Anàlisis exploratori de Cachexia

Andrea Zamora Soria

2024-11-11

Resum i objectius de l'estudi:

La caquèxia és una síndrome metabòlica complexa associada a una malaltia subjacent (com el càncer) i es caracteritza per la pèrdua de massa muscular amb o sense pèrdua de massa grassa (Evans et al., 2008). A causa de la seva naturalesa debilitant, la caquèxia és objecte d'estudi en diferents contextos de malaltia crònica, com ara el càncer. En aquest estudi, es van recollir un total de 77 mostres d'orina, de les quals 47 corresponen a pacients amb caquèxia i 30 a pacients de control (dades del paquet de R "specmine.datasets").

L'objectiu principal és explorar aquest dataset de caquèxia per obtenir una visió general de les dades, transformant-les i preparant-les per a l'anàlisi de resultats. Aquesta exploració inclou la identificació de biomarcadors i patrons associats a la síndrome, amb la finalitat de proporcionar informació rellevant que pugui contribuir a una millor comprensió d'aquesta síndrome metabòlica en el context de malalties com el càncer. A més, es definirà un conjunt de limitacions que permetin comprendre les restriccions i possibles biaixos presents en les dades, per tal de contextualitzar adequadament els resultats i orientar futurs estudis sobre aquesta patologia debilitant

Metodes

Per dur a terme aquesta pràctica, s'ha clonat el repositori facilitat a l'enuncia de la PEC1 i a continuació he treballat amb el dataset **2024-Cachexia.**

Un cop seleccionat el dataset, l'arxiu "**Data_Catalog.xlsx"** del repositori proporciona informació rellevant sobre aquest. Sabem que el dataset forma part del paquet de R **"specmine.datasets",** i que les mostres estan dividides en dos grups no aparellats, tal com s'ha esmentat anteriorment. A més, es constata que tots els valors són numèrics i que no contenen cap valor NA.

Per dur a terme la pràctica hem utilitzar el llenguatge de programació R " RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Paquets utilitzats: -Bioconductor: (Bioconductor: Open Source Software for Bioinformatics." Bioconductor. Accessed April 8, 2024. https://www.bioconductor.org/.)

-knitr: A general-purpose package for dynamic report generation in R. R package version 1.37.R Documentation

-Specmine Dataset: https://bioconductor.org/packages/release/data/experiment/html/specmine.datasets.html

-tidyr: Paquet de R dins de tydiverse per facilitar la transformació de dades i facilitar l'anàlisi i la visualització

Resultats:

Per tal de dur a terme l'analisis exploratori de dades, primer de tot hem instalat a R el paquet bioconductor i hem importat les llibreries necessaries (BiocManager i Biobase)

Tal i com ja hem esmentat, l'arxiu .xlsx ens informa que el dataset es troba en el paquet de R "specmine.datasets", el qual descarregat per poder accedir a les dades.

A continuació mostrem les primeres set columnes per comprobar que la descarrega de dades es correcte:

```
##                             PIF_178 PIF_087 PIF_090 NETL_005_V1 PIF_115 PIF_110
## 1.6-Anhydro-beta-D-glucose   40.85   62.18  270.43      154.47   22.20  212.72
## 1-Methylnicotinamide         65.37  340.36   64.72       52.98   73.70   31.82
## 2-Aminobutyrate              18.73   24.29   12.18      172.43   15.64   18.36
## 2-Hydroxyisobutyrate         26.05   41.68   65.37       74.44   83.93   80.64
## 2-Oxoglutarate               71.52   67.36   23.81     1199.91   33.12   47.94
## 3-Aminoisobutyrate         1480.30  116.75   14.30      555.57   29.67   17.46
##                             NETL_019_V1
## 1.6-Anhydro-beta-D-glucose      151.41
## 1-Methylnicotinamide             36.60
## 2-Aminobutyrate                   8.67
## 2-Hydroxyisobutyrate             42.52
## 2-Oxoglutarate                  223.63
## 3-Aminoisobutyrate               56.26
```

I informació general de Cachexia:

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

```
##  num [1:63, 1:77] 40.9 65.4 18.7 26.1 71.5 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:63] "1.6-Anhydro-beta-D-glucose" "1-Methylnicotinamide" "2-Aminobutyrate" "2-Hydroxyis
##   ..$ : chr [1:77] "PIF_178" "PIF_087" "PIF_090" "NETL_005_V1" ...
```

En analitzar l'estructura del dataset que conté les dades de pacients amb caquexia, podem observar que la matriu Cachexia està composta per un total de 77 columnes, cada una corresponent a un pacient o mostres, i 63 files, que representen els valors de concentració de diversos biomarcadors. Aquestes files són fonamentals per a l'estudi, ja que ofereixen una visió detallada dels factors bioquímics i fisiològics que poden estar implicats en la caquexia.

A més, les metadades associades a aquest dataset també contenen 77 files, correspondents als mateixos pacients o mostres, però només inclouen una sola columna anomenada muscle.loss. Aquesta columna (variable qualitativa) indica si cada pacient és del grup de control o del grup cachexic. La informació sobre si un pacient presenta pèrdua muscular significativa (cachexia) o no (control) és essencial per a l'anàlisi i la interpretació dels resultats. Aquesta classificació permet una comparació directa entre els pacients que experimenten caquexia i aquells que no, ajudant a identificar possibles biomarcadors o mecanismes que contribueixin a la pèrdua de massa muscular.

```
## 'data.frame':    77 obs. of  1 variable:
##  $ Muscle.loss: Factor w/ 2 levels "cachexic","control": 1 1 1 1 1 1 1 1 1 1 ...
```

Ara que hem confirmat que les dades s'han descarregat correctament i hem revisat l'estructura tant de les dades com de les metadades, el següent pas és crear un contenidor de tipus SummarizedExperiment. Aquest objecte serà molt útil per gestionar i organitzar les dades i les metadades d'una manera estructurada,ja

que permet emmagatzemar tant les dades quantificades com les metadades associades (informació sobre el dataset, les files i les columnes), facilitant així la seva manipulació i anàlisi en el context de la investigació biomèdica.

Per dur a terme aquesta part de l'exercici, m'he guiat a través de Bioconductor

```
library(SummarizedExperiment)
```

```
# Convertir les dades a una matriu si no ho és ja
cachexia_SE <- SummarizedExperiment(
  assays=list(counts=as.matrix(cachexia$data)),
  rowData = DataFrame(Compound=rownames(cachexia$data)),
  colData=cachexia$metadata)
```

Un cop creem l'objecte SummarizedExperiment, expliquem l'estructura d'e l'objecte'aquesr, visualitzem les primeres files i comprovem les metadades. En el seguent codi:

-assays: Aquí estem passant les dades de concentració de biomarcadors com una matriu per compatibilitat . cachexia$data ha de contenir les dades que desitgem analitzar.

-rowData: Aquí especifico la informació associada a cada fila (en aquest cas, compounds o biomarcadors) com un dataframe. Utilitzem rownames(cachexia$data) per obtenir els noms dels biomarcadors que seran utilitzats com a identificadors.

-colData: Aquesta part és per les metadades, que contenen informació sobre la perdua de pes de cada pacient. En aquest cas, s'espera que cachexia$metadata sigui un dataframe amb la informació adequada.

```
## 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): Compound
## colnames(77): PIF_178 PIF_087 ... NETL_003_V1 NETL_003_V2
## colData names(1): Muscle.loss
```

Analisi exploratori de les dades

1.Dimensions: Com ja s'ha vist anteriorment, aquest primer pas ajuda a assegurar-se que les dimensions de les dades són les esperades, amb el nombre correcte de mostres (pacients) i variables (biomarcadors). Així, es confirma que el conjunt de dades s'ha carregat correctament i es pot avançar amb confiança en l'anàlisi.

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

2.Analisi estadístic: Proporciona una visió inicial del rang de valors, la mitjana, la mediana i altres estadístiques importants per cada biomarcador. Aquesta informació és útil per identificar valors extrems, distribucions inesperades o valors que podrien necessitar un tractament especial. En aquest cas només mostrem les cinc primeres columnes (pacients) perque no s'extengui molt el resultat.

```
##     PIF_178           PIF_087           PIF_090          NETL_005_V1
##  Min.   :   5.58   Min.   :   7.69   Min.   :   4.44   Min.   :  25.03
##  1st Qu.:  52.72   1st Qu.:  78.66   1st Qu.:  31.50   1st Qu.: 102.51
##  Median : 154.47   Median : 208.51   Median : 141.17   Median : 247.15
##  Mean   : 699.86   Mean   : 708.30   Mean   : 771.79   Mean   :1021.28
```

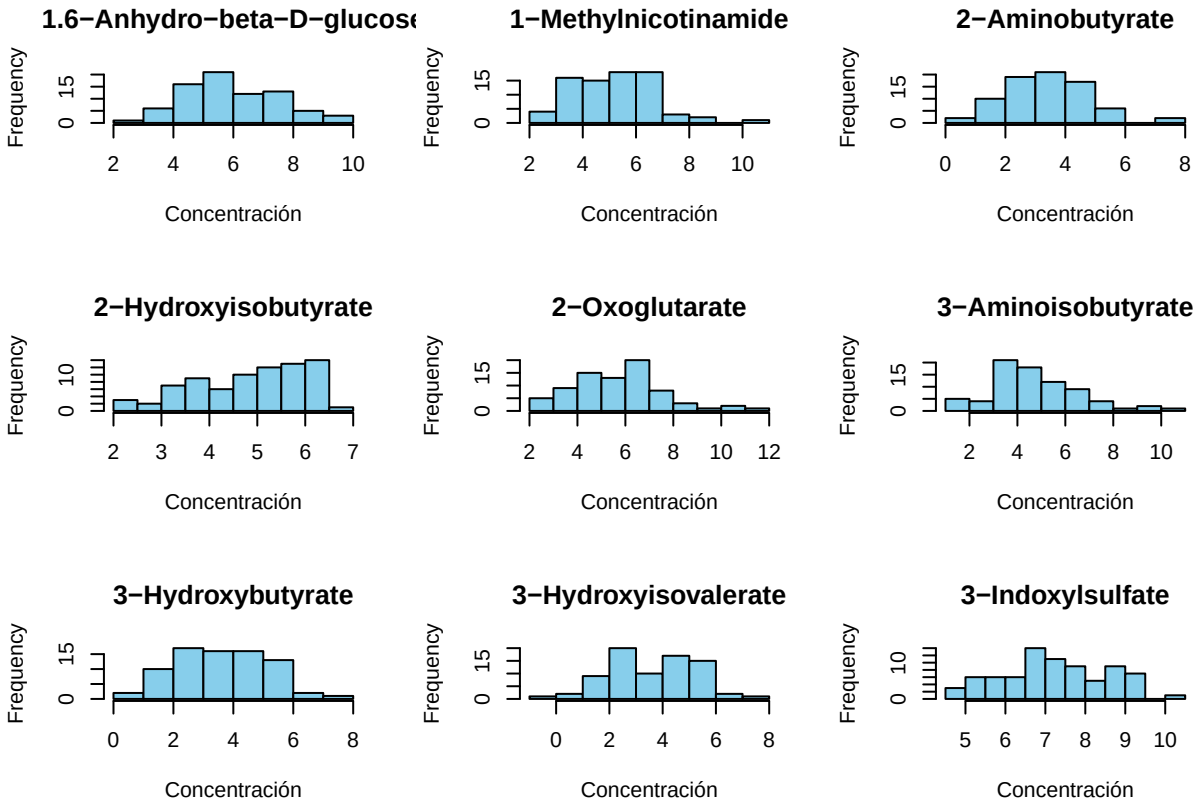
```
##  3rd Qu.:  416.24   3rd Qu.:  412.10   3rd Qu.:  308.03   3rd Qu.:  673.71
##  Max.   :16481.60   Max.   :15835.35   Max.   :24587.66   Max.   :20952.22
##     PIF_115
##  Min.   :   4.53
##  1st Qu.:  44.26
##  Median :  84.77
##  Mean   : 441.22
##  3rd Qu.: 196.62
##  Max.   :6836.29
```

3.**Comprovació de dades mancants:** La presència de valors mancants pot afectar la qualitat de les dades i, per tant, la integritat de les conclusions. Es per això que es molt important assegurar-nos que no existeixen.

```
## Valors mancants al dataset: 0
```

4.**Visualització:** Aquest és essencial per obtenir una visió ràpida dels patrons de dades. Els boxplots mostren la distribució i els possibles valors atípics per cada biomarcador, mentre que els histogrames permeten entendre millor la distribució de cada variable. Això facilita la detecció de possibles valors atípics i desequilibris entre grups.

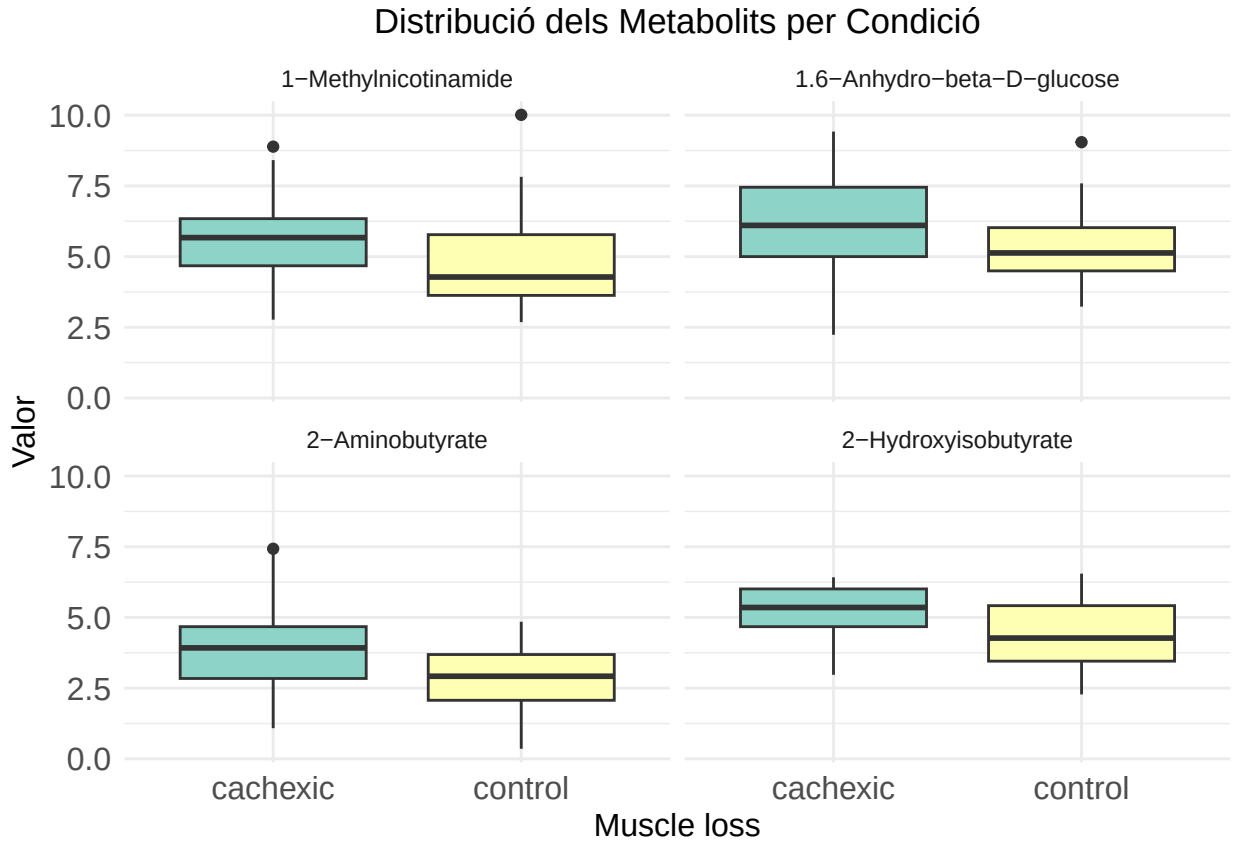
-**Histograma:** en aquest cas transposarem la matriu perque els biomarcadors es presentin com columnes. A continuació visualitzem els 9 primers biomarcadors en un format 3x3 perque sigui més visual. Donat que les dades estan entre el rang 0-33860.35, primer farem una transformació logarítmica (log2).



En aquest gràfic es pot observar que la majoria de les distribucions són força simètriques o lleugerament asimètriques, amb variacions en les concentracions segons el biomarcador. Alguns biomarcadors, "3-Indoxylsulfate," tenen distribucions més concentrades, mentre que altres, com "2-Oxoglutarate," mostren

una major dispersió, amb valors que arriben a concentracions més altes. Aquestes variacions suggereixen diferents nivells de presència entre pacients, proporcionant informació inicial sobre les diferències en perfils metabòlics dins el grup de pacients amb caquèxia.

5.Boxplot: En aquest cas el que busquem comparar la distribució de diversos metabolits (biomarcadors) entre les mostres de caquexia i control. Per fer-ho convertim la matriu a dataframe i seleccionem els metabolits que volem comparar. A continuació creem el gràfic per poder comparar-los. Per això utilitzarem també el paquet tidyr i ggplot:

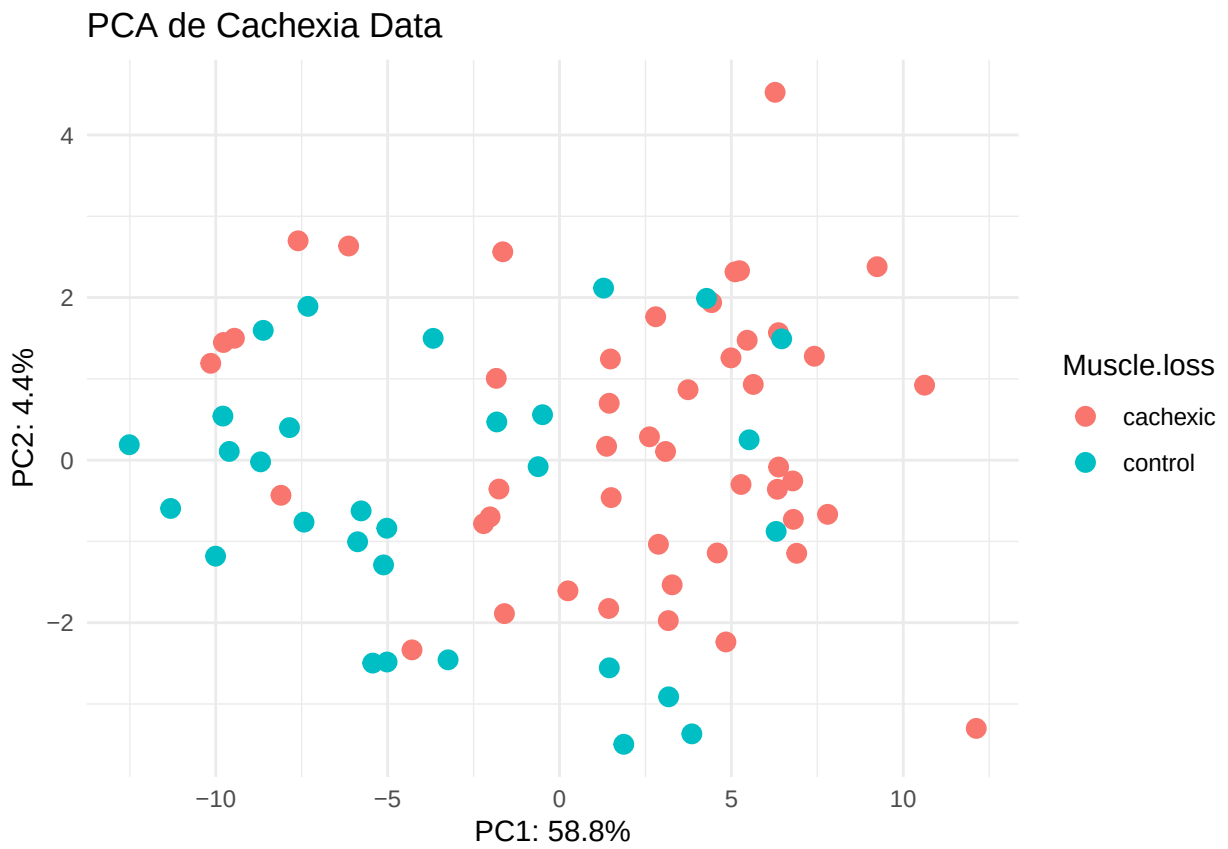


En el recull de boxplot es pot obervar que pels 4 biomarcadors representats, la concentració es més elevada en pacients amb perdua de musculatura. Tot i això, s'haurien d'estudiar tots els biomarcadors per extreure conclusions més esclaridores.

6.PCA: En el nostre cas fem un PCA, posant atenció en el percentatge de variabilitat de les dades explicada per dos components. Ens ajudarà a identificar patrons en les dades i com es relacionen les mostres entre si. En el context de la caquexia, pot ajudar a veure com les concentracions dels biomarcadors varien entre els pacients amb caquexia i els controls, aportant informació útil per a futures investigacions. Per dur a terme la creació del PCA hem instalat el paquet "ggplot2" de Bioconductor.

A continuació podem realitzar l'analisi de PCA sobre la matriu d'expressió transformada. Per fer això utilitzem prcomp sobre la matriu transposada ja que les mostres son les columnes i després extreiem les dades PCA. Per últim agreguem la condició (Muscle.loss) ja que es la nostra variable

A continuació visualitzem les dades per poder entendre com es distribueixen les mostres en l'espai definit per les components principals i quines son les relacions entre els biomarcadors que explica la variabilitat:

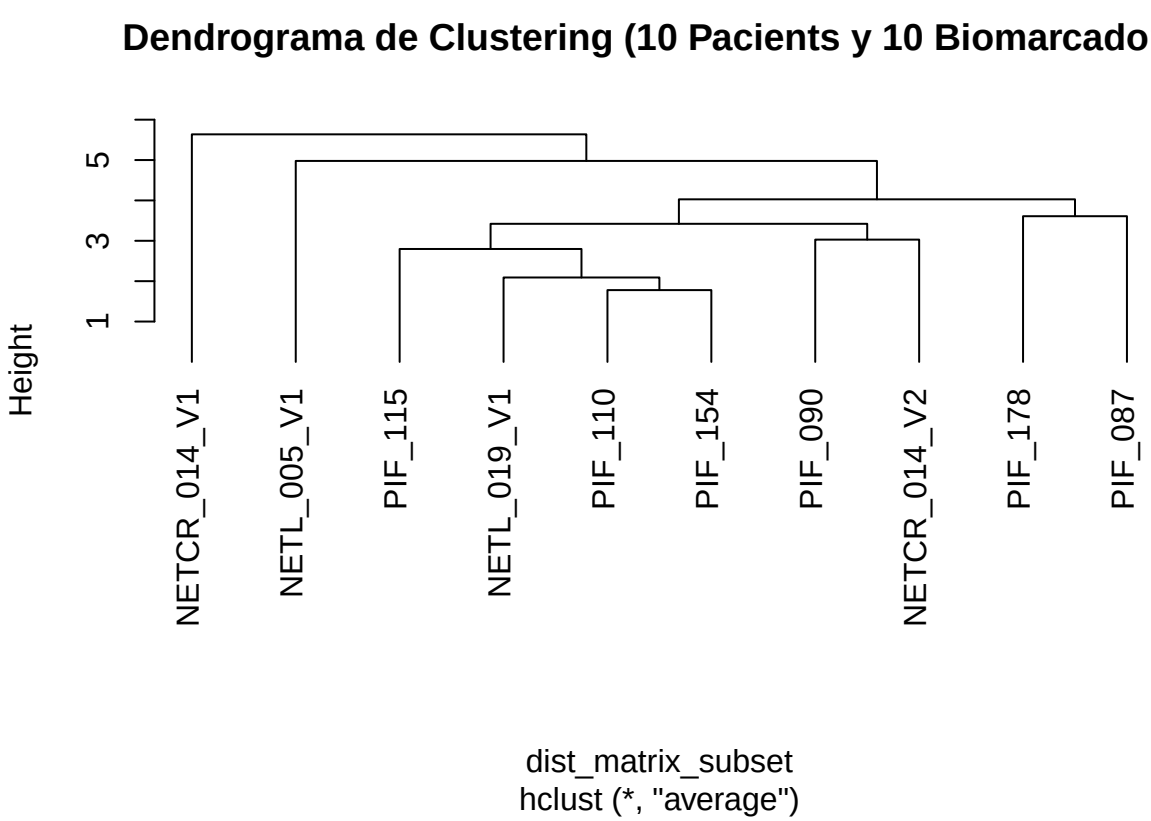


D'aquesta manera podem afirmar que el PC1 (58,5%) es el que millor explica la variabilitat de mostres. En quant als punts, en color vermell, etiquetats com "cachexic", representen mostres que tenen pèrdua muscular (caquèxia) i en color verd blavós, etiquetats com "control", representen mostres de control (sense pèrdua muscular). A simple vista, no sembla haver-hi una separació clara entre les mostres "cachexic" i "control" en els dos primers components principals. Això suggereix que, almenys en termes dels components principals PC1 i PC2, no existeix una diferència òbvia o patró de separació entre aquests dos grups en el conjunt de dades. Per tant, aquesta anàlisi preliminar indica que el PCA en aquests dos components principals no és suficient per diferenciar clarament entre els grups, per tant hauriem, per exemple, d'afegir PC3 i PC4 o altres proves complementaries com per exemple un "Cluster dendogram".

8.Cluster dendogram: Ens podria servir per identificar subgrups de pacients amb patrons metabòlics similars, com diferents graus de la malaltia, cosa que facilita la comprensió de les dades. També a associar biomarcadors específics a certs grups, detectant tant patrons comuns com anomalies, que poden indicar casos especials o errors en les dades. Això proporciona una base més precisa per a anàlisis estadístiques en subgrups homogenis i simplifica la visualització de dades complexes en gràfics més comprensibles.

Per dur a terme l'analisis he seleccionat un subconjunt de dades per fer clustering jeràrquic, en aquest és ideal per simplificar la visualització i facilitar la interpretació dels resultats, ja que un conjunt més petit és més clar i manejable.

```
library(ComplexHeatmap)
#Escalem les dades per igualar les variancies
data_scaled <- scale(t(exp_matrix))
exp_matrix_subset <- data_scaled[1:10, 1:10]
dist_matrix_subset <- dist(exp_matrix_subset)
clust.euclid.average_subset <- hclust(dist_matrix_subset, method = "average")
plot(clust.euclid.average_subset, hang = -1, main = "Dendrograma de Clustering (10 Pacients y 10 Biomar
```



En el dendrograma es mostra com es fusionen les mostres o variables en funció de la seva similitud. Els grups que es fusionen a una altura baixa són més similars (PIF_110,PIF154), mentre que els que es fusionen a una altura alta són més distants (NETCR_014_V1 amb la resta). L'altura de tall permet definir els clústers o grups en què es poden agrupar les mostres. En aquest cas, les mostres o variables agrupades en aquests clústers tenen una gran variabilitat interna i no són gaire similars entre elles. Això vol dir que els grups formats estan relativament dispersos, i no hi ha agrupacions molt fortes basades en la similitud entre les categories definides.

Conclusions i limitacions de l'estudi:

-El conjunt de dades conté un **nombre relativament petit de mostres** (77 mostres), cosa que pot dificultar la generalització dels resultats. A més, la manca d'algunes variables clau podria limitar les conclusions.

-**Falta de separació clara entre grups** (Caquèxia vs Control): Tot i que les anàlisis visuals com el PCA i els boxplots han mostrat algunes diferències, no s'ha observat una separació clara i consistent entre els pacients amb caquèxia i els controls en els dos primers components principals del PCA. Això podria indicar que els biomarcadors utilitzats no són suficients per identificar de manera òbvia les diferències entre els dos grups, o que els biomarcadors no estan sent analitzats en la seva totalitat.

-Alguns **histogrames mostren distribucions amb valors atípics** o una gran dispersió, indicant que les dades no són completament simètriques o segueixen una distribució normal. Això pot dificultar la interpretació de resultats i limitar l'eficàcia d'algunes tècniques estadístiques que assumeixen normalitat.

-**El clustering jeràrquic** no ha revelat clústers molt diferenciats, el que pot suggerir que els biomarcadors no són prou específics per agrupar els pacients en subgrups clars basats en les seves característiques metabòliques.

-**La selecció limitada de biomarcadors en l'anàlisi** (només els primers 9 o 10) pot reduir la capacitat d'identificar patrons metabòlics complets. Una anàlisi més exhaustiva amb un conjunt de biomarcadors més gran podria proporcionar una millor diferenciació entre els grups de pacients.

-**Les metadades disponibles** (únicament la variable "muscle.loss") poden ser insuficients per explicar la variabilitat entre les mostres. Altres factors com l'edat, el gènere, o altres condicions mèdiques dels pacients podrien influir en els resultats i proporcionar una visió més completa.

-**No s'ha realitzat una validació externa de les anàlisi**s, per exemple, mitjançant l'ús de dades de validació independents o tècniques de validació creuada. Això limita la confiança en la robustesa dels resultats obtinguts.

En conclusió, aquestes limitacions suggereixen que, tot i que l'anàlisi exploratòria de les dades pot proporcionar una comprensió inicial de les possibles diferències metabòliques associades a la caquèxia, encara es necessiten estudis més profunds, amb més dades i una selecció més precisa de biomarcadors per obtenir resultats més clarament definitius.

Repositori de Github: https://github.com/azamoras1997/Zamora-Soria-Andrea-PEC1