# PEC 1. Análisis de datos ómicos (M0-157)

Alejandro Gombau García

2025-03-31

## Contents



**Repositorio de GitHub**

https://github.com/alexgomb/Gombau-Garcia-Alejandro-PEC1

## Abstract

En esta PEC se explora de manera integral un conjunto de datos ómicos procedente de muestras de orina de pacientes con cáncer, diferenciando entre individuos caquécticos y controles. Se construyó un objeto *SummarizedExperiment* a partir de datos obtenidos de GitHub, y se aplicaron diversas técnicas de análisis exploratorio—incluyendo resúmenes estadísticos, boxplots, heatmaps, análisis de componentes principales, clustering jerárquico y un *volcano plot*. Este último permitió combinar el log2 fold change y la significación estadística para identificar metabolitos con diferencias notables entre ambos grupos, evidenciando que muchos de estos compuestos se encuentran elevados de forma significativa en los pacientes caquécticos. Aunque los perfiles metabólicos en conjunto no discriminan de forma clara entre los grupos, los hallazgos obtenidos sugieren la necesidad de aplicar enfoques adicionales y análisis más específicos para desentrañar los mecanismos metabólicos subyacentes a la caquexia.

# Objetivos

- Analizar un conjunto de datos ómicos de muestras de orina de pacientes con cáncer, diferenciando entre caquécticos y controles.
- Identificar diferencias en los perfiles metabólicos que puedan servir como biomarcadores o tener relevancia biológica en la caquexia.
- Aplicar técnicas exploratorias (estadísticos, gráficos, PCA, clustering y volcano plot) para detectar patrones significativos en los datos.

# Metodos y flujo de trabajo

Para esta PEC utilizamos el dataset disponible en GitHub: Datasets/2024-Cachexia/human_cachexia.csv

Como deciamos, se trata de un dataset de un conjunto de datos compuesto por concentraciones de 77 muestras de orina de pacientes con cáncer (caquécticos frente a control). La caquexia es un síndrome multifactorial caracterizado por una pérdida significativa de peso y masa muscular, que suele ir acompañada de fatiga, debilidad y alteraciones metabólicas.

Para entender que tipo de pacientes tenemos en el dataset utilizaremos `unique`

```
## [1] "cachexic" "control"
```

## Transformación de los datos

Dado que el dataset tiene 77 filas (pacientes) y 65 columnas de mediciones de metabolitos, para construir un objeto de SummarizedExperiments es necesario organizar la información de forma que las filas representen los features (los metabolitos) y las columnas las muestras (los pacientes). Basicamente, transponer la parte numérica del dataset.

```
## 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): Metabolito
## colnames(77): PIF_178 PIF_087 ... NETL_003_V1 NETL_003_V2
## colData names(2): Patient.ID Muscle.loss
```

SummarizedExperiment se diferencia de ExpressionSet en que permite almacenar múltiples matrices de datos (assays) en un solo objeto. Esto facilita trabajar con versiones diversas de los datos, lo cual facilita la expresión. Los datos en SummarizedExperiments estan organizados en rowData y colData utilizando una clase especial de dataframe que en contraste con el tipo de dataframe de ExpressionSets (`AnnotatedDataFrame`) es bastante mas flexible y adaptable.

## Análisis Exploratorio

En primer lugar exploramos la estructura del objeto SummarizedExperiment

```
## 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): Metabolito
## colnames(77): PIF_178 PIF_087 ... NETL_003_V1 NETL_003_V2
## colData names(2): Patient.ID Muscle.loss
```

Este output nos permite de forma compacta ver caracteristicas del objeto. Vemos que tiene 63 features (los metabolitos) y 77 muestras (los pacientes). La matriz de datos se almacena en el assay "counts", los nombres de las filas corresponden a cada metabolito (por ejemplo, X1.6.Anhydro.beta.D.glucose, X1.Methylnicotinamide, etc.) y en el rowData se guarda información adicional, en este caso solo el nombre del metabolito bajo la columna "Metabolito". Por otro lado, colData contiene dos variables: "Patient.ID" y "Muscle.loss", que representan los metadatos de cada paciente.

Ahora visualizamos los metadatod de las filas (features) y columnas (muestras). Empezando por las primeras filas de RowData (features).

```
## DataFrame with 6 rows and 1 column
##                                        Metabolito
##                                       <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
```

Visualizamos las primeras columnas de colData (muestras).

```
## DataFrame with 6 rows and 2 columns
##              Patient.ID 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
```

A continuación creamos un resumen estadístico de la matriz de datos. Vamos a extraerlo usando la función `assay` de la librería `SummarizedExperiments`.

```
##                              Min. 1st Qu. Median      Mean 3rd Qu.    Max.
## X1.6.Anhydro.beta.D.glucose 4.71   28.79  45.60 105.63039  141.17  685.40
## X1.Methylnicotinamide       6.42   15.80  36.60  71.57364   73.70 1032.77
## X2.Aminobutyrate            1.28    5.26  10.49  18.15974   19.49  172.43
## X2.Hydroxyisobutyrate       4.85   15.80  32.46  37.25065   54.60   93.69
## X2.Oxoglutarate             5.53   22.42  55.15 145.08714   92.76 2465.13
## X3.Aminoisobutyrate         2.61   11.70  22.65  76.75636   56.26 1480.30
```

```
## Output acortado para favorecer la visualización.
```

A continuación graficamos la distribución de los valores para cada metabolito en un boxplot.

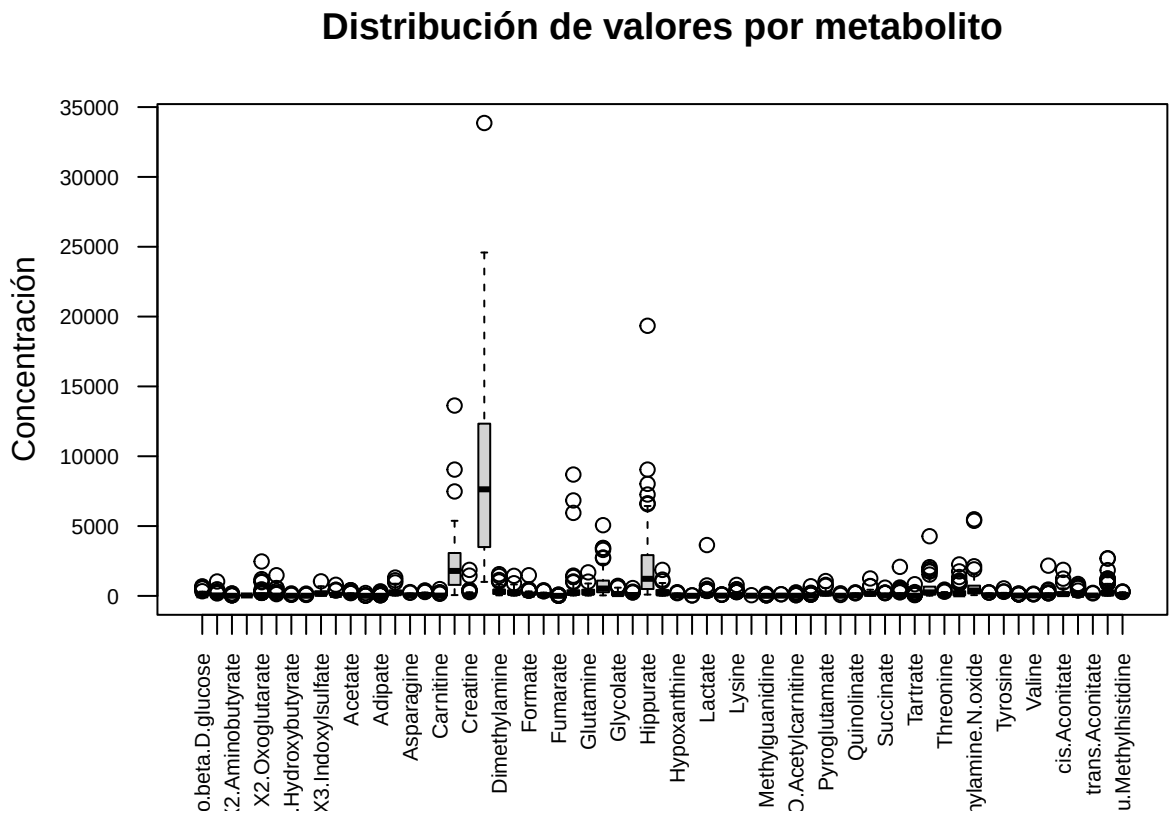


Figure 1: Distribución de los valores de concentración por metabolito

Vemos que hay algunos metabolitos que se comporten de manera similar. A continuación realizaremos un heatmap. Esto nos sirve sirve para visualizar la distribución de los valores de cada metabolito (o grupo de datos) y detectar rápidamente si existen valores atípicos, asimetrías o diferencias importantes entre distintos grupos.

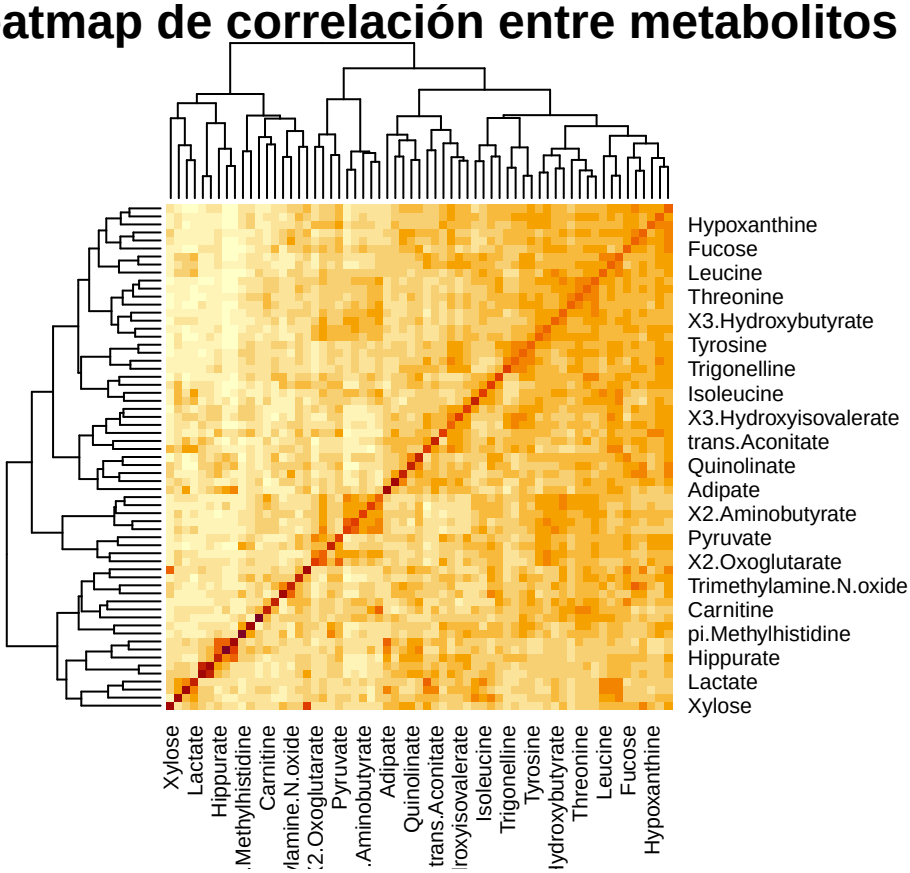


Figure 2: Heatmap de correlación entre metabolitos. Muestra la similitud entre compuestos.

Este heatmap lo que indica que comparten patrones de concentración similares a lo largo de las muestras. Se distinguen regiones con alta correlación (áreas de colores intensos) que sugieren relaciones funcionales o rutas metabólicas comunes, mientras que otros metabolitos muestran correlaciones más débiles, lo que podría reflejar diferencias en su regulación o función.

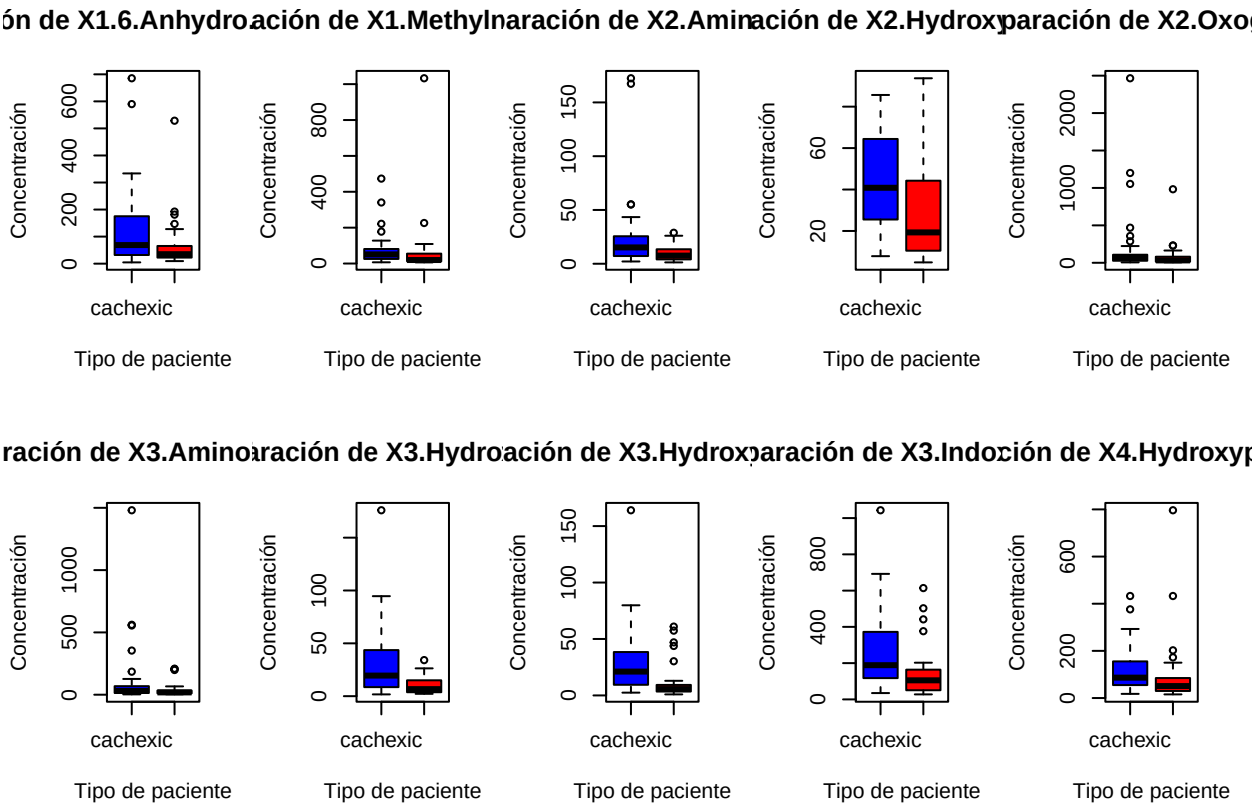


Figure 3: Boxplots comparativos de 10 metabolitos entre pacientes caquécticos y controles.

En la comparación entre grupo `control` y `cachexic` que hemos realizado para comparar algunos metabolitos, vemos que en los 10 seleccionados tenemos mayores concentraciones del metabolito en cuestión para el grupo de caquexia. Sin embargo, este procedimiento no es optimo debido al gran numero de columnas que tenemos. Por lo que una estrategia de reducción de la dimensionalidad podría ser interesante.

Para tratar de captar esa diferencia, reduciremos componentes principales aplicandolo a los metabolitos. De ellos esperamos que expliquen la mayor parte de la variabilidad del conjunto de datos.

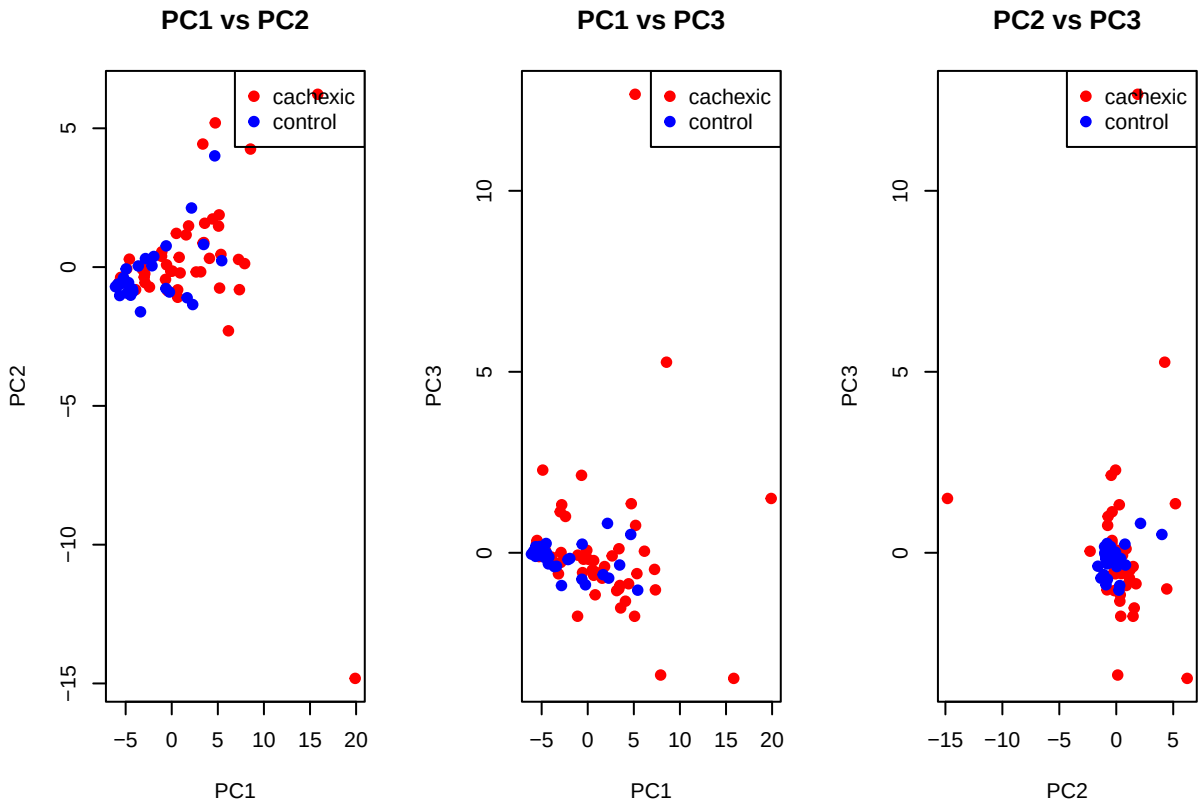


Figure 4: PCA de perfiles metabólicos: proyecciones en PC1 vs PC2, PC1 vs PC3 y PC2 vs PC3, con pacientes coloreados según su condición (cachexic en rojo, control en azul)

Esta grafíca permite observar si los perfiles metabólicos de los pacientes se agrupan de forma distinta según su condición, lo cual puede sugerir diferencias globales en el patrón de metabolitos entre los dos grupos y ayudar a identificar posibles biomarcadores o rutas metabólicas diferenciadas.

A primera vista, se observa que en ninguna de estas combinaciones de componentes principales hay una separación clara entre los grupos, lo que sugiere que, al menos con estas proyecciones lineales, los perfiles metabólicos de cachexic y control no difieren de forma marcada.

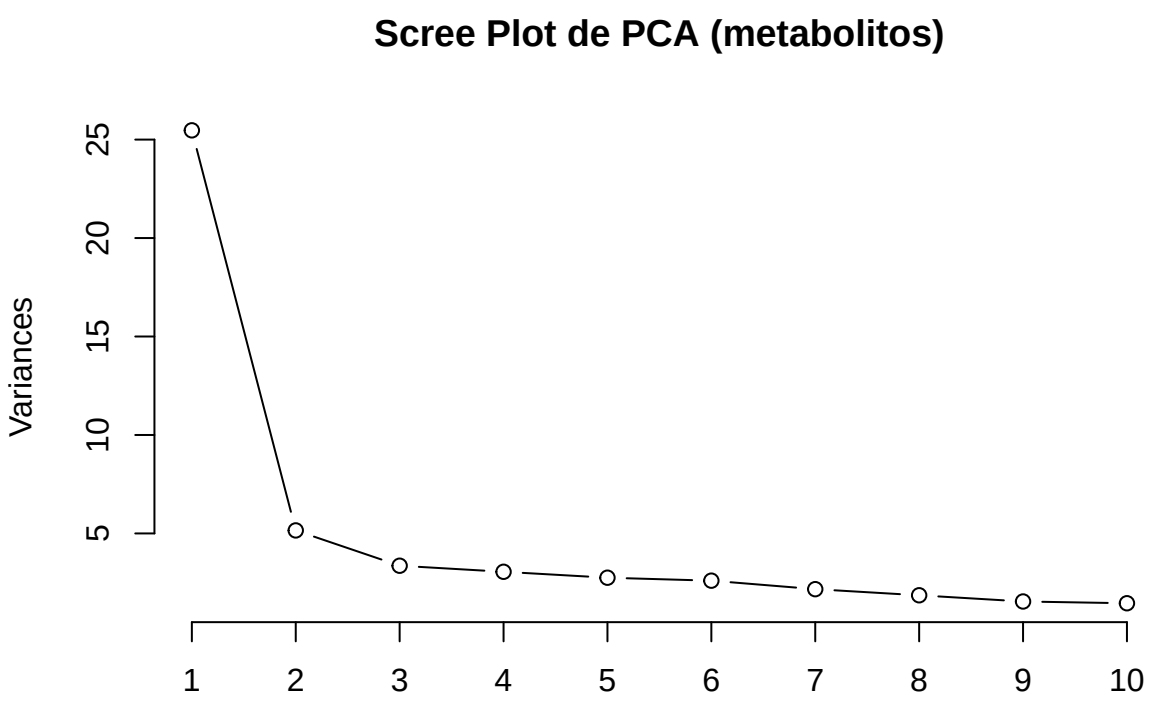


Figure 5: Scree Plot de PCA que muestra la varianza explicada por cada componente principal.

En el Scree Plot se muestra la varianza explicada por los primeros diez componentes principales. Se observa un marcado descenso desde la primera componente, seguida de un aplanamiento a partir de la tercera o cuarta componente. Esto indica que las primeras dos o tres componentes recogen la mayor parte de la variabilidad en los datos, mientras que las componentes posteriores añaden contribuciones cada vez más pequeñas.

Otra opción a explorar puede ser el uso de dendrogramas, en ellos tratamos de ver posibles relaciones no lineales. En este caso tratamos de ver en los pacientes, teniendo en cuenta si son caquexicos o controles, y su perfil metabólico difiere.

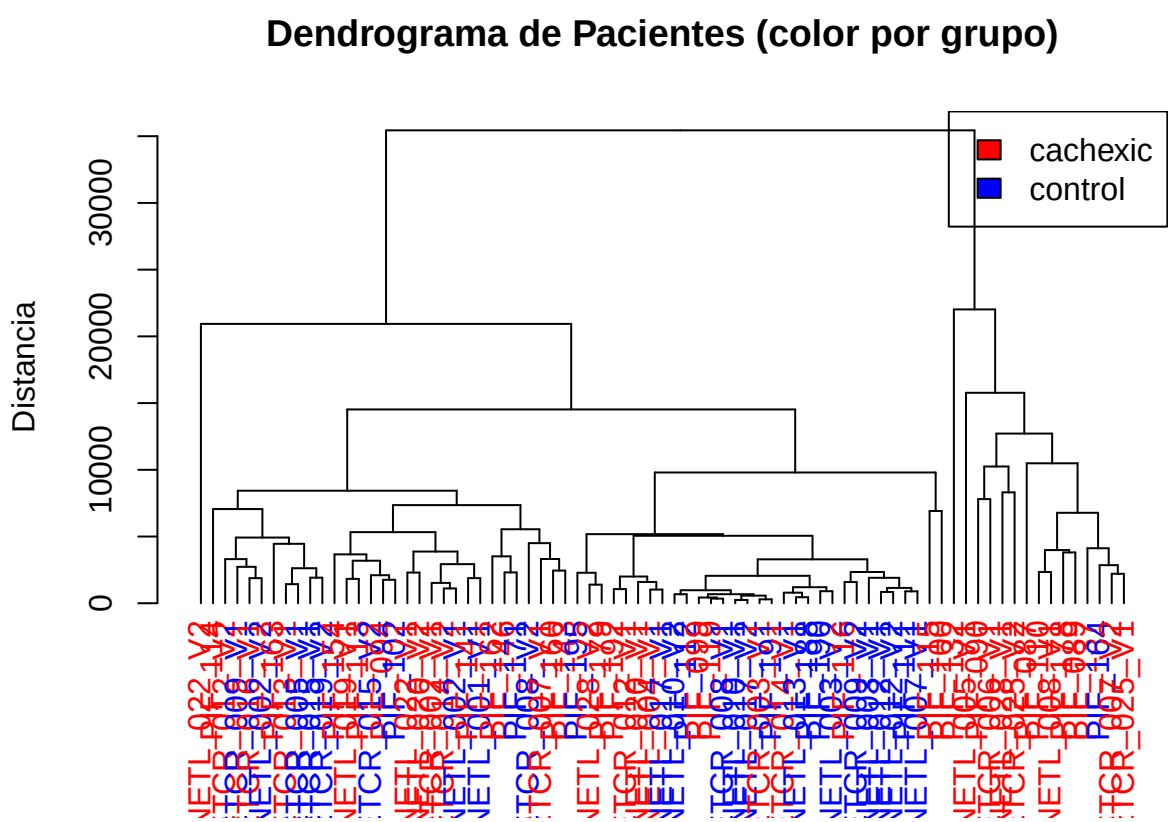


Figure 6: Dendrograma de pacientes: etiquetas coloreadas (rojo: cachexic, azul: control) que reflejan la similitud en los perfiles metabólicos.

A simple vista se observa que los pacientes se agrupan de forma mixta, lo que sugiere que, con la configuración actual de análisis, los perfiles metabólicos no discriminan claramente entre caquexia y control.

El *volcano plot* se utiliza para visualizar de manera simultánea la magnitud del cambio en la concentración de cada metabolito y la significación estadística de ese cambio entre los grupos de pacientes caquécticos y controles.

En este gráfico, el eje x representa el log2 fold change, lo que permite identificar qué tan aumentados o disminuidos están los metabolitos en el grupo caquéctico en comparación con el grupo control, mientras que el eje y muestra el -log10(p-value), indicando la robustez estadística de estas diferencias. Esta representación es particularmente útil en estudios ómicos porque permite resaltar de forma visual aquellos metabolitos que no solo presentan diferencias cuantitativas grandes, sino que también son estadísticamente significativos, sugiriendo que podrían tener un papel biológico relevante en la patología de la caquexia.

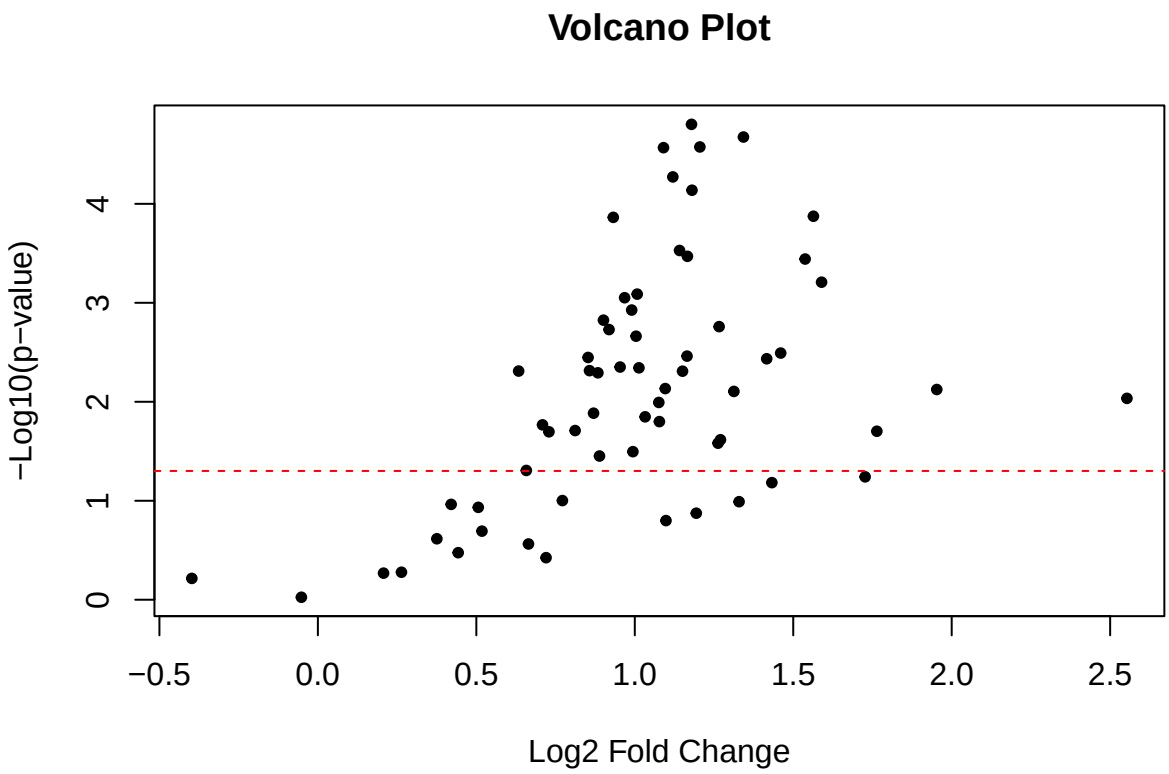


Figure 7: Volcano plot que integra el log2 fold change y la significación (-log10 p-value) para identificar metabolitos diferenciados entre caquécticos y controles. La línea roja indica el umbral de significación (p=0.05).

El eje horizontal (log2 fold change) indica cuán aumentados o disminuidos están los metabolitos en el grupo caquéctico frente al control, mientras que el eje vertical (-log10(p-value)) refleja la robustez estadística de esas diferencias. La mayoría de los puntos se sitúa en la zona de fold change positivo, lo que sugiere que, en promedio, muchos metabolitos muestran concentraciones más elevadas en los pacientes caquécticos. Esto sugiere que algunos metabolitos podrian tener cierta relevancia en el estudio de la caquexia.

## Resultados y discusión

En el análisis exploratorio se observó que el objeto *SummarizedExperiment* contiene 63 metabolitos y 77 muestras, lo que evidencia la heterogeneidad en los niveles de concentración de los metabolitos. Los resúmenes estadísticos y los gráficos de distribución, como los boxplots y heatmaps, mostraron rangos amplios y la presencia de valores atípicos en varios metabolitos. En la comparación univariada de 10 metabolitos se apreció que, en general, los pacientes caquécticos presentan concentraciones mayores que los controles. Además, se realizó un análisis univariado mediante un *volcano plot*, que integró el log2 fold change y la significación estadística, permitiendo identificar aquellos metabolitos con diferencias relevantes entre ambos grupos; este gráfico reveló que la mayoría de los metabolitos tienen fold changes positivos y varios alcanzan significación estadística, indicando que se encuentran elevados en el grupo caquéctico. Sin embargo, los análisis multivariados, como el PCA y el clustering jerárquico, no lograron separar de forma clara a los grupos, sugiriendo que, a pesar de las diferencias puntuales en ciertos metabolitos, es necesario aplicar enfoques adicionales y

técnicas más avanzadas de reducción de dimensionalidad para identificar patrones robustos que distingan entre caquexia y control.

## Conclusiones

- Se detectan diferencias puntuales en la concentración de ciertos metabolitos, con niveles más elevados en caquécticos, según el volcano plot.
- Los análisis multivariados (PCA y clustering) no diferenciaron claramente ambos grupos.
- La alta variabilidad y la presencia de outliers indican la complejidad de los perfiles metabólicos y la necesidad de enfoques adicionales.

## Referencias

*Eisner, R., Stretch, C., Eastman, T. et al.* Learning to predict cancer-associated skeletal muscle wasting from 1H-NMR profiles of urinary metabolites. *Metabolomics* 7, 25–34 (2011). https://doi.org/10.1007/s11306-010-0232-9