PEC1_DO

Adrià Hernández Capell

2025-03-30

Contents

Abstract

La caquexia es un síndrome metabólico que se caracteriza por un estado de desnutrición y tiene como consecuencia la pérdida significativa de masa muscular. Este síndrome suele asociarse a diferentes enfermedades crónicas graves como el cáncer entre otras. En el presente estudio se ha realizado un análisis exploratorio de las concentraciones de 63 metabolitos para 77 sujetos agrupados en los grupos de estudio *cachexic* (47 sujetos) y *control* (30 sujetos). Para ello, se ha trabajado con programación en R v. 4.3.3. utilizando las herramientas que ofrece el lenguaje para estudios de bioinformática, como por ejemplo los objetos de tipo *SummarizedExperiment* para estructurar los datos de forma cómoda. Tras un análisis exploratorio de los datos, se realiza un análisis de componentes principales (PCA) y un análisis de expresión diferencial a través de *VarPlot*, identificando ciertos patrones significativos en estado de caquexia para algunos metabolitos como la Glucosa, la Valina, la Leucina, el Quinolinate o el Adipate, posiblemente debido a procesos consecuentes del síndrome como la degradación muscular, inflamación o el estrés metabólico

Objetivos

El presente estudio tiene como objetivo principal la comparativa de concentraciones metabolímicas entre sujetos pertenecientes al grupo *cachexia* (diagnosticados con caquexia) y sujetos pertenecientes al grupo *control* (sin caquexia) para poder identificar patrones metabólicos que puedan resultar diferenciales entre ambos grupos. Para ello, se requiere el tratamiento y el análisis exploratorio (univariable y/o multivariable) de los datos de metabolómica a partir de un objeto de tipo *SummarizedExperiment*.

Métodos

En el presente estudio se ha utilizado los datos almacenados en el documento *human_cachexia.csv* descargado desde el repositorio GIT con URL *https://github.com/nutrimetabolomics/metaboData/tree/main/Datasets/2024-Cachexia* y se ha importado el archivo a un proyecto de RStudio para la elaboración de un informe dinámico RMarkdown que reproduce los resultados. El tratamiento de datos, así como el análisis de los propios se ha llevado a cabo mediante el lenguaje de programación R v. 4.4.3. Se han utilizado diversas librerías de *base* y la librería *SummarizedExperiment* de *BioConductor* de R v. 4.3.3, con la cual se ha generado un objeto de tipo *SummarizedExperiment* para estructurar los datos del archivo *human_cachexia.csv* de la siguiente forma:

Summarized_Experiment:

- `assays`: En este elemento del objeto estructuraremos los datos de expresión que vamos a tratar y analizar de manera que cada fila corresponda a un metabolito y cada columna a una muestra.
- `colData`: En este elemento del objeto estructuraremos los metadatos relacionados con la las muestras que se tendrán en cuenta para el estudio, de manera que cada fila corresponda a una muestra y cada columna a un tipo de información sobre ella.
- `rowData`: En este elemento del objeto estructuraremos los metadatos relacionados con los datos de expresión con los que trabajaremos.

Tras la correcta estructuración del objeto de tipo *SummarizedExperiment* se procede al análisis exploratorio de los datos para identificar patrones metabólicos diferenciales entre los 2 grupos de estudio. Para ello, se realizará primeramente un análisis estadístico multivariable para observar las distribuciones y la variabilidad de cada metabolito para identificar si se requiere una normalización de los datos.

Tras un primer análisis exploratorio de los datos, se procede a un análisis de componentes principales (PCA) para reducir la dimensionalidad de los datos y observar posibles diferencias entre ambos grupos de estudio

basadas en patrones metabólicos, y una representación de `VolacanoPlot` para identificar los metabolitos que se consideran más diferenciales entre ambos grupos de estudio.

Resultados

Generar un objeto SummarizedExperiment

Como primer paso, una vez se ha descargado el archivo *human_cachexia.csv*, se importa el documento con la instrucción *read.csv()* para poder trabajar con los datos almacenados en el archivo desde RStudio:

```
df= read.csv("human_cachexia.csv")
```

Una vez se han cargado los datos, separaremos la matriz que contiene los datos de expresión de los metabolitos de los que metadatos de las muestras. En este caso, los metadatos de las muestras corresponden a la 1a y 2a columna de *df* (Id del paciente y definición del grupo de estudio al que pertenece, respectivamente), por lo que la matriz con datos de expresión corresponderá a la matriz *df* excluyendo las 2 primeras columnas.

```
metabolitos=df[, -c(1,2)]  #Definimos metabolitos como la matriz de datos de
#expresión de df excluyendo las 2 primeras columnas
rownames(metabolitos)=df$Patient.ID  #Asociamos cada fila de metabolitos con
#los respectivos ID de los pacientes para no perder la relación
```

Asimismo, los metadatos de las muestras (pacientes) corresponden a la clasificación de cada una según el grupo de estudio al que pertenecen (cachexicos o no cachexicos), columna *MuscleLoss* de *df*:

```
col_Data= data.frame(MuscleLoss = df$Muscle.loss, row.names = df$Patient.ID)
#Asociamos cada fila de col_Data con los respectivos ID de los pacientes
#para no perder la relación
```

Es necesario asegurarse que la columna *MuscleLoss* de *col_Data* es un factor con dos niveles correctamente definidos, por lo que:

```
col_Data$MuscleLoss= factor(col_Data$MuscleLoss, levels = c("cachexic", "control"))
```

Adicionalmente, se generan los metadatos de las filas de la matriz de datos de expresión, correspondiente a los nombres de los metabolitos:

```
row_Data=data.frame(nom_metabolitos= colnames(metabolitos), row.names = colnames(metabolitos))
```

Una vez estructurados los datos del archivo origen *human_cachexia*, se define el objeto *Summarized_Experiment*. *SummarizedExperiment* se ha instalado previamente desde el paquete *BiocManager* de BioConductor:

```
BiocManager::install("SummarizedExperiment)
```

```
library(SummarizedExperiment) #Cargamos la librería que incorpora la clase
#SummarizedExperiment

se_human_cachexia=SummarizedExperiment(
```

```
  assays= list(counts=t(as.matrix(metabolitos))), #Se guarda la matriz de datos
  #de expresión como matriz y se transpone para que las columnas de assay
  #correspondan con las filas de colData
  rowData = row_Data,
  colData = col_Data
  #Dado que no se nos notifica otro tipo de información relacionada con los
  #metabolitos,no se genera metadata.
  )
```

Se visualiza el objeto *se_human_cachexia*:

```
se_human_cachexia
```

```
## class: SummarizedExperiment
## dim: 63 77
## metadata(0):
## assays(1): counts
## rownames(63): X1.6.Anhydro.beta.D.glucose X1.Methylnicotinamide ...
##   pi.Methylhistidine tau.Methylhistidine
## rowData names(1): nom_metabolitos
## colnames(77): PIF_178 PIF_087 ... NETL_003_V1 NETL_003_V2
## colData names(1): MuscleLoss
```

```
head(colData(se_human_cachexia)) #Visualizamos las primeras filas de colData
```

```
## DataFrame with 6 rows and 1 column
##             MuscleLoss
##               <factor>
## PIF_178       cachexic
## PIF_087       cachexic
## PIF_090       cachexic
## NETL_005_V1   cachexic
## PIF_115       cachexic
## PIF_110       cachexic
```

```
# head(assay(se_human_cachexia)) No se incluye su visualización en el informe
#debido a su extensión
head(rowData(se_human_cachexia))
```

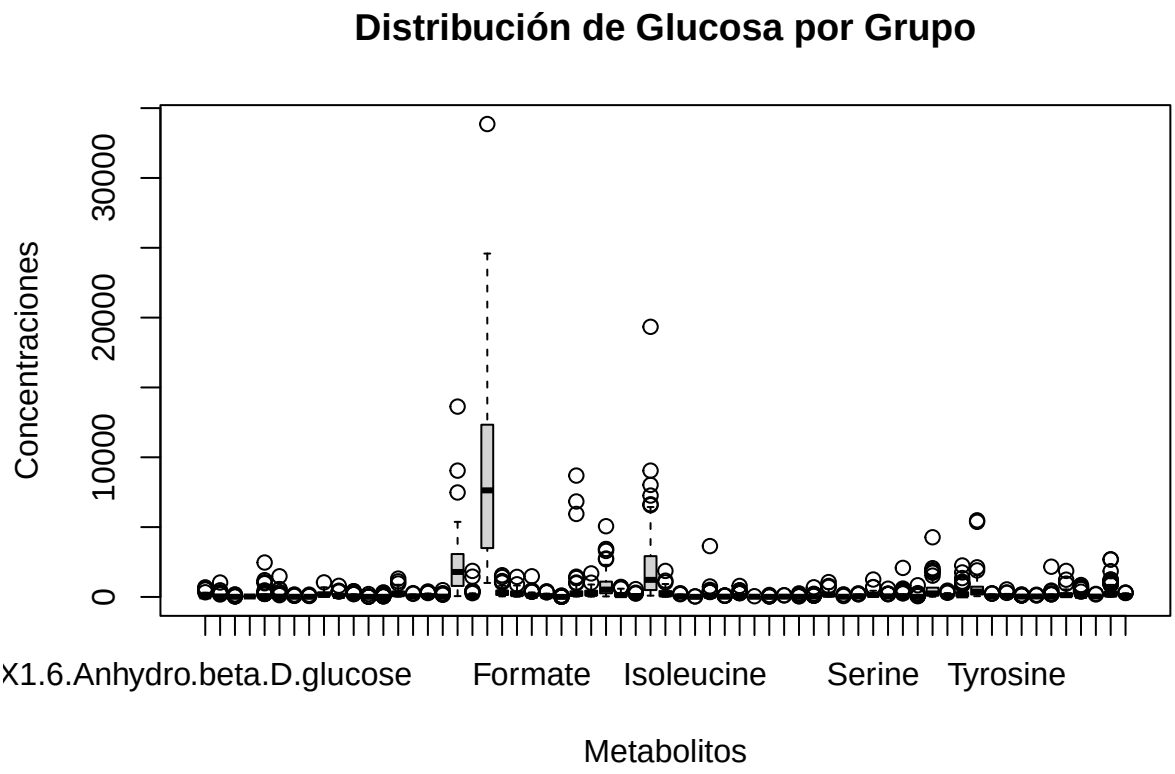
```
## DataFrame with 6 rows and 1 column
##                                    nom_metabolitos
##                                        <character>
## X1.6.Anhydro.beta.D.glucose X1.6.Anhydro.beta.D...
## X1.Methylnicotinamide        X1.Methylnicotinamide
## X2.Aminobutyrate                  X2.Aminobutyrate
## X2.Hydroxyisobutyrate        X2.Hydroxyisobutyrate
## X2.Oxoglutarate                    X2.Oxoglutarate
## X3.Aminoisobutyrate              X3.Aminoisobutyrate
```

Análisis exploratorio

En la siguiente subsección se lleva a cabo un análisis exploratorio de los datos de expresión a partir de la matriz *assay(se_human_cachexia)* y *colData(se_human_cachexia)*.

Primeramente, se realiza un estudio sobre la distribución de los datos de *assay(se_human_cachexia)*, asociados a cada metabolito para observar si existen diferencias de escala y, de ser necesario, realizarse una transformación logarítmica:

```
boxplot(t(assay(se_human_cachexia)),
        main = "Distribución de Glucosa por Grupo",
        xlab = "Metabolitos",
        ylab = "Concentraciones"
        )
```



El boxplot generado representa la distribución y variabilidad de cada metabolito de *assay(se_human_cachexia)*, aportando información sobre valores outliers para cada metabolito y si existe diferencia de escala entre ellos. Visualmente se intuye una diferenciación en la escala entre los diferentes metabolitos evaluados, observando algunos con una mediana (línea central de la caja) próxima a diez mil.

Comprobamos a continuación numéricamente lo que se intuye en el boxplot generado. Para ello, calcularemos el valor de la mediana para todos los metabolitos de *assay(se_human_cachexia)* y visualizaremos los 5 primeros valores y los 5 últimos del cálculo de la mediana para cada metabolito:

```
mediana=apply(assay(se_human_cachexia),1, median) #Calculamos la mediana para
#cada fila de la matriz assay() de nuestro objeto
mediana2=data.frame(rownames(se_human_cachexia), mediana) #Se crea un dataframe
```

```r
#para una mejor visualizzación de los valores máximos y mínimos

head(mediana2[order(-mediana2$mediana),]) #Visualización de los 5 primeros
```

```
##                         rownames.se_human_cachexia. mediana
## Creatinine                               Creatinine 7631.20
## Citrate                                     Citrate 1790.05
## Hippurate                                 Hippurate 1224.15
## Glycine                                     Glycine  528.48
## Trimethylamine.N.oxide      Trimethylamine.N.oxide  383.75
## Dimethylamine                         Dimethylamine  304.90
```

```r
#metabolitos con mediana máxima

tail(mediana2[order(-mediana2$mediana),]) #Visualización de los 5 últimos
```

```
##                   rownames.se_human_cachexia. mediana
## X2.Aminobutyrate            X2.Aminobutyrate   10.49
## Adipate                              Adipate   10.18
## Methylguanidine               Methylguanidine    7.85
## Isoleucine                         Isoleucine    7.17
## Acetone                               Acetone    7.10
## Fumarate                             Fumarate    4.10
```

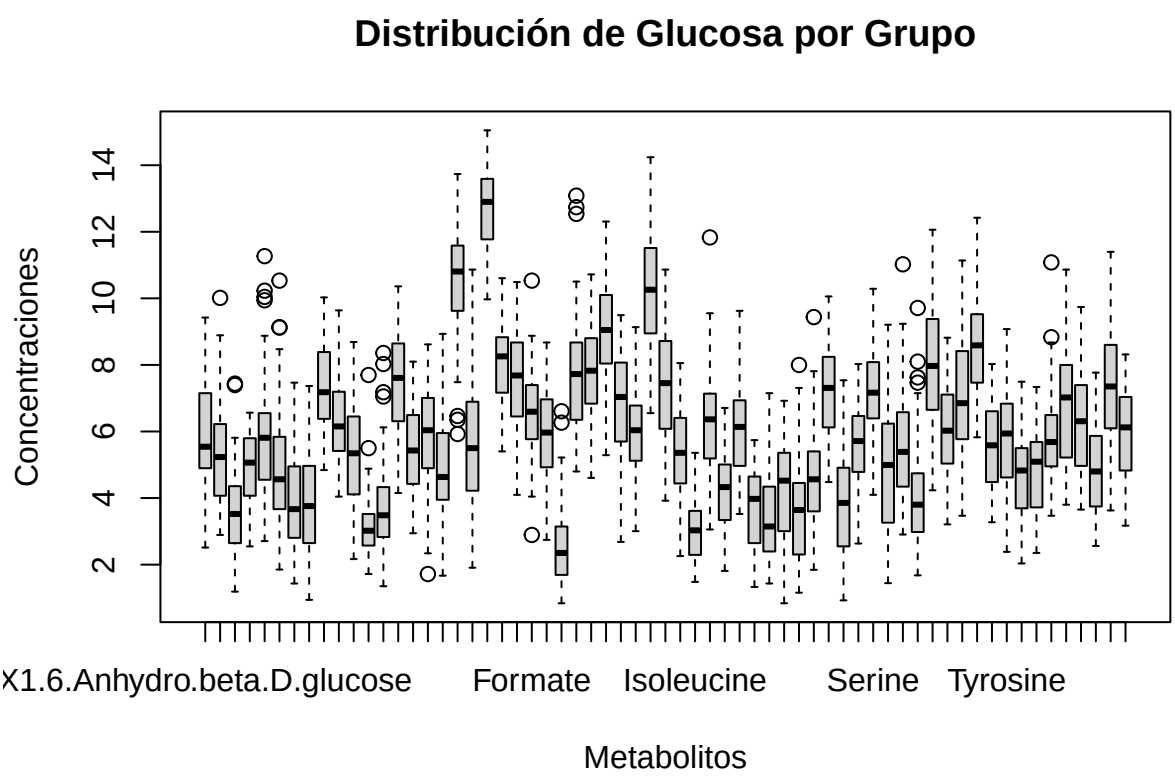
```r
#metabolitos con mediana máxima
```

En consecuencia, con el objetivo de poder continuar un análisis exploratorio y comparativo de los datos, se realiza una normalización logarítmica de los datos de *assay(se_human_cachexia)*:

```r
log_assay=log2(assay(se_human_cachexia)+1) #Transformación logarítmica de los
#datos
```

Se representa a continuación un boxplot de los datos de matabolómica tras realizar la transformación:

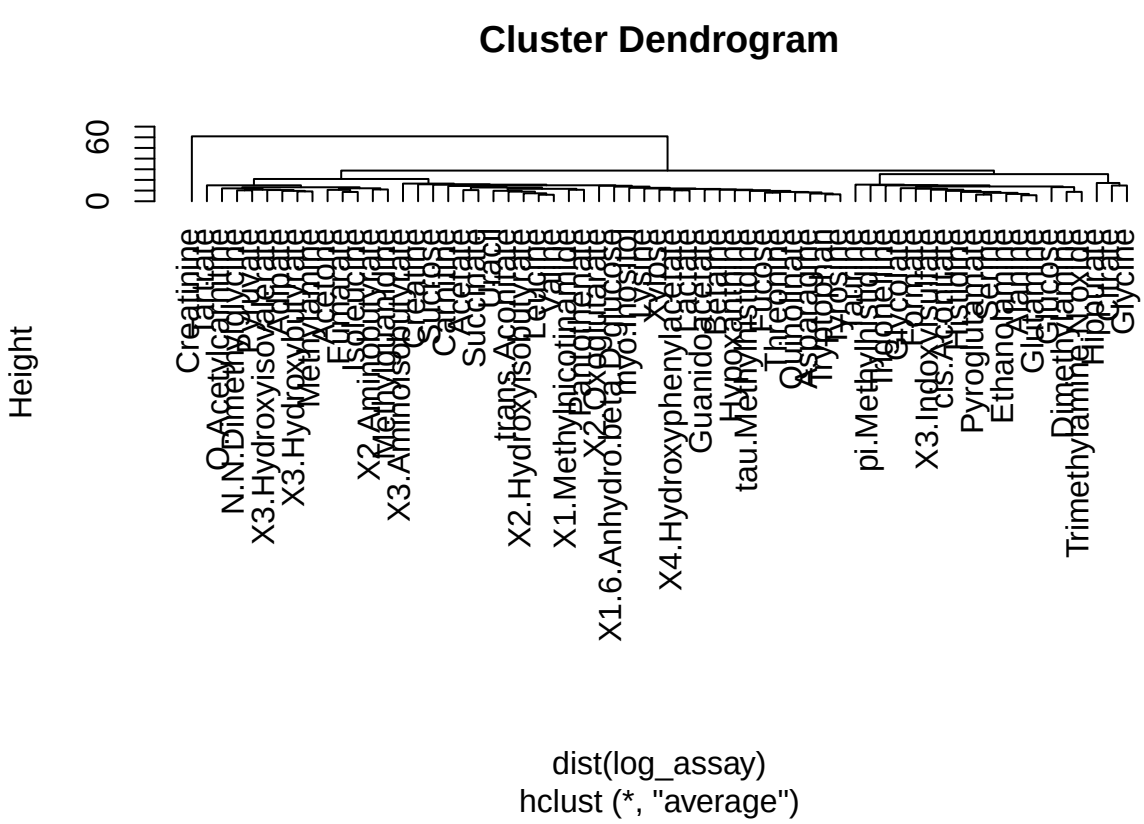
```r
boxplot(t(log_assay),
        main = "Distribución de Glucosa por Grupo",
        xlab = "Metabolitos",
        ylab = "Concentraciones"
        )
```



Se ha conseguido de esta forma que la diferencia de escalas entre metabolitos se haya reducido y la distibución sea homogénea.

Se realiza a continuación un análisis jerárquico de los metabolitos para agrupar jerárquicamente los metabolitos en los que se identifican patrones de similitud. Cabe destacar que las concentraciones de los metabolitos pueden variar en función de factores externos como la dieta de los pacientes.

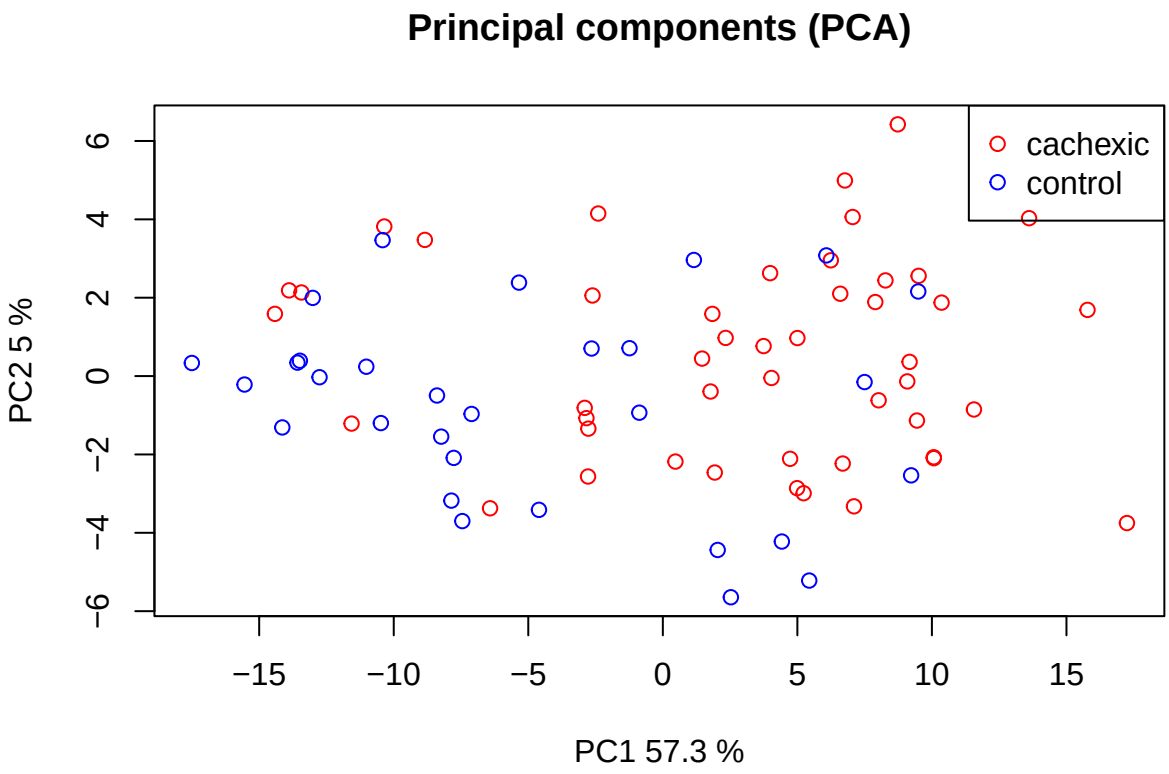
```
clust.euclid.average=hclust(dist(log_assay),method="average") #Se utiliza
#la distancia euclideana por defecto en dist() entre columnas para realizar
#el clustering entre los mentabolitos.

plot(clust.euclid.average, hang=-1) #Visualizamos el dendograma que genera el
```




```
#clustering anterior
```

De la interpretación del dendograma se puede observar un algunos patrones de similitud entre algunos metabolitos. Esta agrupación puede deberse a que existen diversos procesos metabólicos que afectan a la concentración de diversos metabolitos. En estado de cachexia, se producen procesos de degradación muscular o inflamación, los cuales pueden provocar que diferentes metabolitos observen patrones metabólicos similares. Asimismo, estos mismos metabolitos también pueden presentar patrones similares en estado de control.

En este momento, será relevante analizar cuáles son los metabolitos que presentan un mayor factor diferencial durante el estado de cachexia respecto el estado de control. Para ello, se lleva a cabo un análisis de componentes principales (PCA). Este análisis se utiliza como técnica para reducir la dimensionalidad de los datos, es decir, transforma un conjunto de datos posiblemente correlacionados, en un conjunto de menor dimensión con componentes independientes entre sí. De esta forma, se mantiene la información de los datos disminuyendo el número de variables a estudio.

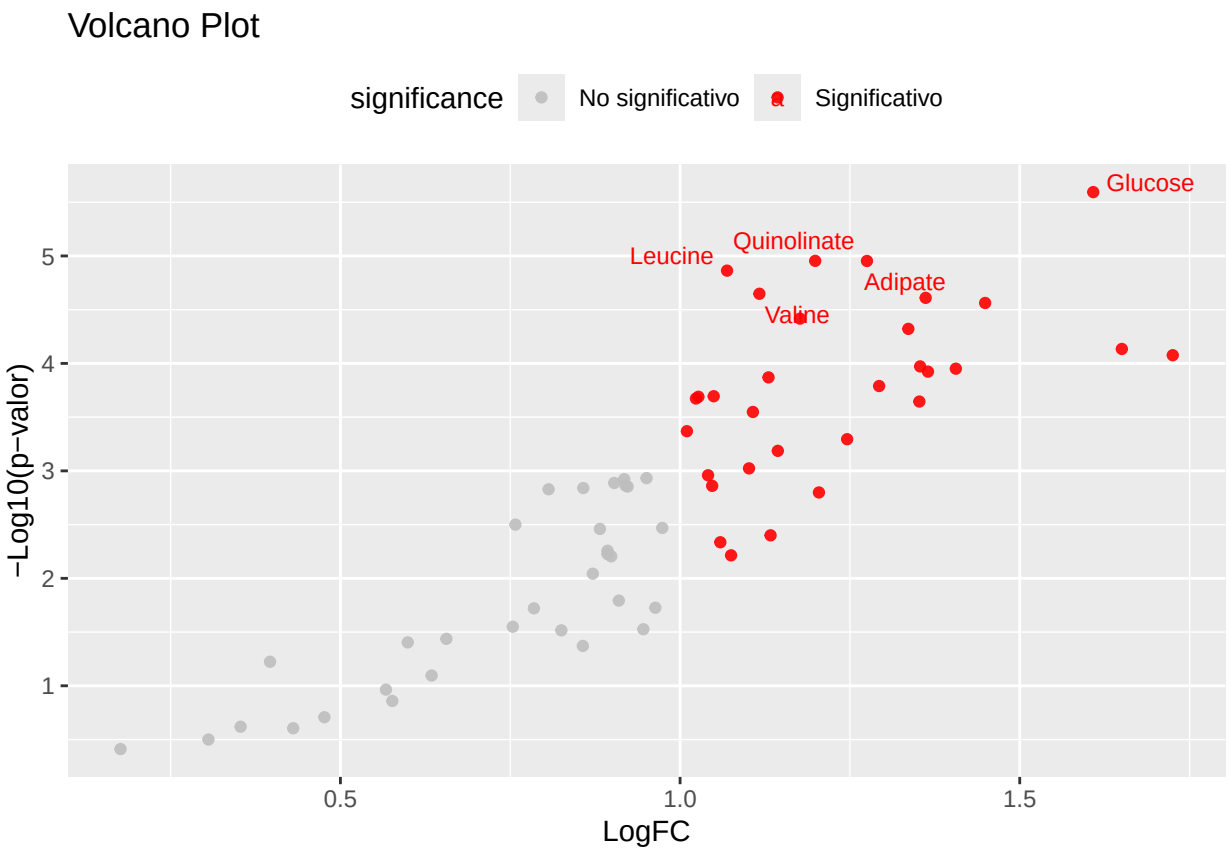


Tras realizar el análisis de componentes principales y plotear las dos principales, se observa que las 2 primeras componentes explican más del 60% de la variabilidad de los datos. Asimismo, se intuye una posible diferenciación de los pacientes cachexicos y los de control en el eje X, es decir, la primera componente principal indica una leve diferenciación entre los dos grupos de estudio. Esto indica que el estado de cachexia puede provocar patrones metabólicos.

Finalmente, se realiza un análisis de expresión diferencial para identificar cuáles son los metabolitos que sufren mayor variabilidad en estado de cachexia respecto el estado de control (sano). Para ello, se genera un `VolcanoPlot`:

En este caso, hemos indicado que se considere como significativos, aquellos metabolitos que tienen un p.valor inferior a 0.05 y un logFC superior en valor absoluto a 1. Los valores de logFC nos proporcionan la información de cuánto varía ese metabolito en los sujetos de un grupo de estudio respecto al otro. Cuanto mayor es el valor, indica una mayor diferenciación entre grupos para ese metabolito. Cabe destacar que un valor negativo de logFC indica que ese metabolito disminuye en el grupo de referencia respecto al otro grupo (en este caso, el grupo de referencia es *cachexic*). Un valor positivo de logFC indica que ese metabolito aumenta en el grupo de referencia respecto al otro grupo.

Visualizamos el gráfico con *ggplot2*:



Discusión

Tras el análisis exploratorio de los datos se puede observar que existe una gran diferencia de escala en los metabolitos que se han estudiado, provocando que sea necesario una normalización para poder realizar análisis diferenciales. El dendograma generado para realizar un análisis de clustering jerárquico tras la normalización de los datos parece indicar que existen patrones de similitud entre varios metabolitos posiblemente debido a una implicación común en diferentes procesos metabólicos.

El análisis de las componentes principales indica que existen metabolitos que se encuentran correlacionados, dado que las 2 primeras componentes principales (reducción de dimensión) pueden explicar más de un 60% de la variabilidad de los datos. Esencialmente, la 1a componente principal parece expresar una leve diferenciación entre los dos grupos de estudio.

Finalmente, tras la representación del *VolcanoPlot* se observan únicamente valores positivos de logFC. En este contexto, esto indicaría que en estado de cachexia, todos los metabolitos evaluados parecen aumentar en concentración respecto el grupo control. **De la siguiente forma, los metabolitos que muestran una expresión diferencial mayor entre los dos grupos de estudio son la Glucosa, la Valina, la Leucina, el Quinolinate y el Adipate.**

Conclusiones

Tras la discusión de los resultados, se puede concluir que existen patrones metabólicos diferenciales que se expresan como una variación de las concentraciones de algunos metabolitos en pacientes en estado de cachexia. En estado de cachexia se producen procesos como degradación muscular, lo que libera aminoácidos en sangre

para obtener energía. Esto provoca un aumento de los metabolitos relacionados con este proceso metabólico, como la Leucina, Valina, Glutamina, Creatina o Creatinina entre otros. Adicionalmente, en estado de cachexia también se observa un estado de estrés metabólico en el que una alteración del metabolismo tiene un impacto como procesos de inflamación (lo cual puede provocar el aumento de metabolitos relacionados con este proceso como el Quinolinate).

Es decir, en estado de cachexia, existen diferentes procesos consecuentes como degradación muscular, inflamación o estrés metabólico que pueden provocar un aumento diferencial en algunos de los metabolitos evaluados para el presente estudio, principalmente en la Glucosa, la Valina, la Leucina, el Quinolinate y el Adipate. No obstante, dado que se realiza el análisis con un tamaño muestral de 77 sujetos, se considera adecuado la realización de estudios adicionales con un tamaño muestral mayor para, de esta forma, considerar los resultados significativos.

Referencias

- Repositorio GitHub del proyecto: https://github.com/adriahc22/Hernandez-Capell-Adria-PEC1
- Fuente origen de los datos *human_cachexia.csv*: https://github.com/nutrimetabolomics/metaboData/tree/main/Datasets/2024-Cachexia
- Análisis exploratorio multivariable: https://aspteaching.github.io/Analisis_de_datos_omicos-Ejemplo_0-Microarrays/ExploreArrays.html#3_Exploratory_Data_Analysis