquèxia podria veure's correlacionada amb una síndrome metabòlica derivada d'altres condicions.

Tot i així, per poder extreure conclusions més fiables i estadísticament robustes, seria millor poder comptar amb un nombre més elevat d'individus, metabòlits analitzats i *metadata* relacionada amb característiques dels individus de l'estudi. L'aplicació d'aquests elements, permetria evitar possibles artefactes o biaixos a l'hora de poder determinar les conclusions.

REFERÈNCIES

Repositori GitHub: https://github.com/cpani6/Panisello-Aranda-Carla-PEC1

1. Clish CB. Metabolomics: an emerging but powerful tool for precision medicine. Cold Spring Harb Mol Case Stud. 2015 Oct;1(1):a000588.
2. Baker Rogers J, Syed K, Minteer JF. Cachexia. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK470208/.
3. Ferrer M, Anthony TG, Ayres JS, Biffi G, Brown JC, Caan BJ, Cespedes Feliciano EM, Coll AP, Dunne RF, Goncalves MD, Grethlein J, Heymsfield SB, Hui S, Jamal-Hanjani M, Lam JM, Lewis DY, McCandlish D, Mustian KM, O'Rahilly S, Perrimon N, White EP, Janowitz T. Cachexia: A systemic consequence of progressive, unresolved disease. Cell. 2023 Apr 27;186(9):1824-1845.
4. Chen Y, Li EM, Xu LY. Guide to Metabolomics Analysis: A Bioinformatics Workflow. Metabolites. 2022 Apr 15;12(4):357.
5. Han, Xikun, and Liming Liang. "metabolomicsR: A Streamlined Workflow to Analyze Metabolomic Data in R." Bioinformatics Advances. 2022. 2 (1): vbac067.
6. Joo J. Processing quantitative metabolomics data with the qmtools Package.2024. Bioconductor[Internet]. Available from: https://www.bioconductor.org/packages/devel/bioc/vignettes/qmtools/inst/doc/qmtools.html#Session_info.
7. Plotting PCA (Principal Component Analysis). Cran-R project[Internet]. Available from: https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html
8. Misra, Biswapriya B. "Data normalization strategies in metabolomics: Current challenges, approaches, and tools." European Journal of Mass Spectrometry 26.3 (2020): 165-174.

9. Blighe K, Rana S, Lewis M. EnhancedVolcano: publication-ready volcano plots with enhanced colouring and labeling. Bioconductor[Internet]. 2024. Available from: https://bioconductor.org/packages/release/bioc/vignettes/EnhancedVolcano/inst/doc/EnhancedVolcano.html.
10. Neshan M, Tsilimigras DI, Han X, Zhu H, Pawlik TM. Molecular Mechanisms of Cachexia: A Review. Cells. 2024 Jan 29;13(3):252.
11. Ferrer M, Anthony TG, Ayres JS, Biffi G, Brown JC, Caan BJ, Cespedes Feliciano EM, Coll AP, Dunne RF, Goncalves MD, Grethlein J, Heymsfield SB, Hui S, Jamal-Hanjani M, Lam JM, Lewis DY, McCandlish D, Mustian KM, O'Rahilly S, Perrimon N, White EP, Janowitz T. Cachexia: A systemic consequence of progressive, unresolved disease. Cell. 2023 Apr 27;186(9):1824-1845.
12. Argilés JM, López-Soriano FJ, Stemmler B, Busquets S. Cancer-associated cachexia - understanding the tumour macroenvironment and microenvironment to improve management. Nat Rev Clin Oncol. 2023 Apr;20(4):250-264.

ANNEX. PROCEDIMENT I CODI EMPRAT PER LA RESOLUCIÓ DE LA PAC1

PREPARACIÓ ENTORN DE TREBALL

```
# Carreguem llibrerias
library(ggplot2)
library(reshape2)
library(SummarizedExperiment)
library(ggfortify)
library(DESeq2)
library(limma)
library(EnhancedVolcano)
library(ComplexHeatmap)

# Primer de tot descarreguem les dades del fitxer de GitHub
url_cachexia <- "https://github.com/nutrimetabolomics/metaboData/raw/main/Datasets/2024-Cachexia/h
uman_cachexia.csv"
download.file(url_cachexia, destfile = "human_cachexia.csv")
cachexia_data <- read.csv("human_cachexia.csv",check.names = FALSE)

str(cachexia_data) # Dades ben carregades

## 'data.frame':    77 obs. of  65 variables:
##  $ Patient ID                : chr  "PIF_178" "PIF_087" "PIF_090" "NETL_005_V1" ...
##  $ Muscle loss               : chr  "cachexic" "cachexic" "cachexic" "cachexic" ...
##  $ 1,6-Anhydro-beta-D-glucose: num  40.9 62.2 270.4 154.5 22.2 ...
##  $ 1-Methylnicotinamide      : num  65.4 340.4 64.7 53 73.7 ...
##  $ 2-Aminobutyrate           : num  18.7 24.3 12.2 172.4 15.6 ...
##  $ 2-Hydroxyisobutyrate      : num  26.1 41.7 65.4 74.4 83.9 ...
##  $ 2-Oxoglutarate            : num  71.5 67.4 23.8 1199.9 33.1 ...
##  $ 3-Aminoisobutyrate        : num  1480.3 116.8 14.3 555.6 29.7 ...
##  $ 3-Hydroxybutyrate         : num  56.83 43.82 5.64 175.91 76.71 ...
##  $ 3-Hydroxyisovalerate      : num  10.1 79.8 23.3 25 69.4 ...
##  $ 3-Indoxylsulfate          : num  567 369 665 412 166 ...
##  $ 4-Hydroxyphenylacetate    : num  120.3 432.7 292.9 214.9 97.5 ...
##  $ Acetate                   : num  126.5 212.7 314.2 37.3 407.5 ...
##  $ Acetone                   : num  9.49 11.82 4.44 206.44 44.26 ...
##  $ Adipate                   : num  38.1 327 131.6 144 15 ...
##  $ Alanine                   : num  314 871 464 590 1119 ...
##  $ Asparagine                : num  159.2 157.6 89.1 273.1 42.5 ...
##  $ Betaine                   : num  110 245 117 279 392 ...
##  $ Carnitine                 : num  265.1 120.3 25 200.3 84.8 ...
##  $ Citrate                   : num  3714 2618 863 13630 854 ...
##  $ Creatine                  : num  196.4 212.7 221.4 85.6 105.6 ...
##  $ Creatinine                : num  16482 15835 24588 20952 6768 ...
##  $ Dimethylamine             : num  633 608 735 1064 242 ...
##  $ Ethanolamine              : num  645 488 407 821 365 ...
##  $ Formate                   : num  441 252 250 469 114 ...
##  $ Fucose                    : num  337 198.3 186.8 407.5 26.1 ...
##  $ Fumarate                  : num  7.69 18.92 7.1 96.54 19.69 ...
##  $ Glucose                   : num  395 8691 1353 863 6836 ...
##  $ Glutamine                 : num  871 602 302 1686 433 ...
##  $ Glycine                   : num  2039 1108 620 5064 395 ...
##  $ Glycolate                 : num  685.4 652 141.2 70.8 26.6 ...
##  $ Guanidoacetate            : num  154 110 183 103 53 ...
##  $ Hippurate                 : num  4582 1737 4316 757 1153 ...
##  $ Histidine                 : num  925 846 284 1043 327 ...
##  $ Hypoxanthine              : num  97.5 82.3 114.4 223.6 66.7 ...
##  $ Isoleucine                : num  5.58 8.17 9.3 37.71 40.04 ...
##  $ Lactate                   : num  107 369 750 369 3641 ...
##  $ Leucine                   : num  42.1 77.5 31.5 103.5 101.5 ...
##  $ Lysine                    : num  146.9 284.3 97.5 290 122.7 ...
##  $ Methylamine               : num  52.5 23.6 18.7 48.9 27.9 ...
##  $ Methylguanidine           : num  9.97 7.69 4.66 141.17 5.31 ...
##  $ N,N-Dimethylglycine       : num  23.3 87.4 24.5 40 46.1 ...
```

```
##  $ O-Acetylcarnitine         : num  52.98 50.4 5.58 254.68 45.6 ...
##  $ Pantothenate              : num  25.8 186.8 145.5 42.5 74.4 ...
##  $ Pyroglutamate             : num  437 437 713 567 185 ...
##  $ Pyruvate                  : num  21.1 37 29.4 64.1 12.3 ...
##  $ Quinolinate               : num  165.7 73 192.5 86.5 38.1 ...
##  $ Serine                    : num  284 392 296 1249 206 ...
##  $ Succinate                 : num  154.5 244.7 142.6 144 68.7 ...
##  $ Sucrose                   : num  45.1 459.4 160.8 111 75.2 ...
##  $ Tartrate                  : num  97.51 32.79 16.28 837.15 4.53 ...
##  $ Taurine                   : num  1920 1261 4273 1525 469 ...
##  $ Threonine                 : num  184.9 198.3 110 376.1 64.1 ...
##  $ Trigonelline              : num  943.9 208.5 192.5 992.3 86.5 ...
##  $ Trimethylamine N-oxide    : num  2122 639 1153 1451 172 ...
##  $ Tryptophan                : num  259.8 83.1 82.3 235.1 103.5 ...
##  $ Tyrosine                  : num  290 167.3 60.3 323.8 142.6 ...
##  $ Uracil                    : num  111 47 31.5 30.6 44.3 ...
##  $ Valine                    : num  86.5 110 59.1 102.5 160.8 ...
##  $ Xylose                    : num  72.2 192.5 2164.6 125.2 186.8 ...
##  $ cis-Aconitate             : num  237 334 330 1863 101 ...
##  $ myo-Inositol              : num  135.6 376.1 86.5 247.2 750 ...
##  $ trans-Aconitate           : num  51.9 217 58.6 75.9 98.5 ...
##  $ pi-Methylhistidine        : num  157.6 308 145.5 249.6 84.8 ...
##  $ tau-Methylhistidine       : num  160.8 130.3 83.9 254.7 79.8 ...
```

```
# Mirem si hi ha valors NA (descripció indica que no, però comprovem)
sum(is.na(cachexia_data))
```

```
## [1] 0
```

Observem al fer estudi previ de les dades, que hi ha un total de 77 pacients no aparellats, i que en total tenim 2 grups amb pèrdua muscular: cachexia i control. Per altra banda, la resta d'elements descrits, son tots valors numèrics de la quantitat de metabòlit que es troba en cada pacient o control. Aquests paràmetres son els que caldrà estudiar per saber quins metabòlits tenen alguna relació amb la cachexia. Observem també que no hi ha cap valor NA.

CREACIÓ DE L'OBJECTE *SUMMARIZEDEXPERIMENT*

```
# Convertim la taula a matriu i eliminem les dues primeres columnes que no contenen elements numèr
ics. Transposem la matriu per tal de tenir les mostres com columnes i els metabolits com a files.
cachexia_assay <- as.matrix(t(cachexia_data[, -c(1:2)]))

# Seguidament definim les matrius de dades i metadades del dataset.
colnames(cachexia_assay) <- cachexia_data$`Patient ID`
rownames(cachexia_assay) <- colnames(cachexia_data[, -c(1:2)])
cachexia_metadata <- cachexia_data[, c("Patient ID", "Muscle loss")]
rownames(cachexia_metadata) <- cachexia_metadata$`Patient ID`

# Creem l'objecte Summarized Experiment.
cachexia_se <- SummarizedExperiment(assays = list(counts = cachexia_assay), colData = cachexia_met
adata)
cachexia_se
```

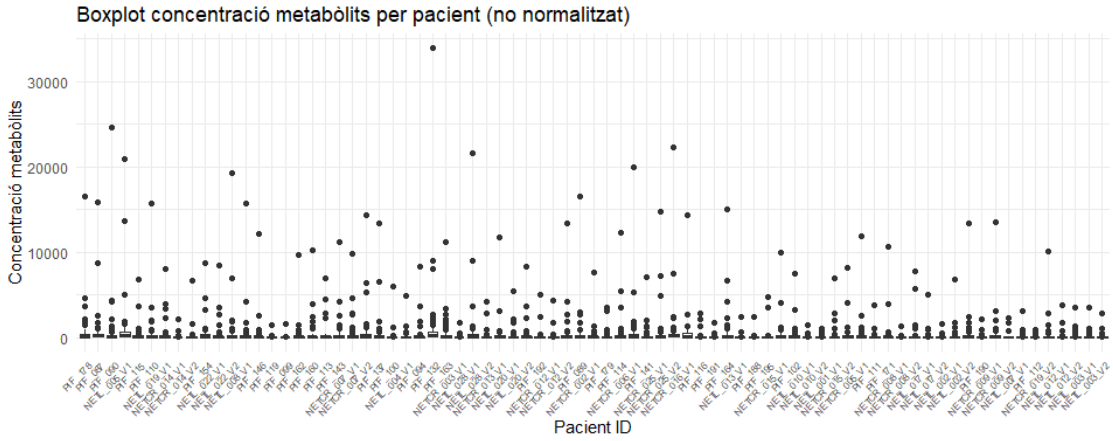
```
## class: SummarizedExperiment
## dim: 63 77
## metadata(0):
## assays(1): counts
## rownames(63): 1,6-Anhydro-beta-D-glucose 1-Methylnicotinamide ...
##   pi-Methylhistidine tau-Methylhistidine
## rowData names(0):
## colnames(77): PIF_178 PIF_087 ... NETL_003_V1 NETL_003_V2
## colData names(2): Patient ID Muscle loss
```

Observem que després de fer els passos addients i obtenir les dades per columnes i files, obtenim un objecte de tipus `SummarizedExperiment` que té 63 files i 77 columnes corresponents a cada pacient.

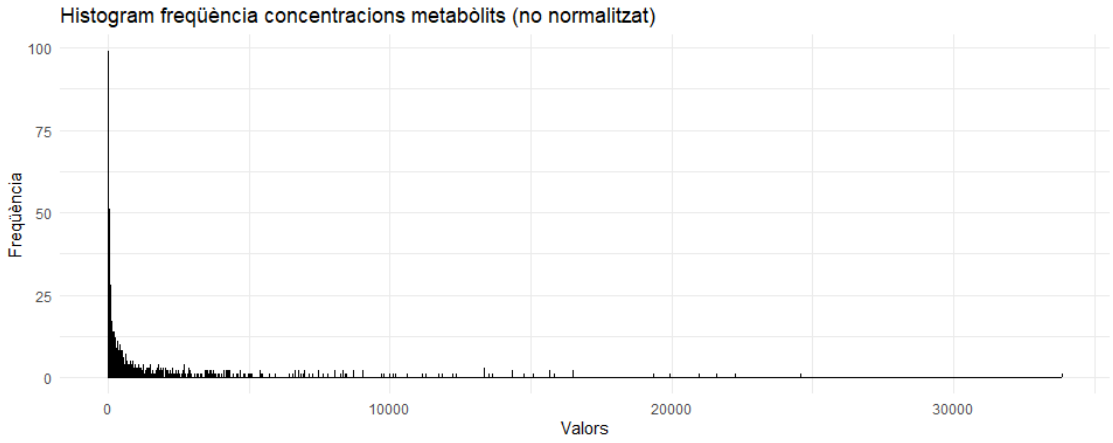
EXPLORACIÓ I CONTROL DE QUALITAT DE LES DADES

Aquest pas és important per assegurar-nos que tot està correctament per poder procedir amb l'anàlisis de les dades.

```
data_long <- reshape2::melt(assay(cachexia_se)) # Cal canviar-ho per poder emprar ggplot2.

# Boxplot per veure distribuciço de metabòlits per pacient sense normalitzar.
ggplot(data_long, aes(x=Var2, y=value))+
  geom_boxplot() +
  theme_minimal() +
  theme(axis.text.x = element_text(hjust = 1, angle = 45, size = 6)) +
  xlab("Pacient ID") + ylab("Concentració metabòlits") + ggtitle("Boxplot concentració metabòlits
per pacient (no normalitzat)")
```



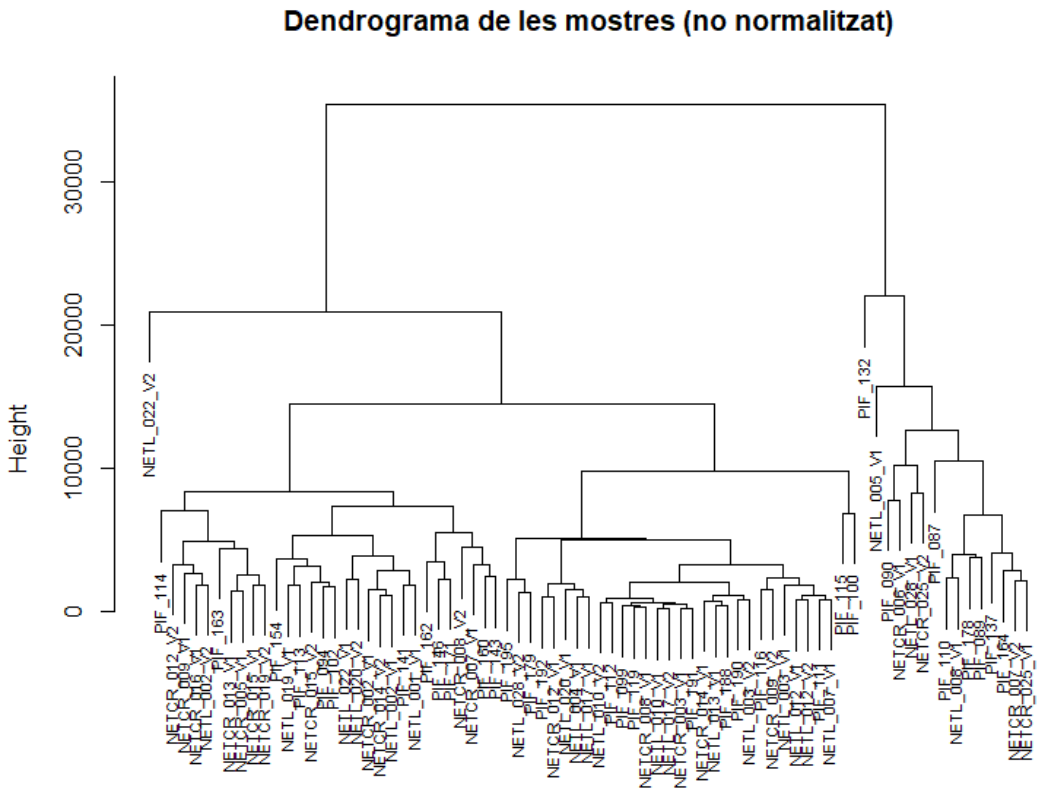

```
# Histograma per veure la distribució de freqüències a nivell d'expressió dels metabòlits sense no
rmalitzar.
ggplot(data_long, aes(x =value )) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black") +
  theme_minimal() +
  xlab("Valors") +
  ylab("Freqüència") +
  ggtitle("Histogram freqüència concentracions metabòlits (no normalitzat)")
```



Observem en el Boxplot la distribució de les dades, i podem veure que existeix variabilitat entre els diferents pacients. En l'histograma observem una distribució altament assimètrica del nivell de metabòlits, amb la majoria de valors concentrats a l'interval més baixi. Això indica que les dades estan bastant esbiaixades, i que per tant, hi ha valors propers a 0 i outliers. Per aquest motiu, caldrà realitzar una normalització de les dades.

També es poden fer altres estudis, com ara l'agrupació jerarquica i estudi per PCAs. Aquest tipus d'estudi ens permet determinar la distància que existeix entre mostres.

```
 # Realitzem agrupació jerarquica
dist_matrix <- dist(t(assay(cachexia_se)), method = "euclidean")
hc_cachexia <- hclust(dist_matrix, method = "complete")
plot(hc_cachexia, main = "Dendrograma de les mostres (no normalitzat)", xlab = "", sub = "",cex=0.
7)
```




```
# Anàlisis per components principals (PCA)
expression_data_cachexia <- assay(cachexia_se)
pca_cachexia <- prcomp(t(expression_data_cachexia), scale = TRUE)
pca_df <- data.frame(pca_cachexia$x)
group <- colData(cachexia_se)$"Muscle loss"

pca_meta <- as.data.frame(colData(cachexia_se))
autoplot(pca_cachexia, data = pca_meta, colour = "Muscle.loss", label = TRUE, label.size = 3) + th
eme_minimal()
```

Observem la distribució segons la distribució jeràrquica i també per PCAs. Observem que no hi ha una tendència clara i diferències en metabòlits entre els grups de cachexia i els controls. Probablement les dades son diferents i la informació es troba esbiaixada, per aquest motiu, serà necessari realitzar una normalització de les dades.

NORMALITZACIÓ DADES (LOGARÍTMICA)

Al tenir unes dades que segueixen una distribució amb gran variabilitat, aplicar una transformació del tipus logarítmica pot ser d'utilitat. Tot i que hi ha altres mètodes, en el nostre cas farem servir aquest tipus de normalització.

```
# Normalització logarítmica
expression_data_cachexia <- assay(cachexia_se) # Extreiem les dades que es troben en SummarizedExp
eriment de cachexia.
cachexia_se_log  <-  log2(expression_data_cachexia + 1)
head(cachexia_se_log)
```

```
##                              PIF_178  PIF_087  PIF_090 NETL_005_V1  PIF_115
## 1,6-Anhydro-beta-D-glucose  5.387156 5.981396 8.084436    7.280492 4.536053
## 1-Methylnicotinamide        6.052459 8.415150 6.038261    5.754353 6.223036
## 2-Aminobutyrate             4.302319 4.660495 3.720278    7.438210 4.056584
## 2-Hydroxyisobutyrate        4.757557 5.415488 6.052459    6.237258 6.408202
## 2-Oxoglutarate              6.180307 6.095080 4.632850   10.229912 5.092546
## 3-Aminoisobutyrate         10.532648 6.879583 3.935460    9.120419 4.938756
##                              PIF_110 NETL_019_V1 NETCR_014_V1 NETCR_014_V2
## 1,6-Anhydro-beta-D-glucose 7.739578    7.251814     5.022368     5.712045
```

```
## 1-Methylnicotinamide       5.036503    5.232661     2.967169     4.966707
## 2-Aminobutyrate            4.275007    3.273516     2.372952     3.094236
## 2-Hydroxyisobutyrate       6.351204    5.443607     3.801159     5.162291
## 2-Oxoglutarate             5.612942    7.811407     4.702103     6.351204
## 3-Aminoisobutyrate         4.206331    5.839456     3.273516     4.247168
##                             PIF_154 NETL_022_V1 NETL_022_V2 NETL_008_V1
## 1,6-Anhydro-beta-D-glucose 6.893848    4.439623    7.008317    5.924575
## 1-Methylnicotinamide       5.740388    7.797078    7.481234    5.697941
## 2-Aminobutyrate            4.356848    4.016140    3.773996    2.967169
## 2-Hydroxyisobutyrate       6.194560    4.896756    4.002703    5.556429
## 2-Oxoglutarate             6.223036    8.487076    6.109152    6.807999
## 3-Aminoisobutyrate         5.881909    6.565140    6.736605    3.182692
##                             PIF_146  PIF_119  PIF_099  PIF_162  PIF_160
## 1,6-Anhydro-beta-D-glucose 6.493775 4.618826 5.401221 9.206843 6.822348
## 1-Methylnicotinamide       5.078524 2.980025 3.273516 4.522307 4.715893
## 2-Aminobutyrate            3.508429 1.641546 1.831877 4.016140 4.043519
## 2-Hydroxyisobutyrate       5.050502 3.145677 3.145677 5.556429 5.612942
## 2-Oxoglutarate             5.064366 3.221877 2.980025 5.078524 4.896756
## 3-Aminoisobutyrate         5.471838 1.989139 2.879706 5.036503 4.097611
##                             PIF_113  PIF_143 NETCR_007_V1 NETCR_007_V2  PIF_137
## 1,6-Anhydro-beta-D-glucose 7.395234 7.524267     7.710875     5.162291 8.386380
## 1-Methylnicotinamide       4.384741 6.522307     5.768714     6.593653 5.204376
## 2-Aminobutyrate            3.853996 3.313246     2.646163     4.618826 3.157044
## 2-Hydroxyisobutyrate       5.008541 6.023921     5.612942     6.109152 5.797013
## 2-Oxoglutarate             5.612942 4.425594     7.739578     8.170676 4.425594
## 3-Aminoisobutyrate         6.322649 4.302319     5.683696     6.722193 6.009661
##                             PIF_100 NETL_004_V1  PIF_094  PIF_132  PIF_163
## 1,6-Anhydro-beta-D-glucose 5.064366    2.513491 6.123501 7.753952 8.256916
## 1-Methylnicotinamide       3.416840    3.600508 3.894333 7.008317 4.743623
## 2-Aminobutyrate            2.292782    5.471838 3.720278 5.022368 4.813012
## 2-Hydroxyisobutyrate       3.587365    4.994580 4.702103 5.120186 5.373300
## 2-Oxoglutarate             7.423914    6.722193 4.868884 6.479457 6.166113
## 3-Aminoisobutyrate         1.989139    5.782671 6.208868 6.023921 6.994014
##                            NETCR_003_V1 NETL_028_V1 NETL_028_V2 NETCR_013_V1
## 1,6-Anhydro-beta-D-glucose     5.274634    5.542258    5.134221     6.765137
## 1-Methylnicotinamide           3.560715    8.890051    6.550901     4.138323
## 2-Aminobutyrate                2.584963    4.111031    3.209453     4.799087
## 2-Hydroxyisobutyrate           3.209453    6.009661    4.138323     5.064366
## 2-Oxoglutarate                 3.666757    7.797078    5.811214     5.995485
## 3-Aminoisobutyrate             3.234195    4.043519    2.134221     4.938756
##                            NETL_020_V1 NETL_020_V2  PIF_192 NETCR_012_V1
## 1,6-Anhydro-beta-D-glucose    3.840967    4.854993 7.151473     3.907852
## 1-Methylnicotinamide          5.697941    6.351204 6.109152     5.556429
## 2-Aminobutyrate               1.970854    4.070389 5.387156     4.910733
## 2-Hydroxyisobutyrate          5.387156    6.038261 3.787641     4.674122
## 2-Oxoglutarate                5.584662    6.479457 4.757557     6.023921
## 3-Aminoisobutyrate            4.549669    3.666757 4.508429     3.814550
##                            NETCR_012_V2  PIF_089 NETCR_002_V1  PIF_179  PIF_114
## 1,6-Anhydro-beta-D-glucose     7.940695 6.965438     7.151473 5.176323 9.422906
## 1-Methylnicotinamide           6.879583 6.365448     4.882643 4.785551 5.218394
## 2-Aminobutyrate                5.358959 5.811214     4.412104 2.634593 5.064366
## 2-Hydroxyisobutyrate           5.967169 6.166113     3.935460 4.966707 6.436795
## 2-Oxoglutarate                 7.452530 6.550901     6.622198 3.068671 4.702103
## 3-Aminoisobutyrate             5.768714 9.134837     3.234195 3.234195 7.538616
##                            NETCR_006_V1  PIF_141 NETCR_025_V1 NETCR_025_V2
## 1,6-Anhydro-beta-D-glucose     8.127530 4.070389     4.952334     4.165912
## 1-Methylnicotinamide           5.373300 4.618826     6.607922     6.850874
## 2-Aminobutyrate                5.811214 4.247168     2.916477     1.819668
## 2-Hydroxyisobutyrate           5.712045 5.260778     6.052459     6.294253
## 2-Oxoglutarate                 6.237258 4.480911    10.042521    11.268033
## 3-Aminoisobutyrate             8.472691 4.799087     3.921246     4.356848
##                            NETCR_016_V1  PIF_116  PIF_191   PIF_164 NETL_013_V1
## 1,6-Anhydro-beta-D-glucose     8.199427 4.938756 4.316146  7.008317    5.162291
## 1-Methylnicotinamide           5.881909 6.151981 4.674122 10.013700    3.733354
## 2-Aminobutyrate                7.395234 2.718088 2.100978  3.260026    2.780310
## 2-Hydroxyisobutyrate           6.379725 4.302319 3.522307  6.066520    4.016140
```

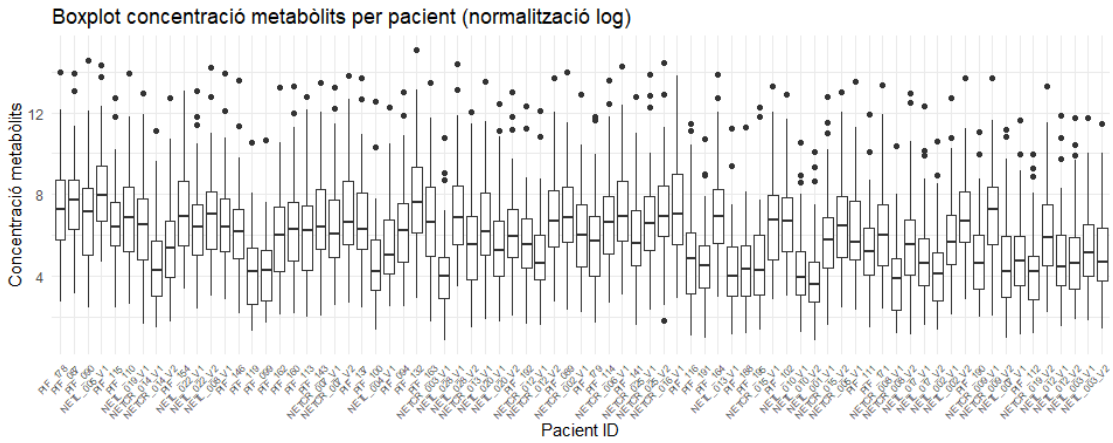
```
## 2-Oxoglutarate                8.875657 2.707083 3.416840  5.288728    4.152183
## 3-Aminoisobutyrate            5.768714 1.851999 4.799087  6.080871    3.614710
##                             PIF_188  PIF_195 NETCR_015_V1  PIF_102 NETL_010_V1
## 1,6-Anhydro-beta-D-glucose 6.052459 4.016140     6.166113 4.715893    5.148527
## 1-Methylnicotinamide       4.646739 6.579391     6.265662 6.679339    3.787641
## 2-Aminobutyrate            2.513491 3.627607     4.563768 3.221877    2.257011
## 2-Hydroxyisobutyrate       4.070389 3.196922     5.953032 5.910493    3.221877
## 2-Oxoglutarate             3.042644 2.731183     7.854494 6.479457    3.935460
## 3-Aminoisobutyrate         2.046142 2.805292     5.768714 4.563768    4.660495
##                            NETL_010_V2 NETL_001_V1 NETCR_015_V2 NETCR_005_V1
## 1,6-Anhydro-beta-D-glucose    4.288359    5.260778     5.120186     4.549669
## 1-Methylnicotinamide          3.234195    5.811214     5.768714     5.811214
## 2-Aminobutyrate               2.257011    3.068671     4.260778     4.439623
## 2-Hydroxyisobutyrate          2.548437    5.218394     5.570766     5.302685
## 2-Oxoglutarate                3.182692    6.265662     6.365448     7.366497
## 3-Aminoisobutyrate            4.577127    3.442280     5.514122     7.696550
##                             PIF_111  PIF_171 NETCR_008_V1 NETCR_008_V2
## 1,6-Anhydro-beta-D-glucose 7.208868 6.023921     5.064366     6.836682
## 1-Methylnicotinamide       3.468583 2.891419     3.907852     5.471838
## 2-Aminobutyrate            2.867896 4.896756     1.989139     2.500802
## 2-Hydroxyisobutyrate       4.854993 4.316146     2.622930     4.813012
## 2-Oxoglutarate             4.646739 6.436795     3.182692     4.549669
## 3-Aminoisobutyrate         3.989139 5.036503     2.805292     4.813012
##                            NETL_017_V1 NETL_017_V2 NETL_002_V1 NETL_002_V2
## 1,6-Anhydro-beta-D-glucose    4.536053    5.570766    7.596041    9.048432
## 1-Methylnicotinamide          4.439623    3.430285    6.779391    7.825786
## 2-Aminobutyrate               3.145677    2.035624    3.132577    3.853996
## 2-Hydroxyisobutyrate          4.370862    3.364572    5.556429    6.565140
## 2-Oxoglutarate                5.302685    3.534809    5.811214    7.854494
## 3-Aminoisobutyrate            3.364572    3.827819    3.005400    3.560715
##                             PIF_190 NETCR_009_V1 NETCR_009_V2 NETL_007_V1
## 1,6-Anhydro-beta-D-glucose 4.896756     7.509933     5.599020    4.084064
## 1-Methylnicotinamide       3.351911     5.627023     3.119356    4.097611
## 2-Aminobutyrate            2.707083     3.313246     2.339137    1.550901
## 2-Hydroxyisobutyrate       4.220330     5.726286     3.364572    4.070389
## 2-Oxoglutarate             3.948601     9.941635     6.052459    4.715893
## 3-Aminoisobutyrate         4.043519     7.639087     5.683696    3.853996
##                             PIF_112 NETCR_019_V2 NETL_012_V1 NETL_012_V2
## 1,6-Anhydro-beta-D-glucose 4.577127     5.176323    4.165912    3.377124
## 1-Methylnicotinamide       3.508429     5.740388    4.070389    3.907852
## 2-Aminobutyrate            1.189034     3.894333    3.522307    2.622930
## 2-Hydroxyisobutyrate       2.718088     5.500165    4.549669    4.618826
## 2-Oxoglutarate             3.247928     6.650765    5.995485    5.584662
## 3-Aminoisobutyrate         3.881665     7.710875    3.574102    3.840967
##                            NETL_003_V1 NETL_003_V2
## 1,6-Anhydro-beta-D-glucose    5.274634    5.302685
## 1-Methylnicotinamide          4.260778    3.760221
## 2-Aminobutyrate               4.757557    4.002703
## 2-Hydroxyisobutyrate          4.002703    3.760221
## 2-Oxoglutarate                4.605257    4.536053
## 3-Aminoisobutyrate            5.106432    4.480911
```

CONTROL DE QUALITAT DESPRÉS DE LA NORMALITZACIÓ (LOGARÍTMICA)

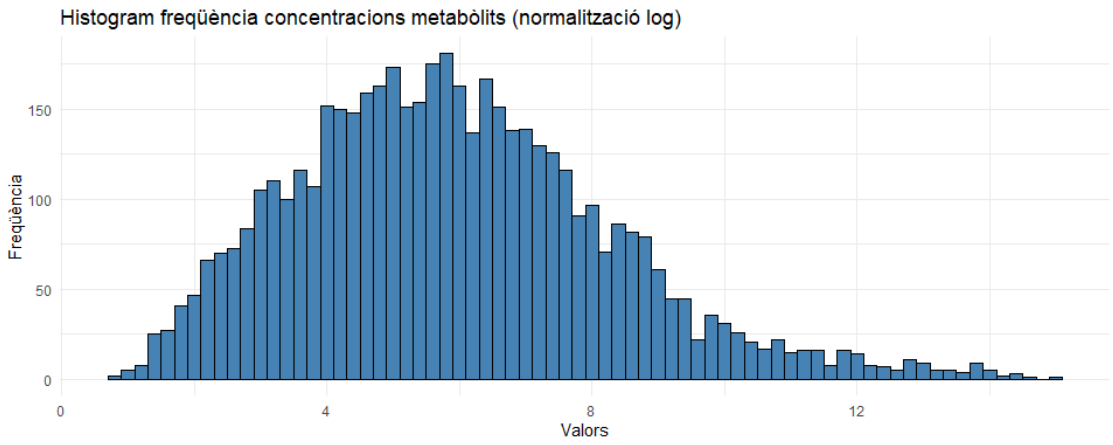
Cal comprovar que ara els gràfics surten millor per poder seguir fent els anàlisis.

```
data_long_log <- reshape2::melt(cachexia_se_log) # Cal canviar-ho per poder emprar ggplot2.

# Boxplot per veure distribuciço de metabòlits per pacient (normalització log).
ggplot(data_long_log, aes(x=Var2, y=value))+
  geom_boxplot() +
  theme_minimal() +
  theme(axis.text.x = element_text(hjust = 1, angle = 45, size = 6)) +
```

```
  xlab("Pacient ID") + ylab("Concentració metabòlits") + ggtitle("Boxplot concentració metabòlits
per pacient (normalització log)")
```

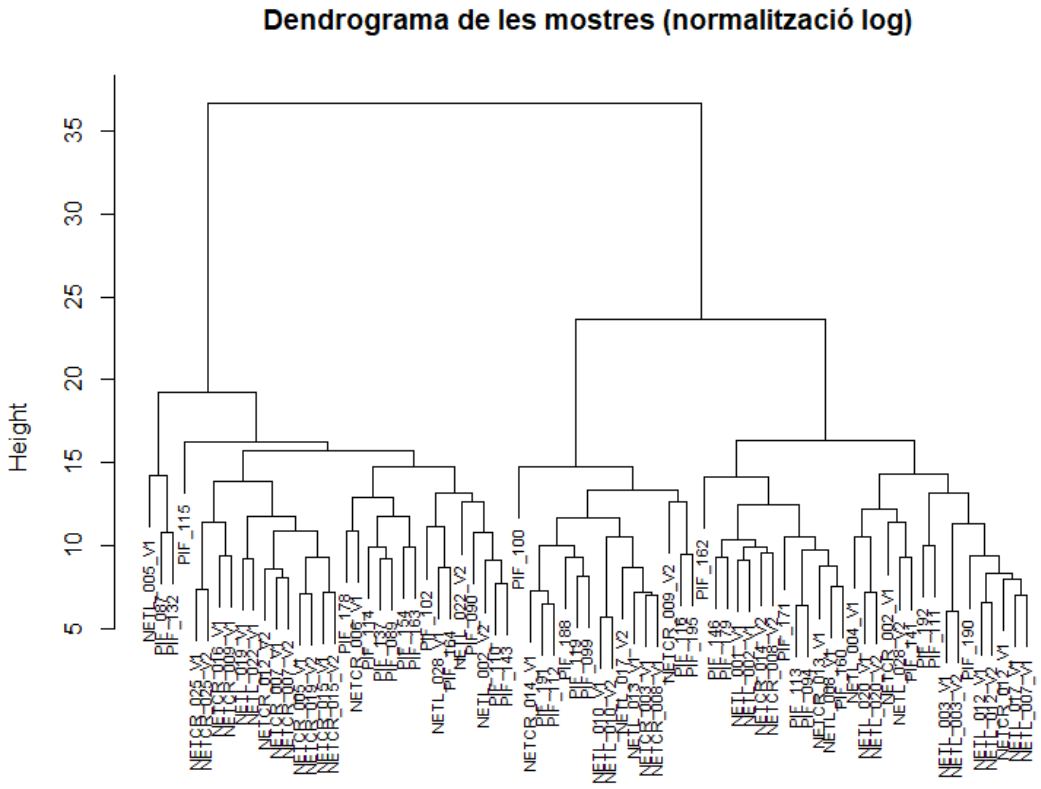



```
# Histograma per veure la distribució de freqüències a nivell d'expressió dels metabòlits (normali
tzació log).
ggplot(data_long_log, aes(x =value )) +
  geom_histogram(binwidth = 0.2, fill = "steelblue", color = "black") +
  theme_minimal() +
  xlab("Valors") +
  ylab("Freqüència") +
  ggtitle("Histogram freqüència concentracions metabòlits (normalització log)")
```



Observem que conseguim una millor distribució de les mostres utilitzant aquest tipus de normalització. Son més homogènies. Podem seguir indagant en l'estructura mitjançant l'agrupació jerarquica i l'anàlisis de PCA.

```
 # Realitzem agrupació jerarquica
dist_matrix_log <- dist(t(cachexia_se_log), method = "euclidean")
hc_cachexia_log <- hclust(dist_matrix_log, method = "complete")
plot(hc_cachexia_log, main = "Dendrograma de les mostres (normalització log)", xlab = "", sub = ""
,cex=0.7)
```




```
# Anàlisis per components principals (PCA)
expression_data_cachexia_log <- cachexia_se_log
pca_cachexia_log <- prcomp(t(expression_data_cachexia_log), scale = TRUE) # Fiquem la transposada.

autoplot(pca_cachexia_log, data = pca_meta, colour = "Muscle.loss", label = TRUE, label.size = 3)
+ theme_minimal()
```

Observem que en tots els casos, es troba distribuït d'una forma més homogènia per tots els gràfics realitzats durant el control.

Al realitzar la normalització logarítmica, el que ens permet és tenir unes dades amb asimetria reduïda i amb control dels valors extrems, i a més a més. Això és útil perquè ens permet comparar patrons d'expressió de metabòlits, fer anàlisis multivariants (PCA, clustering) i fer tests estadístics. Per tant, a partir d'ara, els anàlisis que es realitzaran, seran amb aquestes mostres normalitzades.

ANÀLISIS DE METABÒLITS DIFERENCIALS

Aquests anàlisis ens permeten veure quines diferències hi ha entre els grups de cachexia i els grups control. Ho farem amb el paquet de *limma* de Bioconductor.

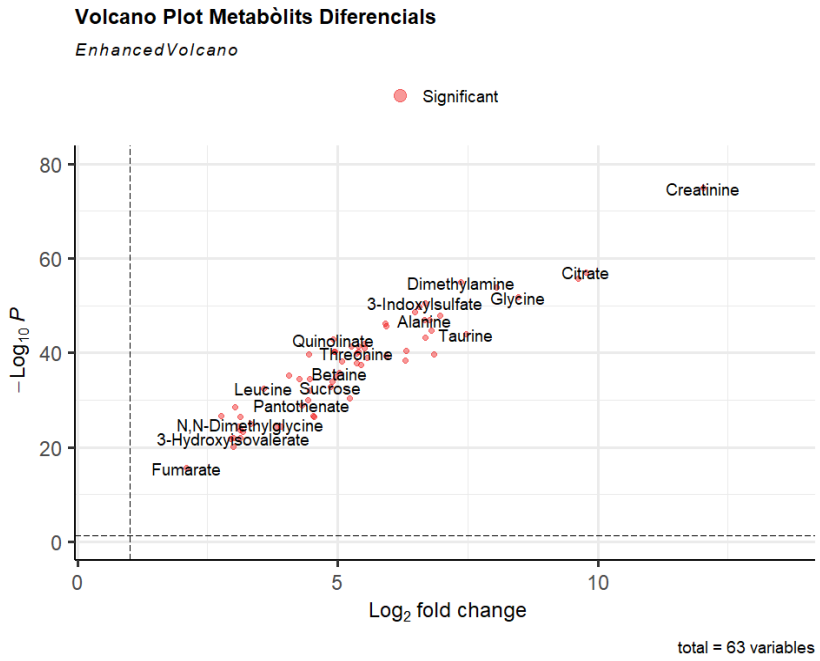
```
# Fem que el paràmetre diferencial sigui el grup de pèrdua muscular.
exp_cachexia <- model.matrix(~ 0 + factor(colData(cachexia_se)$`Muscle loss`))
colnames(exp_cachexia) <- c("Control", "Cachexia")

# Realitzem la prova diferencial amb limma
fit <- lmFit(cachexia_se_log, exp_cachexia)
fit2 <- eBayes(fit)
results_limma <- topTable(fit2, coef = "Cachexia", number = Inf) # Volem veure aquells que tenen u
n valor p significatiu.
head(results_limma)
```

```
##                             logFC   AveExpr        t      P.Value    adj.P.Val
## Creatinine              12.023892 12.640402 55.18811 2.222346e-77 1.400078e-75
## Citrate                  9.769064 10.425267 35.76354 2.994067e-59 9.431312e-58
## Hippurate                9.627429 10.273985 34.47599 9.276207e-58 1.948003e-56
## Dimethylamine            7.368670  7.995535 33.72091 7.309429e-57 1.151235e-55
## Trimethylamine N-oxide   8.053095  8.598352 32.67613 1.356702e-55 1.709445e-54
## Glycine                  8.462143  9.017474 31.00583 1.706269e-53 1.791583e-52
##                                B
## Creatinine              163.0416
## Citrate                 123.8836
## Hippurate               120.5763
## Dimethylamine           118.5830
## Trimethylamine N-oxide  115.7569
## Glycine                 111.0657
```

Inicialment observem que els 6 primers metabòlits que sembla que mostren diferències serien Creatinina, Citrat, Hippurat, Dimethylamina i Glycine. Podem seguir explorant amb altres tipus d'anàlisis com veurem a continuació. De fet, per poder-ho veure de forma més visual, es poden realitzar Volcano plots o bé Heatmaps. Ho fem a continuació.

```
# Fem un Volcano plot i visualitzem quins metabòlits presenten diferències.
EnhancedVolcano(results_limma,
                lab = rownames(results_limma),
                x = 'logFC',
                y = 'adj.P.Val',
                pCutoff = 0.05,
                FCcutoff = 1,
                title = 'Volcano Plot Metabòlits Diferencials',
                colAlpha = 0.4,
                legendLabels = c('NS', 'LogFC', 'p-value', 'Significant'),
                legendPosition = 'top')
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_vline()`).
```

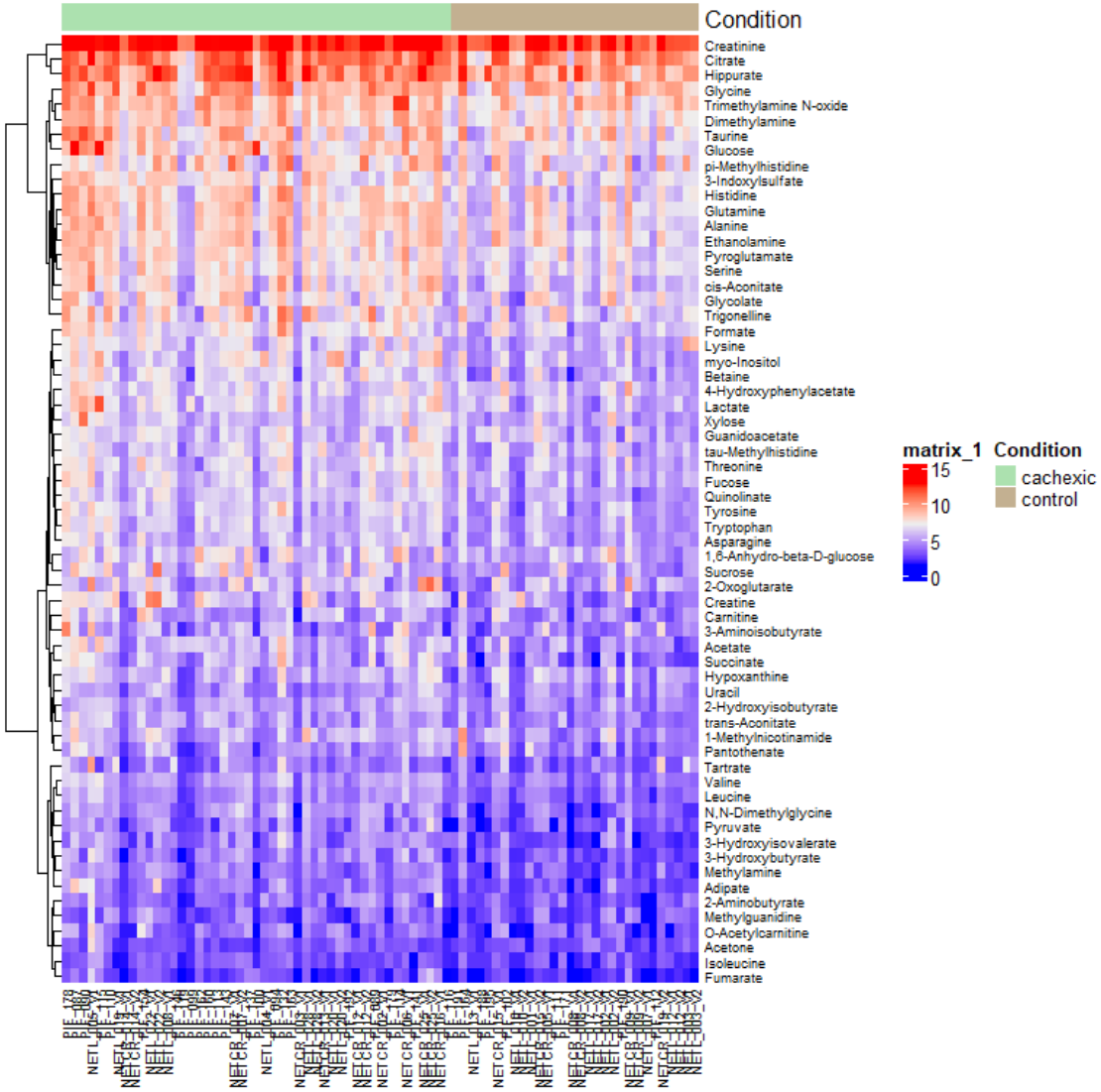


Observem que els que sembla que tinguin una p significativa son els que surten remarcats en color vermell. Veiem alguns exemples escrits, com ara Creatinina (vista anteriorment), citrat, taurina, fumarat, etc.

També podem veure, concretament, si aquests elements que presenten diferències en l'expressió es situen més en pacients amb cachexia o bé en els controls.

Una altra manera de mirar perfils d'expressió és mitjançant la realització d'un Heatmap, com veiem a continuació.

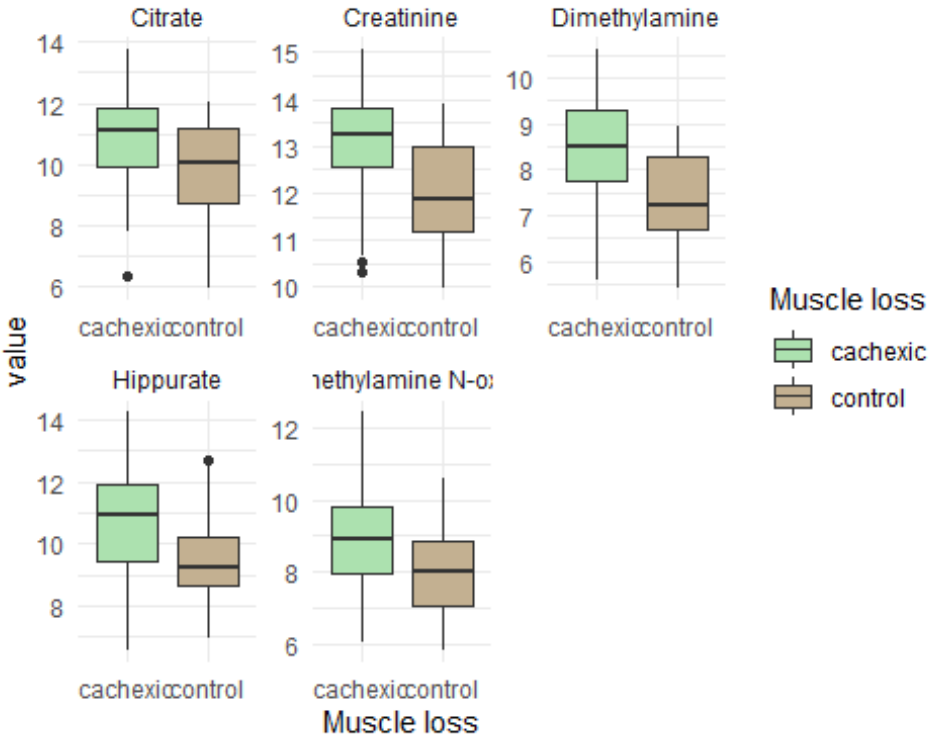
```
# Heatmap Control vs Cachexia
ha = HeatmapAnnotation(Condition = cachexia_data$`Muscle loss`,
                       col = list(Condition = c("cachexic" = "#ACE1AF", "control" ="#C3B091")))
Heatmap(cachexia_se_log,top_annotation = ha, cluster_columns = FALSE,row_names_gp = gpar(fontsize
= 7), column_names_gp = gpar(fontsize = 7))
```



Podem mirar si realitzant boxplots on es compari expressió de cada metabòlit en funció del grup de pacients. Com tenim 63 metabòlits, el que farem serà representar aquells que presenten diferències més grans (Top5) que hem obtingut amb Limma.

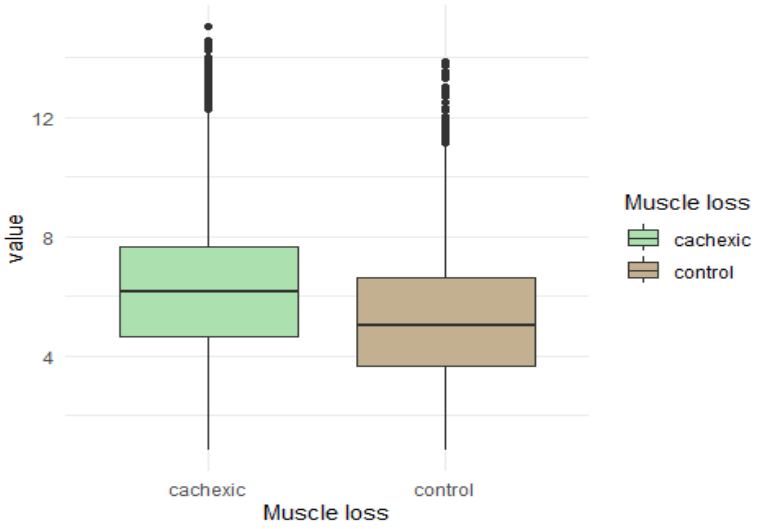
```
# Boxplot TOP 5 metabòlits
cachexia_sig<- melt(cachexia_se_log)

cachexia_sig_meta <- merge(cachexia_sig,cachexia_data[,c(1,2)], by.x="Var2", by.y="Patient ID")

sig_metabolites <- rownames(results_limma[order(results_limma$adj.P.Val),])[1:5]
```

```
ggplot(cachexia_sig_meta[cachexia_sig_meta$Var1 %in% sig_metabolites,], aes(x= `Muscle loss`, y=va
lue, fill=`Muscle loss`)) + geom_boxplot() + scale_fill_manual(values = c("cachexic" = "#ACE1AF",
"control" ="#C3B091"))+ theme_minimal() + facet_wrap(~Var1,scales="free")
```



També es pot realitzar un Boxplot general per confirmar que en general hi ha més metabòlits en orina en els pacients amb caquèxia en comparaicó amb els grups control.

```
# Boxplot General
ggplot(cachexia_sig_meta, aes(x= `Muscle loss`, y=value, fill=`Muscle loss`)) + geom_boxplot() + s
cale_fill_manual(values = c("cachexic" = "#ACE1AF", "control" ="#C3B091")) + theme_minimal()
```

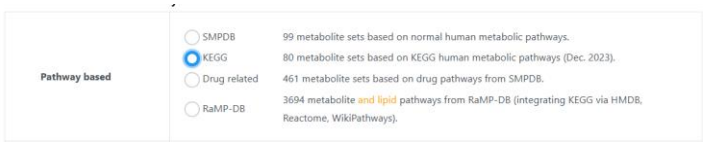


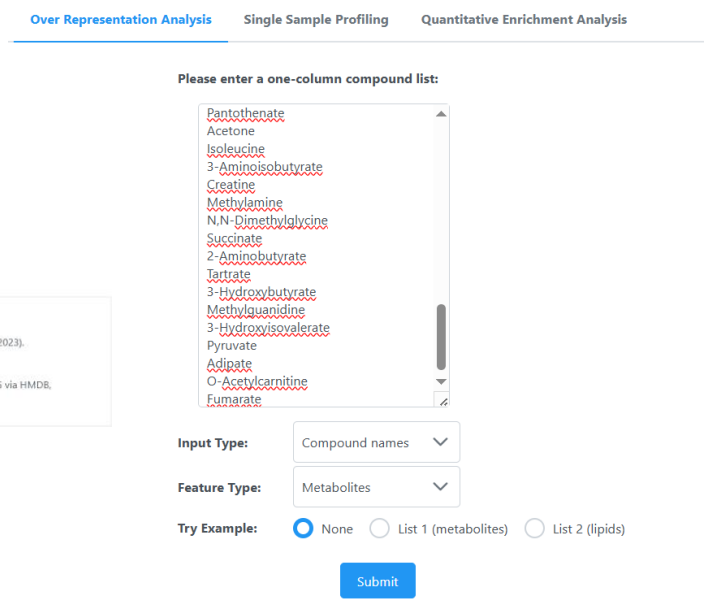
ANÀLISIS DE SIGNIFICÀNCIA BIOLÒGICA

Degut amb incompatibilitats amb R, es realitzar aquest estudi amb la Web de MetaboloAnalyst (https://dev.metaboanalyst.ca/ModuleView.xhtml). En aquesta fem els estudis situats a Annotated Features – Enrichment Analysis.

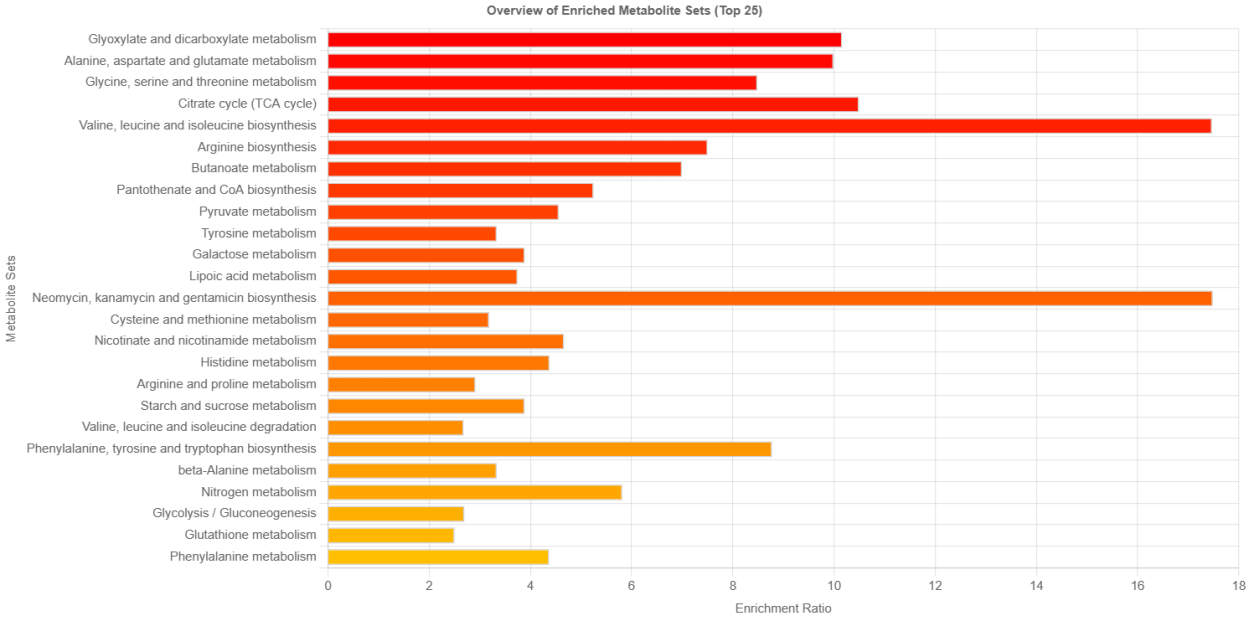
Input Data Type	Available Modules (click on a module to proceed, or scroll down to explore a total of 18 modules including **utilities**)				
LC-MS Spectra (mzML, mzXML or mzData)			Spectra Processing [LC-MS w/wo MS2]		
MS Peaks (peak list or intensity table)		Peak Annotation [MS2-DDA/DIA]	Functional Analysis [LC-MS]	Functional Meta-analysis [LC-MS]	
Generic Format (.csv or .txt table files)	Statistical Analysis [one factor]	Statistical Analysis [metadata table]	Biomarker Analysis	Statistical Meta-analysis	Dose Response Analysis
Annotated Features (metabolite list or table)		Enrichment Analysis	Pathway Analysis	Network Analysis	
Link to Genomics & Phenotypes (metabolite list)			Causal Analysis [Mendelian randomization]		

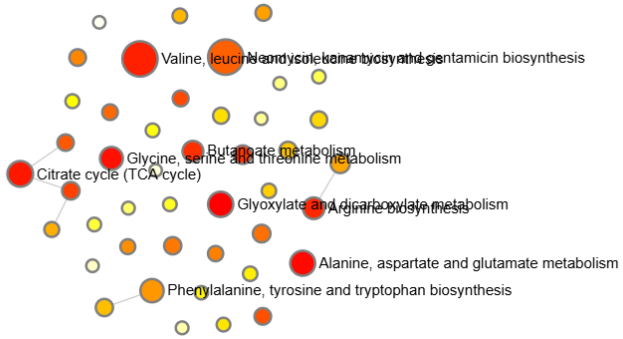
Fiquem la llista de metabòlits augments en caquèxia (en el nostre cas tots) i això ens permetrà veure quins metabòlits están enriquits, i quines vies i xarxes están associades. Podem seleccionar diferents tipus d'estudis, però comencem amb el KEGG basat en vies.

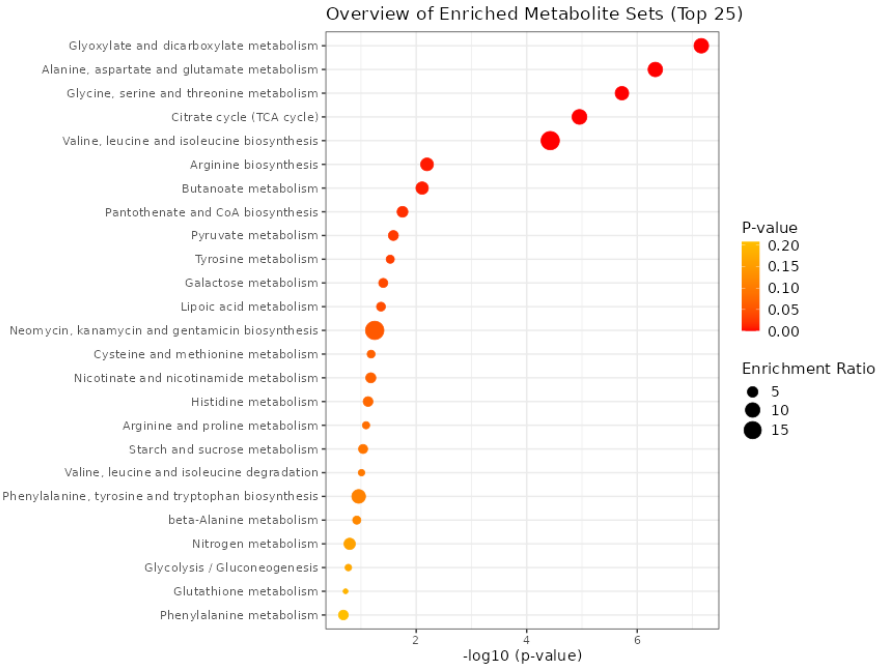




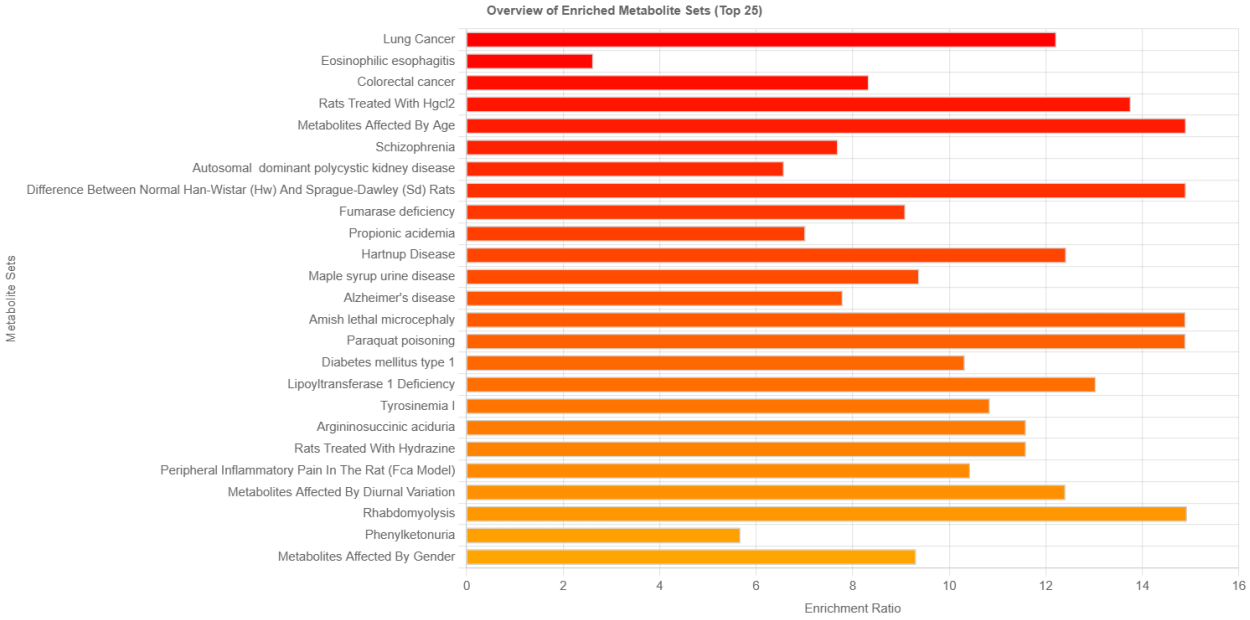
Un cop realitzat l'estudi, ens trobem amb diferents gràfics. Un que ens mostra una Overview de les 25 vies més upregulades i un altre amb la xarxa.

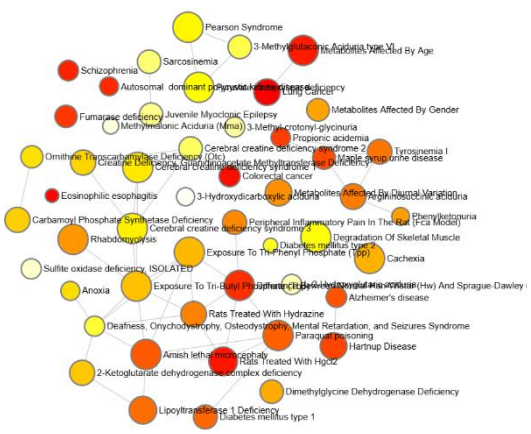

Overview of Enriched Metabolite Sets (Top 25)
Glyoxylate and dicarboxylate metabolism
Alanine, aspartate and glutamate metabolism
Glycine, serine and threonine metabolism
Citrate cycle (TCA cycle)
Valine, leucine and isoleucine biosynthesis
Arginine biosynthesis
Butanoate metabolism
Pantothenate and CoA biosynthesis
Pyruvate metabolism
Tyrosine metabolism
Galactose metabolism
Lipoic acid metabolism
Neomycin, kanamycin and gentamicin biosynthesis
Cysteine and methionine metabolism
Nicotinate and nicotinamide metabolism
Histidine metabolism
Arginine and proline metabolism
Starch and sucrose metabolism
Valine, leucine and isoleucine degradation
Phenylalanine, tyrosine and tryptophan biosynthesis
beta-Alanine metabolism
Nitrogen metabolism
Glycolysis / Gluconeogenesis
Glutathione metabolism
Phenylalanine metabolism
Metabolite Sets
0
2
4
6
8
10
12
14
16
18
Enrichment Ratio

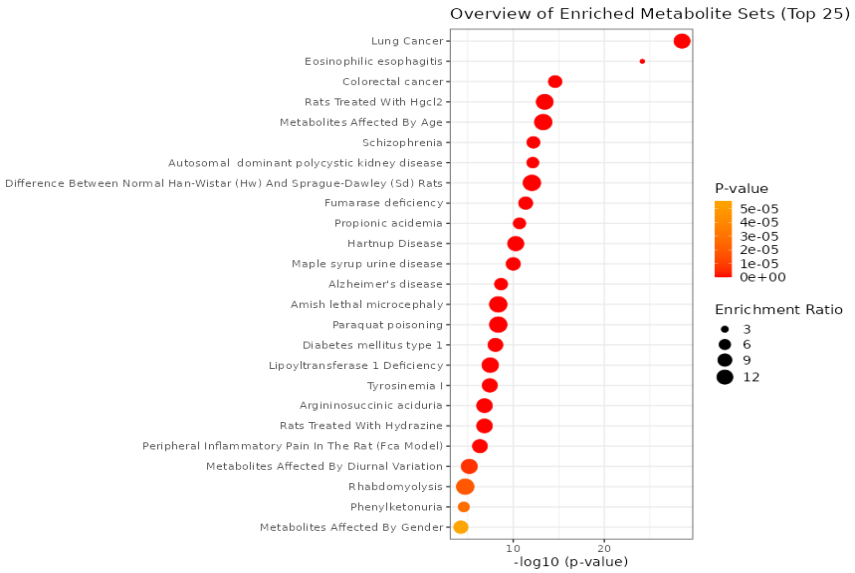

Valine, leucine and isoleucine biosynthesis
Neomycin, kanamycin and gentamicin biosynthesis
Butanoate metabolism
Glycine, serine and threonine metabolism
Citrate cycle (TCA cycle)
Glyoxylate and dicarboxylate metabolism
Arginine biosynthesis
Alanine, aspartate and glutamate metabolism
Phenylalanine, tyrosine and tryptophan biosynthesis


Overview of Enriched Metabolite Sets (Top 25)
Glyoxylate and dicarboxylate metabolism
Alanine, aspartate and glutamate metabolism
Glycine, serine and threonine metabolism
Citrate cycle (TCA cycle)
Valine, leucine and isoleucine biosynthesis
Arginine biosynthesis
Butanoate metabolism
Pantothenate and CoA biosynthesis
Pyruvate metabolism
Tyrosine metabolism
Galactose metabolism
Lipoic acid metabolism
Neomycin, kanamycin and gentamicin biosynthesis
Cysteine and methionine metabolism
Nicotinate and nicotinamide metabolism
Histidine metabolism
Arginine and proline metabolism
Starch and sucrose metabolism
Valine, leucine and isoleucine degradation
Phenylalanine, tyrosine and tryptophan biosynthesis
beta-Alanine metabolism
Nitrogen metabolism
Glycolysis / Gluconeogenesis
Glutathione metabolism
Phenylalanine metabolism
2
4
6
-log10 (p-value)
P-value
0.20
0.15
0.10
0.05
0.00
Enrichment Ratio
5
10
15

Un altre estudi que es pot fer, és veure les signatures de la malaltia, estudiant els metabòlits en orina. En aquest cas, enlloc de via, podem veure amb quines malalties es podrien associar amb els mateix tipus de gràfics.







Amb aquest programa, es poden determinar molts més elements, però nosaltres amb el que observem en aquest estudi será suficiente.