ANÀLISI DE DADES ÒMIQUES - PEC1

Ànnia Castillo Niell

PEC 1

TAULA DE CONTINGUTS

INTRODUCCIÓ

Aquesta PEC està estructurada de tal manera que en els primers apartats trobareu explicat el que s'ha realitzat en l'anàlisi de les dades, i al final, per evitar embrutar la visualització, hi podreu trobar el codi de R amb les sortides corresponents i gràfics. L'enllaç al respositori de Github és el següent: https://github.com/acniell/Castillo-Niell-Annia-PEC1.git

RESUM EXECUTIU

S'ha realitzat un anàlisi metabolòmic a una mostra de 77 pacients amb l'objectiu de detectar si hi ha diferències en els nivells de metabòlits en orina entre els pacients caquèctics i un grup control. S'ha aplicat una *Anàlisi de Components Principals* i s'ha avaluat la variabilitat dels 63 metabòlits entre els dos grups.

Gràficament s'observen certes diferències entre els dos grups, però estadísticament significatives a nivell de la segona component principal (PC2). Avaluant les càrregues de PC2, s'han seleccionat els 10 metabòlits amb més pes i s'han trobat diferències estadísticament significatives entre el grup de caquèctics i el control en 9 de 10.

Aquest anàlisi presenta vàries limitacions, com ara manca d'informació dels pacients inclosos, i altres factors que podrien causar una gran variabilitat metabòlica entre els individus. Tot i així, amb els resultats obtinguts, es podria suggerir futures línies d'investigació orientats a identificar biomarcadors fiables i precoços per a la detecció de caquèxia en pacients oncològics.

OBJECTIUS DE L'ESTUDI

Amb el dataset proporcionat de human_cachexia, no tenim massa informació exceptuant que s'han analitzat 77 mostres d'orina determinant el nivell de 63 metabòlits diferents. La principal diferènicia rau en què una

part de les mostres prové de pacients caquèctics i l'altra meitat de pacients control.

Per tant, en el meu cas, he decidit plantejar-me:

- H0: no hi ha diferències en els nivells de metabòlits en orina entre els pacients control i els caquèctics.
- H1: sí que hi ha diferències.

MATERIALS I MÈTODES

S'ha realitzat un anàlisi metabolòmic en una mostra de 77 pacients, que amb el resum facilitat a *Data_catalog* s'entén que una part d'aquests són pacients neoplàsics. La mostra conté 47 caquèctics i 30 de control i no tenim més dades al respecte (ni sexe, dades antropomètriques, tipus de neoplàsia, etc). Consta que la mostra de pacients és de Evans et al., 2008 però no he pogut identificar l'article d'on prové. Les mostres recollides són d'orina i s'ha analitzat la presència de 63 metabòlits diferents. El dataset ha estat facilitat per l'assignatura d'*Ànàlisi de Dades Òmiques* de la *Universitat Oberta de Catalunya (UOC)*, que especifica que el dataset prové de *"specmine.datasets"* dels paquets de R.

El conjunt de dades s'ha importat al programari R i s'ha creat l'element *Summarized Experiment* per facilitar la manipulació de dades. En aquest es divideixen les dades en:

- El dataset de metabòlits: que hi ha la concentració de 62 metabòlits.

- Les metadades que inclouen: un identificador del pacient i la variable de pèrdua muscular muscle_loss; aquesta última ens permet dividir els pacients en dues categories: caquèctics i controls.

- La informació dels noms dels metabòlits .

Una vegada visualitzades gràficament les dades, s'ha aplicar una transformació logarítmica per reduir la variabilitat de les dades i també s'ha aplicat un procés de normalització i estandarització per poder comparar-les i fer l'anàlisi posterior.

Veure Gràfic 1: boxplot comparatiu de cada metabòlit entre caquèctic i control

Veure Gràfic 2: Gràfic de densitat i distribució dels metabòlits

S'ha realitzat una *Anàlisi de Components Principals (PCA)* per explorar la varabilitat de les dades i veure si hi ha diferències entre el grup de caquèctics i control. S'ha estimat el nombre òptim de components principals (s'ha utilitzat la primera component princpal (PC1) i la segona (PC2) finalment) i s'han representat gràficament per veure la seva distribució.

Veure Gràfic 4: Variància explicada per cada PC

Veure Gràfic 5: Representació gràfica de PC1 i PC2

Veure Gràfic 6: Càrregues de PC1 i PC2

Posteriorment s'ha analitzat si hi ha diferències significatives entre els dos grups (caquèctics/controls) amb *test t de Welch*. S'han analitzat els 10 metabòlits amb les càrregues més elevades de PC2, que ha estat on hem trobat diferències significatives entre els dos grups. Finalment S'ha aplicat de nou un *t test de Welch* d'aquests 10 metabòlits per determinar si les diferents concentracions entre els dos grups eren estadísticament significatives.

Veure Gràfic 7: Càrregues dels metabòlits a PC2

RESULTATS

La realització de PCA ha permès veure la variabilitat dels metabòlits entre els dos grups de pacients: caquèctics i controls. La representació gràfica dels PCA sí que mostra certa possible separació entre els dos grups. PC1 explica una proporcio significativa de la variabilitat, amb un colze que gràficament s'observa entre PC1 i PC2. El *test t de Welch* aplicat per veure si hi havia diferències significatives de PC1 i PC2 entre els dos grups, no va ser significativa per PC1 però sí per PC2. Tot i així, entent que treballem amb PC2,

aquestes diferències trobades poden ser molt petites i insuficients per diferenciar de forma efectiva els dos grups. S'han analitzat les càrregues de PC2 per veure els metabòlits amb més influència, i s'ha aplicat un *t test de Welch* per a cada un d'ells, mostrant que 6 dels 10 metabòlits presenten diferències estadísticament significatives. Aquest resultat podria suggerir que alguns metabòlits podrien ser biomarcadors per identificar els pacient amb caquèxia, tenint en compte però les limitacions de l'estudi (veure a continuació).

DISCUSSIÓ I LIMITACIONS I CONCLUSIONS DE L'ESTUDI

Després de l'anàlisi realitzat el que sobretot encara predomina és la variabilitat entre els dos grups de pacients, cosa que dificulta la identificació d'un patró clar, i possiblement evidencia que el procés de caquèxia afecta moltes mes vies que encara no estan identificades, incloses ni valorades en aquest context.

El fet d'haver trobat diferències estadísticament significatives entre els dos grups a nivell de PC2 pot suggerir que hi ha podria haver processos metabòlics que estiguin estretament lligats a la caquèxia.

Per exemple: el succinat és un metabòlit que participa en el cicle de Krebs per a la creació final d'energia (ATP) i aigua. S'ha vist que en pacients en situació de caquèxia poden tenir nivells més elevats. Això també quadraria amb una glucosa més elevada, ja que els pacients amb caquèxia (el pacient estrella és el pacient amb anorèxia) tenen un estat proinflamatori constant que genera una resistència perifèrica a la insulina i que per tant, disminueix la captació de glucosa intracel · lular. Per últim, el lactat també apunta cap a aquesta direcció: és el biomarcador estrella que es fa servir en medicina com a indicador d'un canvi del metabolisme aeròbic a anaeròbic (per múltiples motius). En el pacients caquèctics, amb un augment de la glucosa sanguínia, també hi ha un augment de la glicòlisi anaeròbica, que en conseqüència augmenta els nivells de lactat. Aquests biomarcadors però, són molt genèrics i també es podrien trobar elevats per exemple en un pacient crític sèptic.

Aquests resultats no són definitius i poden estar subjectes a altres factors que actualment potser no tenim controlats: no sabem el tipus de neoplàsia dels pacients (les pancreàtiques són molt més agressives i segurament porten a nivell de caquèxia major), tipus de dieta dels pacients, nivell d'activitat física, règim de tractament oncoespecífic, sexe, etc.

Tot i així, els resultats obtinguts, poden suggerir futures línies d'investigació, ja que si es pogués trobar un patró que ho suggerís de forma precoç, i s'una manera tan poc invasiva com una mostra d'orina, en un futur es podrien engegar programes per a la prevenció de la caquèxia en pacients hematooncològics i així aconseguit millor tolerància a les teràpies antineoplàsiques i millors resultats de remissió.

ANNEX: CODI DE R I GRÀFICS

1. IMPORTAR I CARREGAR LES DADES

Importem el dataset de cachexia. Hem descarregat el paquet de github i ara el pugem localment al nostre entorn

```
getwd() #M'asseguro d'estar trebllant on desitjo.
library(readr)
human_cachexia <- read_csv("2024-Cachexia/human_cachexia.csv")
View(human_cachexia)
```

2. EXPLORACIÓ MÍNIMA ORIENTATIVA

Ara a nem a veure l'aspecte del nostre dataset:

```
head(human_cachexia)
class(human_cachexia)
names(human_cachexia)
str(human_cachexia)
dim(human_cachexia)
```

Veiem que té les dues primeres columnes que són els ID i a quin grup pertany la mostra (caquèctic o no), i posteriorment la resta són quantificació dels diferents metabolomes en cada pacient. Té unes dimencions de 77x65.

3. CREACIÓ DEL SUMMARIZED EXPERIMENT

Per poder crear el nostre SummarizedExperiment ens hem d'assegurar que tenim tot el necessari instal · lat (i per posteriorment):

```
library(GEOquery)
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
suppressWarnings({
    BiocManager::install("SummarizedExperiment")
    BiocManager::install("mixOmics")
})
library(SummarizedExperiment)
library(Biobase)
library(ggplot2)
```

El que farem doncs, amb les dades que tenim, és filtrar que són metadades i què és la matriu de dades, i tot això guardar-ho en un SummarizedExperiment.

```
x<-as.matrix(human_cachexia[, -c(1, 2)]) #elimino les dues primeres columnes.
columnesMeta <- human_cachexia[, 1:2] #creem les metadades
#els noms de les columnes
colnames(columnesMeta) <- c("ID pacient", "Muscle loss")
rownames(x) <- columnesMeta$`ID pacient` #aparellem que la columna d'identificador ara sigui el nom de

#Creem l'element que s'ha demanat, SummarizedExperiment
SumExp <- SummarizedExperiment(
    assays = list(counts = t(x)),  # Transposem perquè les mostres/individus siguin columnes
    colData = DataFrame(columnesMeta), #incloem les metadades
    rowData = DataFrame(Variable = colnames(x))  #els noms dels metabòlits analitzats
)
SumExp
```

```
## 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(1): Variable
## colnames(77): PIF_178 PIF_087 ... NETL_003_V1 NETL_003_V2
## colData names(2): ID.pacient Muscle.loss
```

Ens hem de fixar que de la matriu assay, cada columna és un individu i les files són els metabòlits. Ara cridarem cada element per veure que tot sigui correcte, és a dir, a mode comprovació que s'han creat els elements correctament:

```
#Amago el resultat ja que és un output molt gros.
dataset<-assay(SumExp)
dataset
```

Aquestes són les metadades

```
colData(SumExp)
```

```
## DataFrame with 77 rows and 2 columns
##                ID.pacient Muscle.loss
##               <character> <character>
## PIF_178           PIF_178    cachexic
## PIF_087           PIF_087    cachexic
## PIF_090           PIF_090    cachexic
## NETL_005_V1   NETL_005_V1    cachexic
## PIF_115           PIF_115    cachexic
## ...                   ...         ...
## NETCR_019_V2 NETCR_019_V2     control
## NETL_012_V1   NETL_012_V1     control
## NETL_012_V2   NETL_012_V2     control
## NETL_003_V1   NETL_003_V1     control
## NETL_003_V2   NETL_003_V2     control
```

I la informació de les dades dels metabòlits:

```
rowData(SumExp)
```

```
## DataFrame with 63 rows and 1 column
##                                        Variable
##                                     <character>
## 1,6-Anhydro-beta-D-glucose 1,6-Anhydro-beta-D-g..
## 1-Methylnicotinamide         1-Methylnicotinamide
## 2-Aminobutyrate                   2-Aminobutyrate
## 2-Hydroxyisobutyrate         2-Hydroxyisobutyrate
## 2-Oxoglutarate                     2-Oxoglutarate
## ...                                          ...
## cis-Aconitate                     cis-Aconitate
## myo-Inositol                       myo-Inositol
## trans-Aconitate                 trans-Aconitate
## pi-Methylhistidine           pi-Methylhistidine
## tau-Methylhistidine         tau-Methylhistidine
```

4. EXPLORACIÓ DEL SUMMARIZED EXPERIMENT (SE)

En aquest apartat hi ha l'exploració del SE així com la transformació i normalització de les dades.

Ja ho hem vist força quan hem muntat SummarizedExperiment però anem a veure com s'estructuren, si hi ha missing values, etc.

```
#la dimensió
dim(SumExp)
```

```
## [1] 63 77
```

Tenim 63 files i 77 columnes.

A continuació veiem l'inici de les metadades

```
head(colData(SumExp))
```

```
## DataFrame with 6 rows and 2 columns
##              ID.pacient Muscle.loss
##             <character> <character>
## PIF_178         PIF_178    cachexic
```

```
## PIF_087         PIF_087    cachexic
## PIF_090         PIF_090    cachexic
## NETL_005_V1 NETL_005_V1    cachexic
## PIF_115         PIF_115    cachexic
## PIF_110         PIF_110    cachexic
```

Abans de seguir, mirem si hi ha missing values.

```
any(is.na(dataset))
```

```
## [1] FALSE
```

No n'hi han pel que de moment podem seguir endavant.

Anem a mirar un resum:

```
#amago resultats per gran output
summary(dataset)
sd(dataset)
```

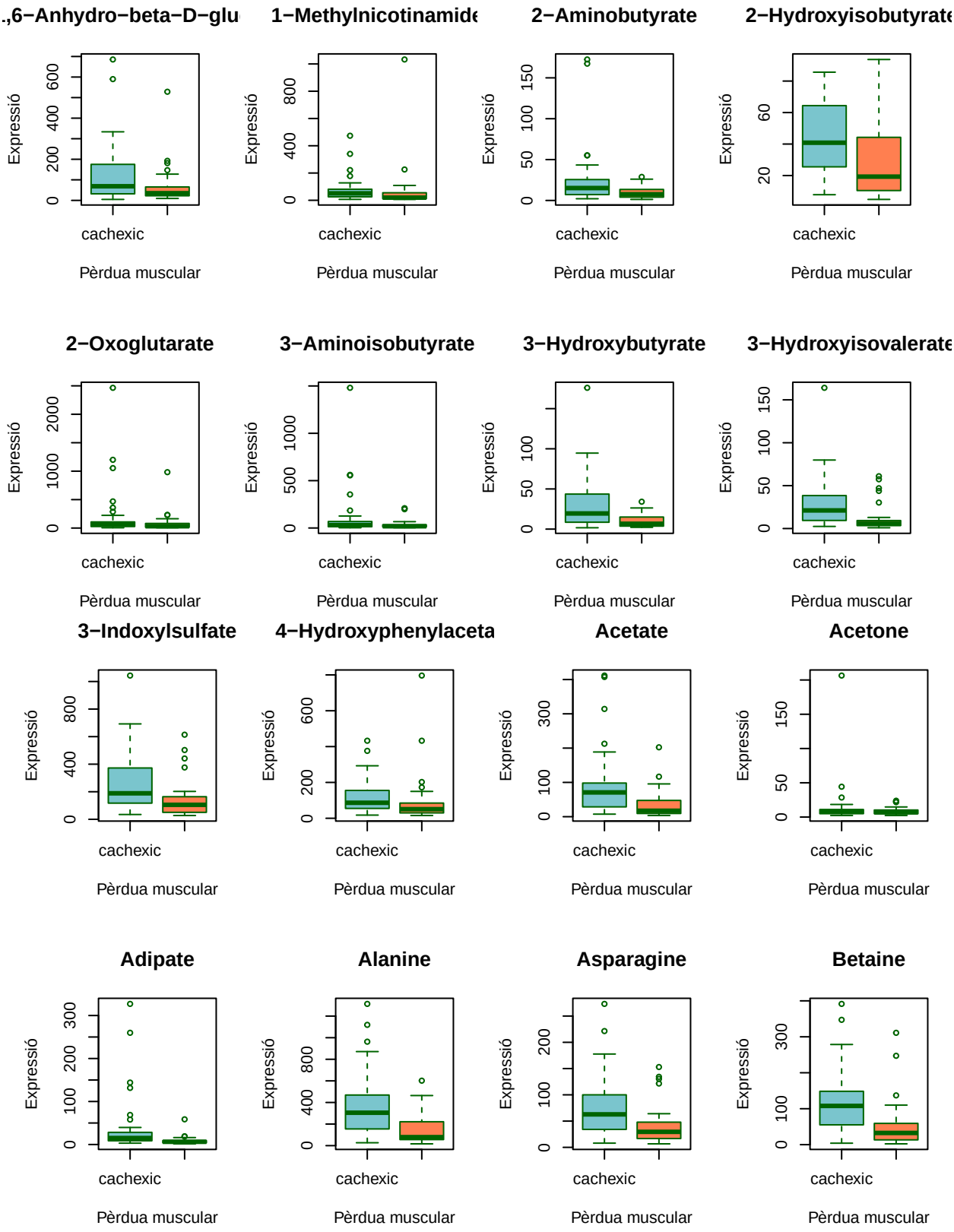
Ens hem de fixar que aquí el resum que ens dona és per cada individu i no cada metabòlit. Si volem veure de cada metabòlit les mateixes dades que l'anterior, podem transposar la matriu.

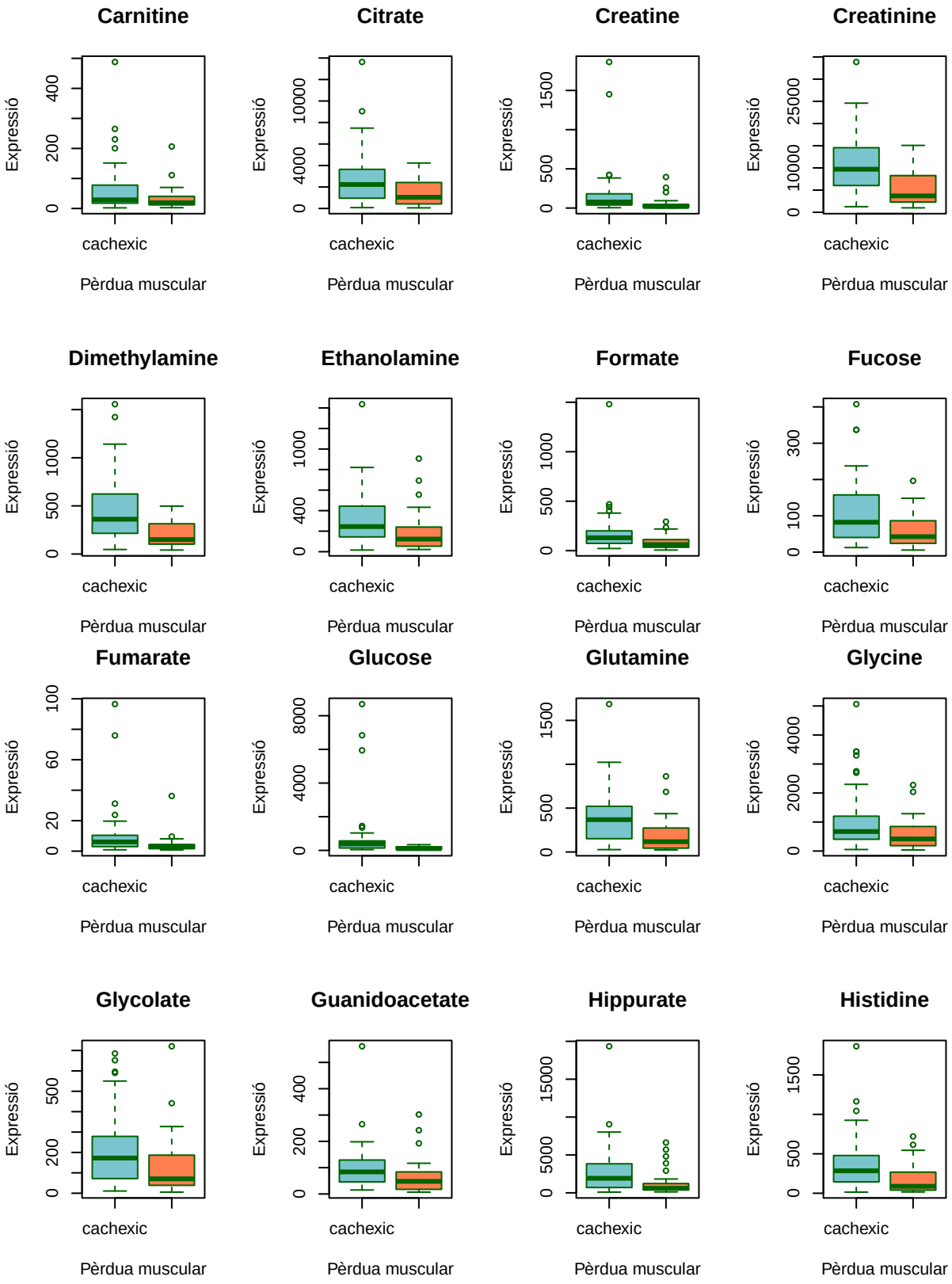
```
#amago resultats per gran output
summary(t(dataset))
sd(t(dataset))
```

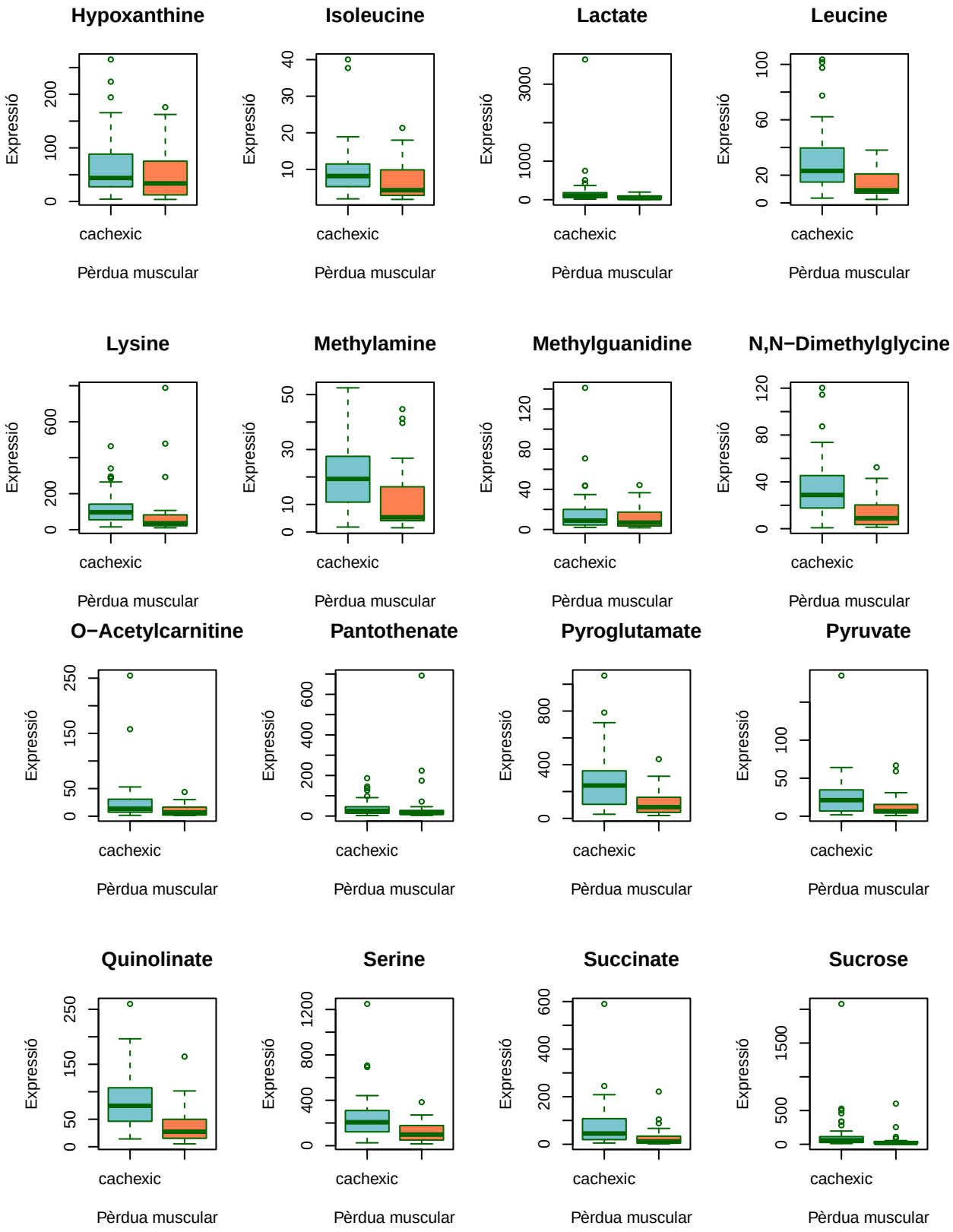
Anem a veure-ho gràficament, comparant els caquètics i control.

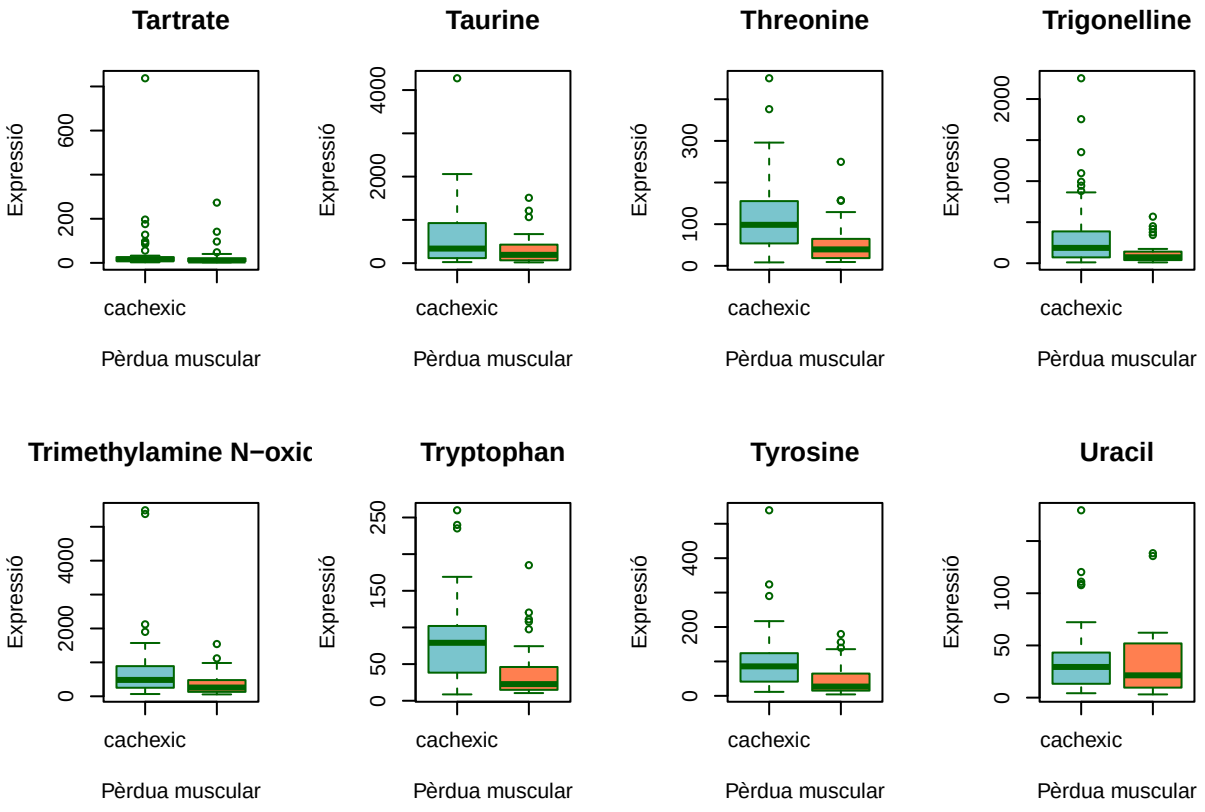
```
par(mfrow=c(2, 4))
for (i in 1:63) {
  boxplot(x[, i] ~ columnesMeta$`Muscle loss`,
          main = colnames(x)[i],  # Nom del metabolit
          xlab = "Pèrdua muscular",
          ylab = "Expressió",
          col = c("cadetblue3", "coral"),
          border ="darkgreen")
}
```

Gràfic 1: boxplot comparatiu de cada metabòlit entre caquèctic i control

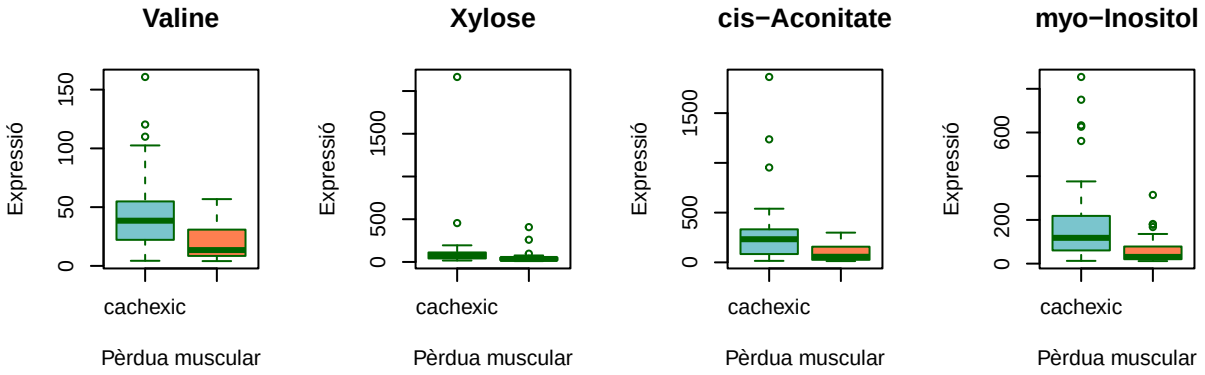

.,6−Anhydro−beta−D−glu
1−Methylnicotinamide
2−Aminobutyrate
2−Hydroxyisobutyrate
2−Oxoglutarate
3−Aminoisobutyrate
3−Hydroxybutyrate
3−Hydroxyisovalerate
3−Indoxylsulfate
4−Hydroxyphenylaceta
Acetate
Acetone
Adipate
Alanine
Asparagine
Betaine
Expressió
cachexic
Pèrdua muscular

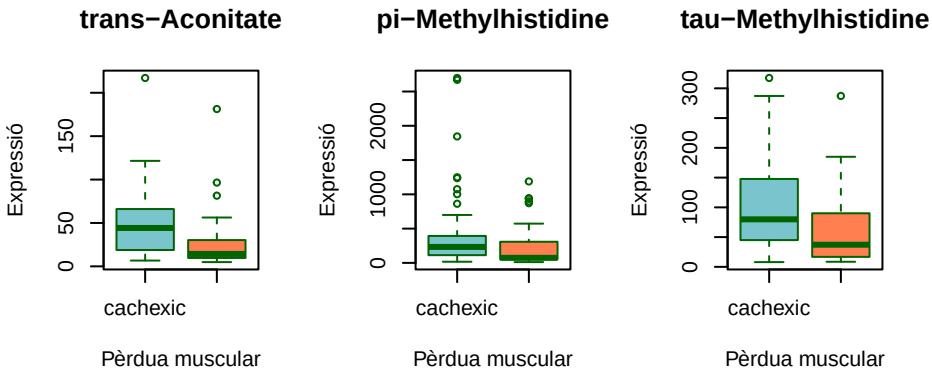

Carnitine
Citrate
Creatine
Creatinine
Dimethylamine
Ethanolamine
Formate
Fucose
Fumarate
Glucose
Glutamine
Glycine
Glycolate
Guanidoacetate
Hippurate
Histidine
Expressió
cachexic
Pèrdua muscular


Hypoxanthine
Isoleucine
Lactate
Leucine
Lysine
Methylamine
Methylguanidine
N,N−Dimethylglycine
O−Acetylcarnitine
Pantothenate
Pyroglutamate
Pyruvate
Quinolinate
Serine
Succinate
Sucrose
Expressió
cachexic
Pèrdua muscular




```
par(mfrow=c(1, 1))
```





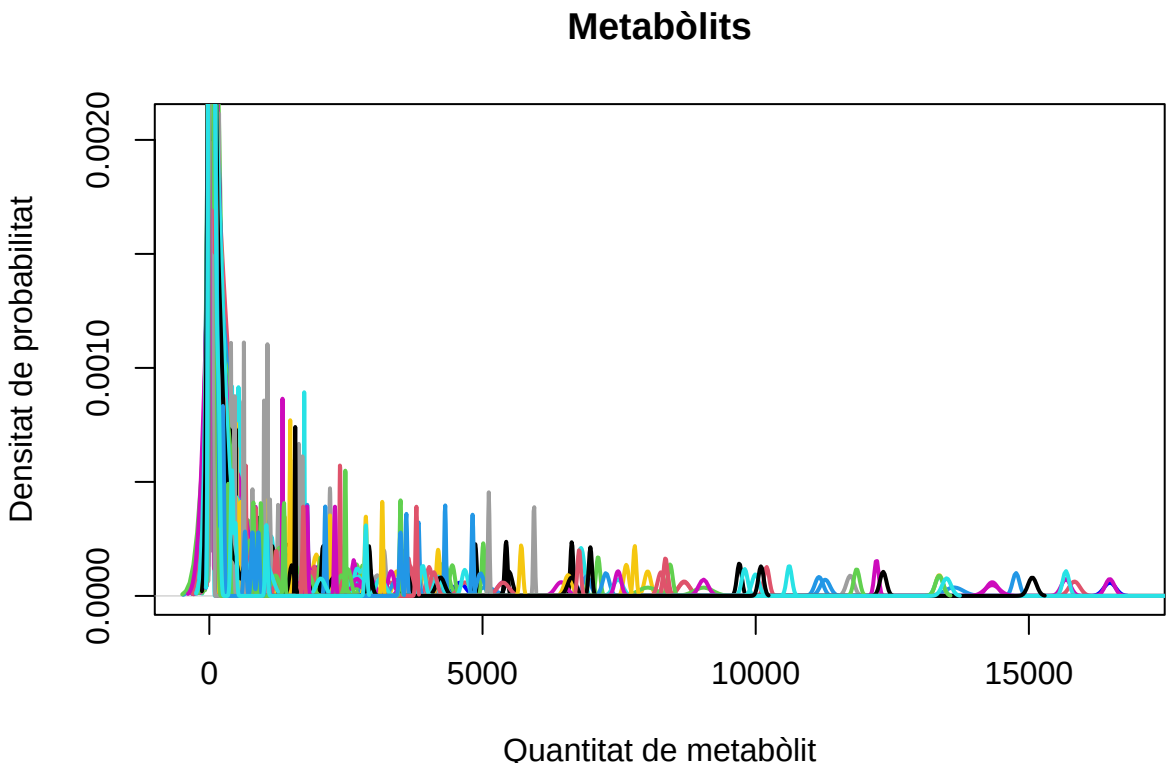
Això és complicat i poc útil perquè he hagut de generar 63 gràfics per veure com es comporta cada metabòlit

en funció de si el pacient està caquèctic o no. Alguns però ja mostren que no varien entre els dos grups, i alguns sí, que és el que anirem a buscar.

Mirem amb un gràfic de densitat com es distribueixen les dades:

```
plot(density(dataset[, 1]), main = "Metabòlits", xlab = "Quantitat de metabòlit", ylab = "Densitat de p
#generem un bucle per representar tots els metabòlits
for (i in 2:ncol(dataset)) {
  lines(density(dataset[, i]), col = i, lwd = 2)  }
```

Gràfic 2: Gràfic de densitat i distribució dels metabòlits



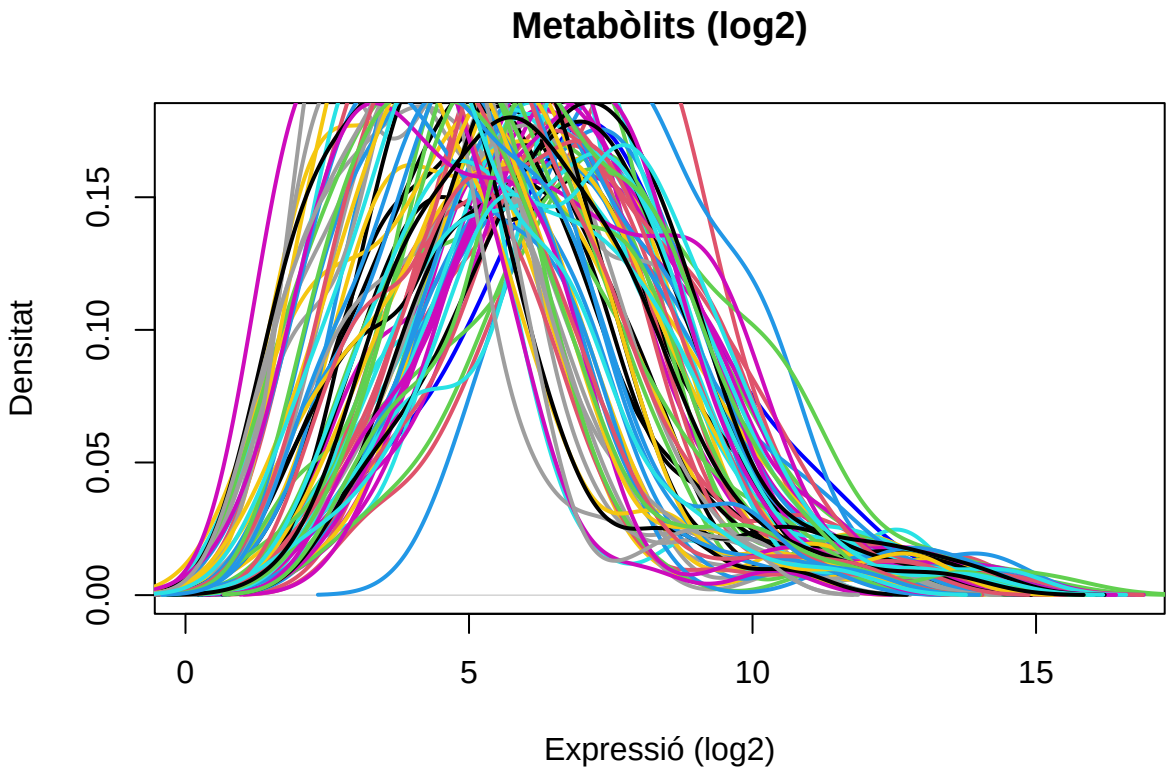
Veiem vàries coses: - La primera és una asimetria clara amb cua a la dreta; serà interessant pensar en una transformació logarítmica. - La segona que tenim un pic al voltant de 0, cosa que suggereix que la majoria de metabòlits tenen concentracions baixes en les mostres, amb només alguns altres que tenen concentracions més elevades.

Tot i així, continua essent un gràfic molt saturat.

Procedim a fer la transformació logarítmica per intentar guanyar simetria i compactar els valors.

```
datasetLog2<- log2(assays(SumExp)$counts + 1)
plot(density(datasetLog2[, 1]), main = "Metabòlits (log2)", xlab = "Expressió (log2)", ylab = "Densitat
#repetim bucle anterior
for (i in 2:ncol(datasetLog2)) {
  lines(density(datasetLog2[, i]), col = i, lwd = 2)
}
```

Gràfic 3: Gràfic de densitat i distribució dels metabòlits amb la transformació logarítmica



Encara hi ha una cua a la dreta però els valors estan més centrats. Tot i així, és possible que encara hi hagi valors extrems i de cares a l'análisi de PCA podria ser beneficiós una normalització i escala z-score.

```
# Normalització z score
# Aplica la normalització z-score als valors log-transformats
log2_dataset_normalitzat <- scale(datasetLog2, center = TRUE, scale = TRUE)
```

5. ANÀLISI DE PCA

Fins ara hem estat intentant fer representació gràfica i anàlisi de 63 metabòlits diferents (i 77 entrades de cada un), cosa molt farregosa i que ens dificulta fer l'anàlisi de dades. L'anàlisi de PCA en permetrà reduir aquesta gran dimensió de les dades i combinar les variables per generar les components principals, no perdent excessiva informació en el procés. A més, esperem que així poguem identificar diferents patrons segons si pertanyen al grup de control o caquèxia.

```
#té un output molt gran, pel que queda amagat el resultat per no embrutar.
PCA_calcul <- prcomp(t(log2_dataset_normalitzat))
summary(PCA_calcul)
```

```
## Importance of components:
##                           PC1     PC2     PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.2042 0.99752 0.93704 0.8859 0.85115 0.79473 0.77889
## Proportion of Variance 0.1133 0.07772 0.06859 0.0613 0.05659 0.04934 0.04739
## Cumulative Proportion  0.1133 0.19099 0.25957 0.3209 0.37746 0.42680 0.47418
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.75682 0.69166 0.6645 0.62553 0.61280 0.58632 0.56371
## Proportion of Variance 0.04474 0.03737 0.0345 0.03056 0.02933 0.02685 0.02482
## Cumulative Proportion  0.51892 0.55629 0.5908 0.62135 0.65069 0.67754 0.70236
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.54594 0.51376 0.49874 0.47893 0.45729 0.44854 0.42826
## Proportion of Variance 0.02328 0.02062 0.01943 0.01792 0.01633 0.01571 0.01433
```

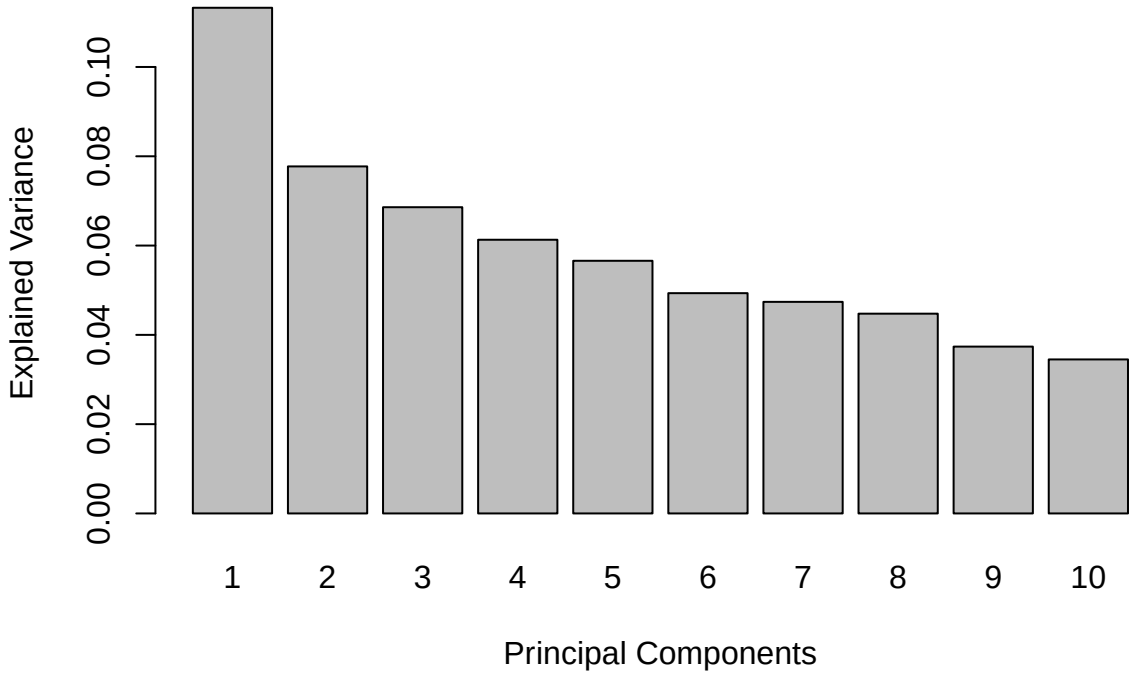
```
## Cumulative Proportion  0.72564 0.74626 0.76569 0.78361 0.79994 0.81566 0.82998
##                           PC22    PC23    PC24   PC25    PC26    PC27    PC28
## Standard deviation     0.40917 0.40110 0.38280 0.3631 0.36093 0.34254 0.32547
## Proportion of Variance 0.01308 0.01257 0.01145 0.0103 0.01018 0.00916 0.00827
## Cumulative Proportion  0.84306 0.85563 0.86707 0.8774 0.88755 0.89671 0.90499
##                           PC29    PC30    PC31    PC32    PC33    PC34    PC35
## Standard deviation     0.31550 0.30933 0.29680 0.29425 0.28251 0.27294 0.26350
## Proportion of Variance 0.00778 0.00747 0.00688 0.00676 0.00623 0.00582 0.00542
## Cumulative Proportion  0.91276 0.92024 0.92712 0.93388 0.94012 0.94593 0.95136
##                           PC36    PC37    PC38    PC39    PC40    PC41   PC42
## Standard deviation     0.24682 0.23932 0.22971 0.21906 0.21086 0.20510 0.1894
## Proportion of Variance 0.00476 0.00447 0.00412 0.00375 0.00347 0.00329 0.0028
## Cumulative Proportion  0.95612 0.96059 0.96471 0.96846 0.97193 0.97522 0.9780
##                           PC43    PC44    PC45    PC46    PC47    PC48   PC49
## Standard deviation     0.18224 0.16969 0.16076 0.15717 0.15037 0.14948 0.1384
## Proportion of Variance 0.00259 0.00225 0.00202 0.00193 0.00177 0.00175 0.0015
## Cumulative Proportion  0.98062 0.98286 0.98488 0.98681 0.98858 0.99032 0.9918
##                           PC50    PC51    PC52    PC53    PC54    PC55    PC56
## Standard deviation     0.13435 0.12554 0.12017 0.11527 0.09870 0.08784 0.07899
## Proportion of Variance 0.00141 0.00123 0.00113 0.00104 0.00076 0.00060 0.00049
## Cumulative Proportion  0.99323 0.99446 0.99559 0.99663 0.99739 0.99799 0.99848
##                           PC57    PC58    PC59   PC60    PC61    PC62      PC63
## Standard deviation     0.07262 0.06937 0.06489 0.0511 0.04054 0.03036 5.329e-16
## Proportion of Variance 0.00041 0.00038 0.00033 0.0002 0.00013 0.00007 0.000e+00
## Cumulative Proportion  0.99889 0.99927 0.99960 0.9998 0.99993 1.00000 1.000e+00
```

```
#amago output per gran sortida
PCA_calcul$x
```

El resultat no és gaire encoratjador i veiem que la PC1 té només el 11% de la variabilitat, PC2 un 7.7%, pel que puja acumulat al 19%, i si incloem la tercera, puja l'acumulada fins al 25.9%. Una altra manera de visualitzar-ho seria fent servir el paquet de mixOmics:

```
library(mixOmics)
PCA_tune<-tune.pca(t(log2_dataset_normalitzat), ncomp=10, scale=FALSE) #triem que es vegin només les pr
plot(PCA_tune)
```

Gràfic 4: Variància explicada per cada PC



Al gràfic veiem que el colze se situa amb les PC1 i PC2.

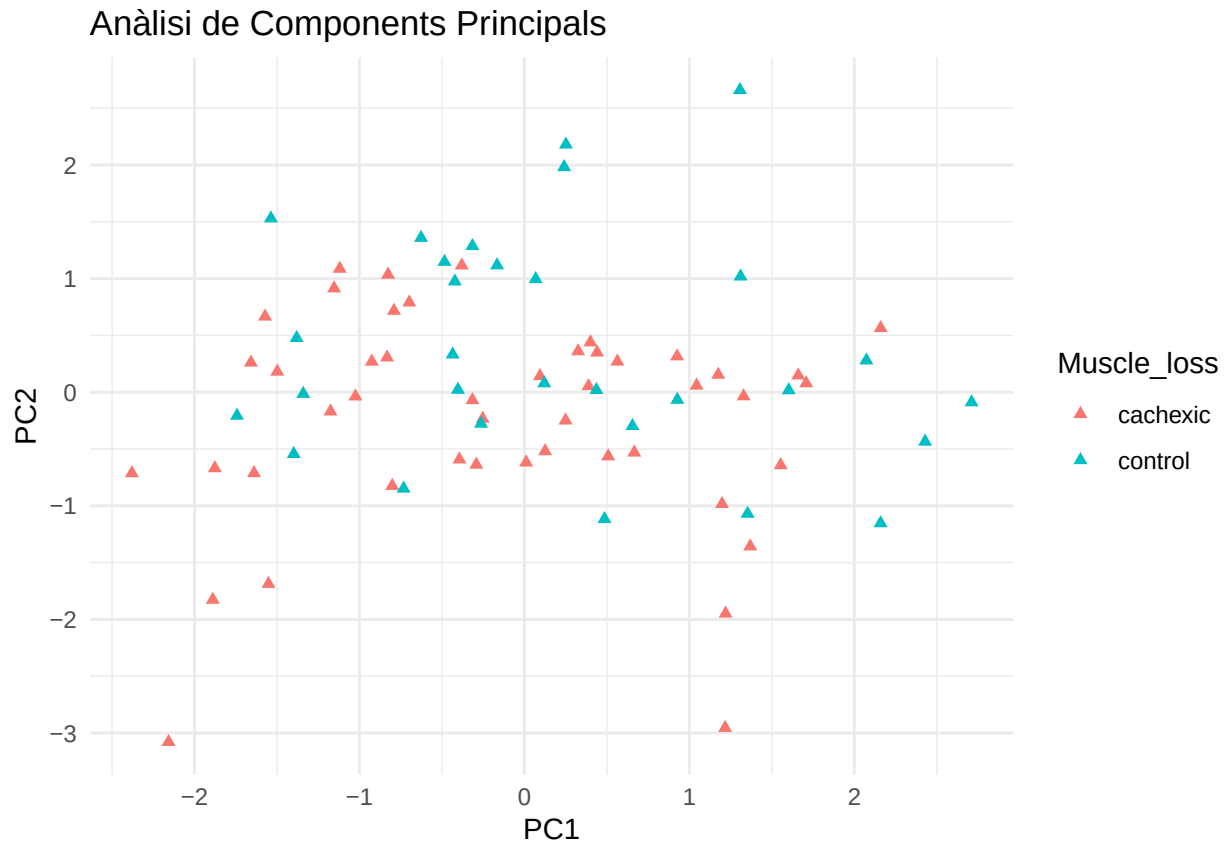
Representem les dues primeres components gràficament.

```
#les coordenades de les components principals
coord <- PCA_calcul$x
#Ho passem a format dataframe
PCA_resultat <- as.data.frame(coord)

#Afegim la variable Muscle_loss de SE a la taula de resultats PCA
PCA_resultat$Muscle_loss <- colData(SumExp)$Muscle.loss

suppressMessages({
  ggplot(PCA_resultat, aes(x = PC1, y = PC2, color = Muscle_loss)) +
    geom_point(shape = 17) +
    theme_minimal() +
    labs(title = "Anàlisi de Components Principals", x = "PC1", y = "PC2")
})
```

Gràfic 5: Representació gràfica de PC1 i PC2



El gràfic podria insinuar que hi hagi diferències marcades entre els dos grups en aquests dos components principals, especialment veient que hi ha uns metabòlits de distribució més superior i d'altres més inferior. O bé també podria ser que no n'hi hagin o bé que hi ha una gran variabialitat encara entre els metabòlits que ens enmascara poder veure diferències entre grups.

Anem a mirar-ho amb t test de Welch

```
t_test_PC1 <- t.test(PC1 ~ Muscle_loss, data = PCA_resultat)
t_test_PC1
```

```
##
##  Welch Two Sample t-test
##
## data:  PC1 by Muscle_loss
## t = -1.3324, df = 60.456, p-value = 0.1877
## alternative hypothesis: true difference in means between group cachexic and group control is not equa
## 95 percent confidence interval:
##  -0.9395442  0.1882101
## sample estimates:
## mean in group cachexic  mean in group control
##             -0.1463638              0.2293033
```

```
t_test_PC2 <- t.test(PC2 ~ Muscle_loss, data = PCA_resultat)
t_test_PC2
```

```
##
##  Welch Two Sample t-test
##
## data:  PC2 by Muscle_loss
```

```
## t = -2.7428, df = 59.282, p-value = 0.008045
## alternative hypothesis: true difference in means between group cachexic and group control is not equa
## 95 percent confidence interval:
##  -1.0736430 -0.1679361
## sample estimates:
## mean in group cachexic  mean in group control
##             -0.2418661              0.3789235
```

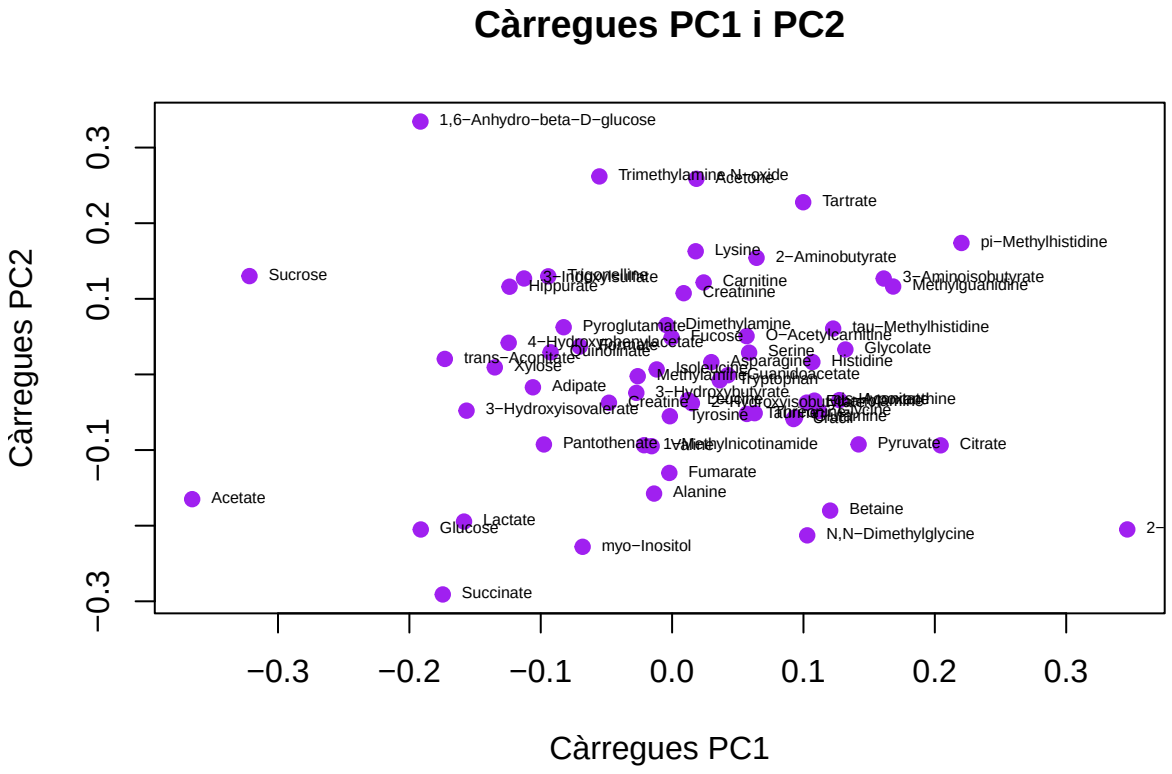
Veiem amb aquests resultats que: - PC1: no s'observa una diferència clarament significativa entre els dos grups de pacients (caquèctics i controls) en quant a PC1. Per tant, en termes de variabilitat podríem dir que els dos grups són similars.

- PC2: sí que veiem una diferència estadísticament significativa entre els dos grups, pel que podríem epnsar que els metabòlits que contribueixen a aquesta component són diferents entre els dos grups. Amb aquest resultat podríem anar a mirar a veure si hi ha algun patró distintiu.

5.1 CÀRREGUES DELS METABÒLITS A PC1 i PC2 Ara veiem de cada metabòlit, quina contribució té a cada component.

```r
loads <- PCA_calcul$rotation
# Gràfic de càrregues per a tots els metabòlits en PC1 i PC2
plot(loads[, 1], loads[, 2],
     xlab = "Càrregues PC1",
     ylab = "Càrregues PC2",
     main = "Càrregues PC1 i PC2",
     pch = 19, col = "purple")
text(loads[, 1], loads[, 2], labels = rownames(loads), pos = 4, cex = 0.5)
```

Gràfic 6: Càrregues de PC1 i PC2



Ara mirem de PC2 els metabòlits més influents i una vegada obtinguts els representarem gràficament.

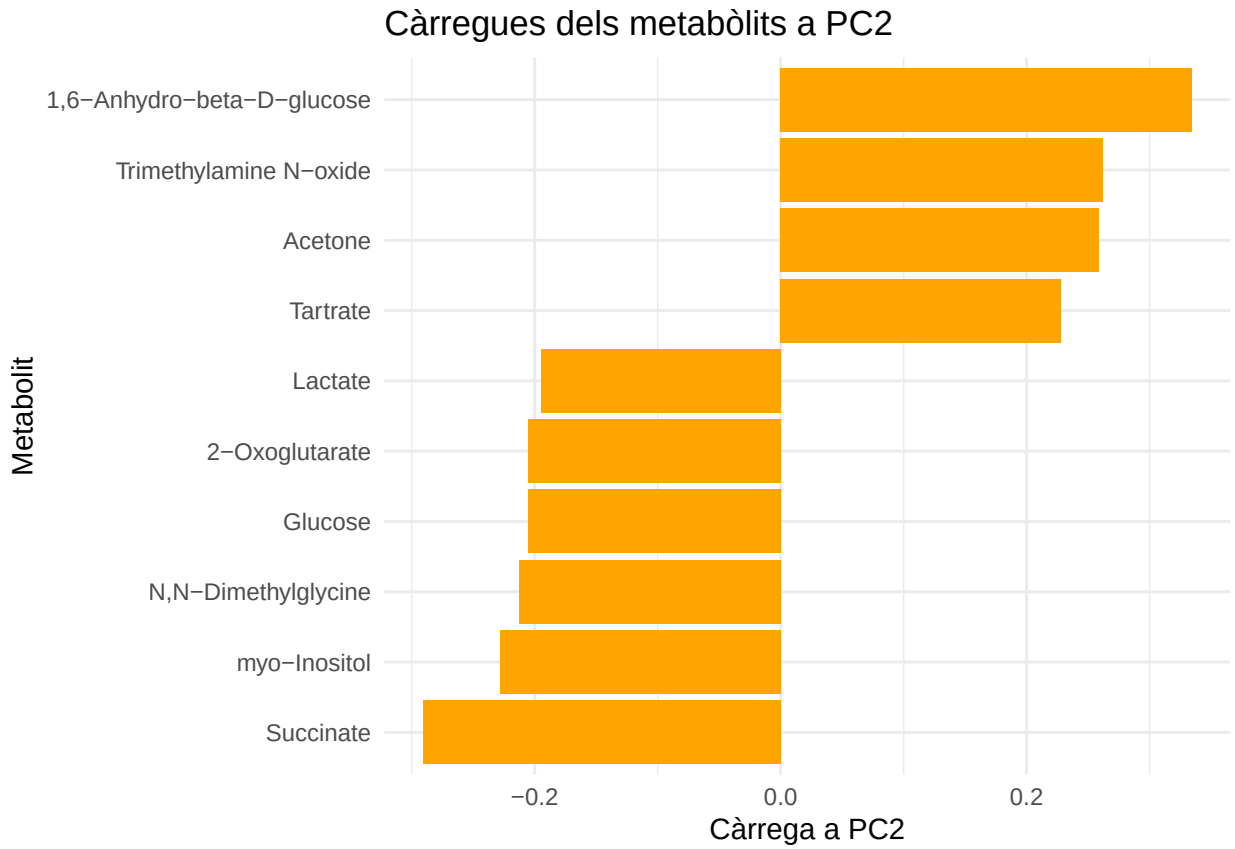
```
# Agafem les càrregues de PC2 i les ordenem segons valors absoluts.
PC2_loads <- sort(abs(loads[, "PC2"]), decreasing = TRUE)
metabolits_PC2 <- names(PC2_loads[1:10]) # Agafem els 10 primers metabòlits més rellevants
data.frame(Metabolit = metabolits_PC2, Carrega_PC2 = loads[metabolits_PC2, "PC2"])
```

```
##                                          Metabolit Carrega_PC2
## 1,6-Anhydro-beta-D-glucose 1,6-Anhydro-beta-D-glucose   0.3344061
## Succinate                                Succinate  -0.2909275
## Trimethylamine N-oxide         Trimethylamine N-oxide   0.2618561
## Acetone                                    Acetone   0.2587005
## myo-Inositol                           myo-Inositol  -0.2279403
## Tartrate                                   Tartrate   0.2277066
## N,N-Dimethylglycine               N,N-Dimethylglycine  -0.2128031
## Glucose                                    Glucose  -0.2050374
## 2-Oxoglutarate                       2-Oxoglutarate  -0.2049703
## Lactate                                    Lactate  -0.1945552
```

```
#agafem el dataframe creat anteriorment
PC2_dataframe <- data.frame(Metabolit = metabolits_PC2, Carrega = loads[metabolits_PC2, "PC2"])

# Gràfic de barres
ggplot(PC2_dataframe, aes(x = reorder(Metabolit, Carrega), y = Carrega)) +
  geom_bar(stat = "identity", fill = "orange") +
  coord_flip() +
  labs(title = "Càrregues dels metabòlits a PC2",
       x = "Metabolit",
       y = "Càrrega a PC2") +
  theme_minimal()
```

Gràfic 7: Càrregues dels metabòlits a PC2



Ara que tenim els 10 metabòlits més influents a PC2, mirarem un per un si hi ha diferències estadísticament significatives entre els dos grups: caquèctics i control. De nou, apliquem un t test per veure-ho:

```
# Ajuntem els metabòlits i la variable muscle_loss
metabolit_data <- data.frame(Muscle_loss = columnesMeta$`Muscle loss`)
#obrim un bucle per crear els test t a tots els metabòlits
for (metabolit in metabolits_PC2) {
  metabolit_data[[metabolit]] <- dataset[metabolit, ]
}

results <- lapply(metabolits_PC2, function(metabolit) {
  t.test(metabolit_data[[metabolit]] ~ metabolit_data$Muscle_loss, var.equal = FALSE)
})
#segon bucle per imprimir els noms dels metabòlits i saber quins estem comparant
for (i in 1:length(metabolits_PC2)) {
  cat(metabolits_PC2[i], ":\n") #el nom del metabòlit i salt de línia
  print(results[[i]])
  cat("\n")
}
```

```
## 1,6-Anhydro-beta-D-glucose :
##
##  Welch Two Sample t-test
##
## data:  metabolit_data[[metabolit]] by metabolit_data$Muscle_loss
## t = 2.1438, df = 74.31, p-value = 0.03532
## alternative hypothesis: true difference in means between group cachexic and group control is not equa
## 95 percent confidence interval:
```

```
##    4.179922 114.187284
## sample estimates:
## mean in group cachexic  mean in group control
##              128.68894               69.50533
##
##
## Succinate :
##
##  Welch Two Sample t-test
##
## data:  metabolit_data[[metabolit]] by metabolit_data$Muscle_loss
## t = 3.0068, df = 69.228, p-value = 0.003677
## alternative hypothesis: true difference in means between group cachexic and group control is not equa
## 95 percent confidence interval:
##  16.75887 82.82701
## sample estimates:
## mean in group cachexic  mean in group control
##               79.62894               29.83600
##
##
## Trimethylamine N-oxide :
##
##  Welch Two Sample t-test
##
## data:  metabolit_data[[metabolit]] by metabolit_data$Muscle_loss
## t = 2.4815, df = 60.708, p-value = 0.01587
## alternative hypothesis: true difference in means between group cachexic and group control is not equa
## 95 percent confidence interval:
##   83.79158 779.55170
## sample estimates:
## mean in group cachexic  mean in group control
##               820.3406               388.6690
##
##
## Acetone :
##
##  Welch Two Sample t-test
##
## data:  metabolit_data[[metabolit]] by metabolit_data$Muscle_loss
## t = 1.1077, df = 51.123, p-value = 0.2732
## alternative hypothesis: true difference in means between group cachexic and group control is not equa
## 95 percent confidence interval:
##  -4.001479 13.854245
## sample estimates:
## mean in group cachexic  mean in group control
##               13.34638                8.42000
##
##
## myo-Inositol :
##
##  Welch Two Sample t-test
##
## data:  metabolit_data[[metabolit]] by metabolit_data$Muscle_loss
## t = 3.7738, df = 62.209, p-value = 0.0003612
```

```
## alternative hypothesis: true difference in means between group cachexic and group control is not equa
## 95 percent confidence interval:
##   56.06222 182.33043
## sample estimates:
## mean in group cachexic  mean in group control
##              181.83766               62.64133
##
##
## Tartrate :
##
##  Welch Two Sample t-test
##
## data:  metabolit_data[[metabolit]] by metabolit_data$Muscle_loss
## t = 0.89137, df = 68.089, p-value = 0.3759
## alternative hypothesis: true difference in means between group cachexic and group control is not equa
## 95 percent confidence interval:
##  -22.98702  60.10439
## sample estimates:
## mean in group cachexic  mean in group control
##               47.23468               28.67600
##
##
## N,N-Dimethylglycine :
##
##  Welch Two Sample t-test
##
## data:  metabolit_data[[metabolit]] by metabolit_data$Muscle_loss
## t = 4.5518, df = 71.926, p-value = 2.114e-05
## alternative hypothesis: true difference in means between group cachexic and group control is not equa
## 95 percent confidence interval:
##  11.74284 30.04340
## sample estimates:
## mean in group cachexic  mean in group control
##               34.48979               13.59667
##
##
## Glucose :
##
##  Welch Two Sample t-test
##
## data:  metabolit_data[[metabolit]] by metabolit_data$Muscle_loss
## t = 2.7164, df = 46.474, p-value = 0.009239
## alternative hypothesis: true difference in means between group cachexic and group control is not equa
## 95 percent confidence interval:
##   177.8635 1194.6584
## sample estimates:
## mean in group cachexic  mean in group control
##               827.2189               140.9580
##
##
## 2-Oxoglutarate :
##
##  Welch Two Sample t-test
##
```

```
## data:  metabolit_data[[metabolit]] by metabolit_data$Muscle_loss
## t = 1.4255, df = 67.941, p-value = 0.1586
## alternative hypothesis: true difference in means between group cachexic and group control is not equa
## 95 percent confidence interval:
##  -39.01949 234.20568
## sample estimates:
## mean in group cachexic  mean in group control
##              183.11043               85.51733
##
##
## Lactate :
##
##  Welch Two Sample t-test
##
## data:  metabolit_data[[metabolit]] by metabolit_data$Muscle_loss
## t = 1.948, df = 47.59, p-value = 0.05733
## alternative hypothesis: true difference in means between group cachexic and group control is not equa
## 95 percent confidence interval:
##   -4.92217 308.68933
## sample estimates:
## mean in group cachexic  mean in group control
##              217.63191               65.74833
```

Veiem que tenim un resultat estadísticament significatiu a 6 dels 10 metabòlits. Si per exemple ens fixem amb el succinat i la glucosa, podríem raonar el perquè poden existir diferències entre els dos grups: El succinat és un metabòlit que participa en el cicle de Krebs per a la creació final d'energia (ATP) i aigua. S'ha vist que en pacients en situació de caquèxia poden tenir nivells més elevats (degut q la necessitat constant de crear energia quan hi ha un dèficit de la ingesta per exemple). Això també quadraria amb una glucosa més elevada, ja que els pacients amb caquèxia (el pacient estrella és la pacient amb anorèxia nerviosa) tenen un estat proinflamatori constant que genera una resistència perifèrica a la insulina i que per tant, disminueix la captació de glucosa intracel · lular. Per últim, el lactat també apunta cap a aquesta direcció: és el biomarcador estrella que es fa servir en medicina com a indicador d'un canvi del metabolisme aeròbic a anaeròbic (per múltiples motius). En el pacients caquèctics, amb un augment de la glucosa sanguínia, també hi ha un augment de la glicòlisi anaeròbica, que en conseqüència augmenta els nivells de lactat.