# Informe de la Prueba de Evaluación Continua 1

Análisis de Datos Ómicos

Samuel Pis Vigil

28/03/2025

## Contents



## Introducción

En enlace a la página de GitHub con toda la información necesaria para la realización de esta PEC es este : https://github.com/spisv/Pis_Vigil_Samuel_PEC1

Para realizar esta PEC he escogido el archivo de datos *Human cachexia* del repositorio de GitHub que se nos ha proporcionado en este ejercicio. He escogido este set porque me pareció un ejemplo perfecto de caso en el que se puede realizar un análisis diferencial entre muestras pertenecientes a pacientes de una enfermedad (la caquexia) y muestras pertenecientes a controles sanos. La caquexia es una enfermedad asociada a una severa desnutrición, lo cual provoca una serie de síntomas como la pérdida de músculo. Este hecho es fundamental para que podamos interpretar apropiadamente los datos en el contexto biomédico en el cual han sido recogidos.

## Descarga del archivo

```
# Descargamos los datos desde usando el link al repositorio de GitHub
Cachexia <- read.csv("https://raw.githubusercontent.com/nutrimetabolomics/metaboData/main/Datasets/2024-
```

## Generación del objeto *Summarized Experiment*

Vamos a generar un objeto *SummarizedExperiment* con los datos del set:

```
library(SummarizedExperiment)

# Extraemos los datos como una matriz, descartando las dos primeras columnas
# (que se corresponden a los metadatos)
matriz <- as.matrix(Cachexia[, -c(1,2)])

# Asignamos el ID del paciente o control como nombre de fila
rownames(matriz) <- Cachexia[,1]

# Extraemos los metadatos y les asignamos el nombre de lo que son
# (IDs de pacientes y grupo [paciente o control])
columnas <- data.frame(
  PatientID = Cachexia[,1],
  Group = Cachexia[,2],
  row.names = Cachexia[,1]
)
filas <- data.frame(
  Metabolite = colnames(matriz),
  row.names = colnames(matriz)
)

# Creamos el objeto SummarizedExperiment, con una trasposición de la matriz
# para que las filas sean las variables y las columnas los pacientes o controles
sumario <- SummarizedExperiment(
  assays = list(counts = t(matriz)),
  colData = columnas,
  rowData = filas
)

sumario
```

```
## 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): Metabolite
## colnames(77): PIF_178 PIF_087 ... NETL_003_V1 NETL_003_V2
## colData names(2): PatientID Group
```

```
save(sumario, file = "sumario.rda")
```

La principal diferencia de un objeto SummarizedExperiment es que recopila tanto la matriz con los datos del experimento como los metadatos asociados al mismo, lo que nos permite tener en un mismo objeto de R toda la información del mismo experimento. Por contra,un expression set recopila solamente los datos en sí mismos, sin incorporar los metadatos, que en este caso son los IDs de los pacientes y las diferentes variables metabólicas analizadas

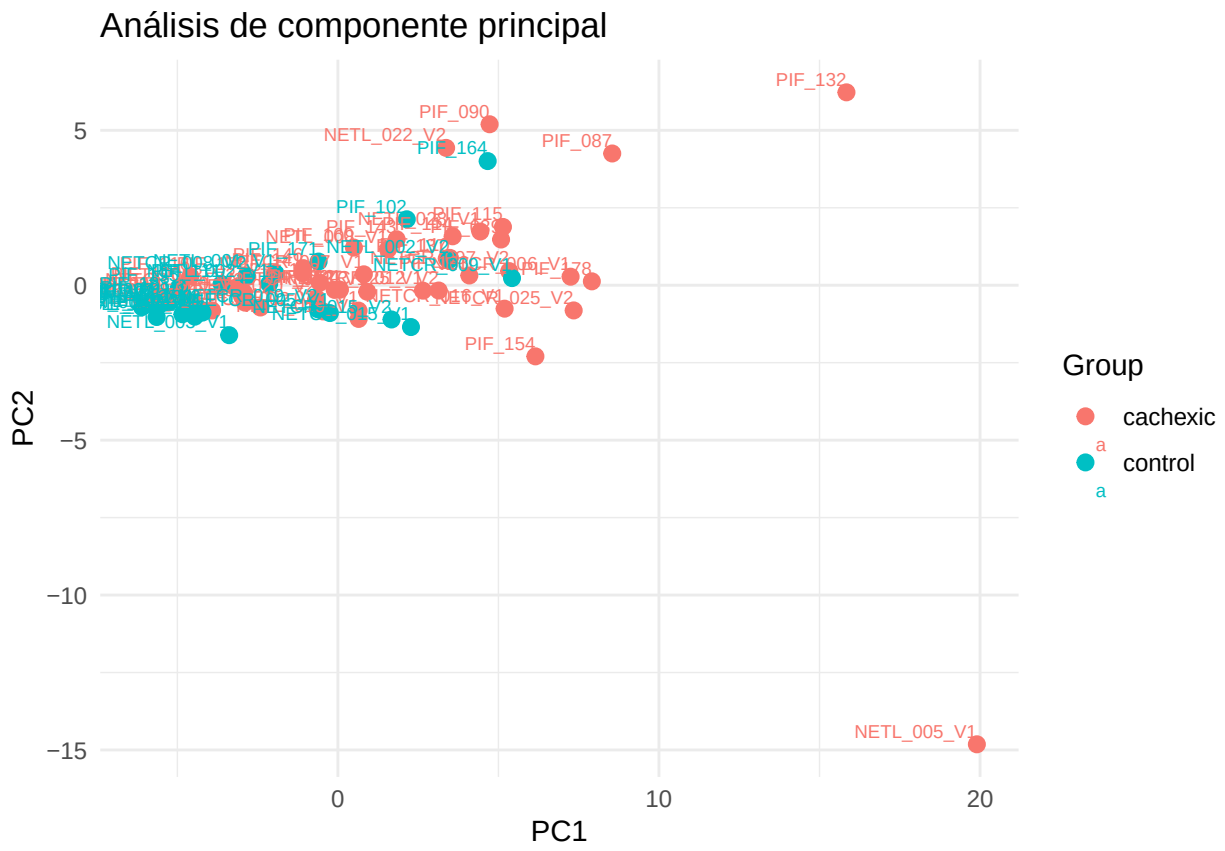
# Análisis exploratorio de los datos

## Análisis de componente principal y Heatmap

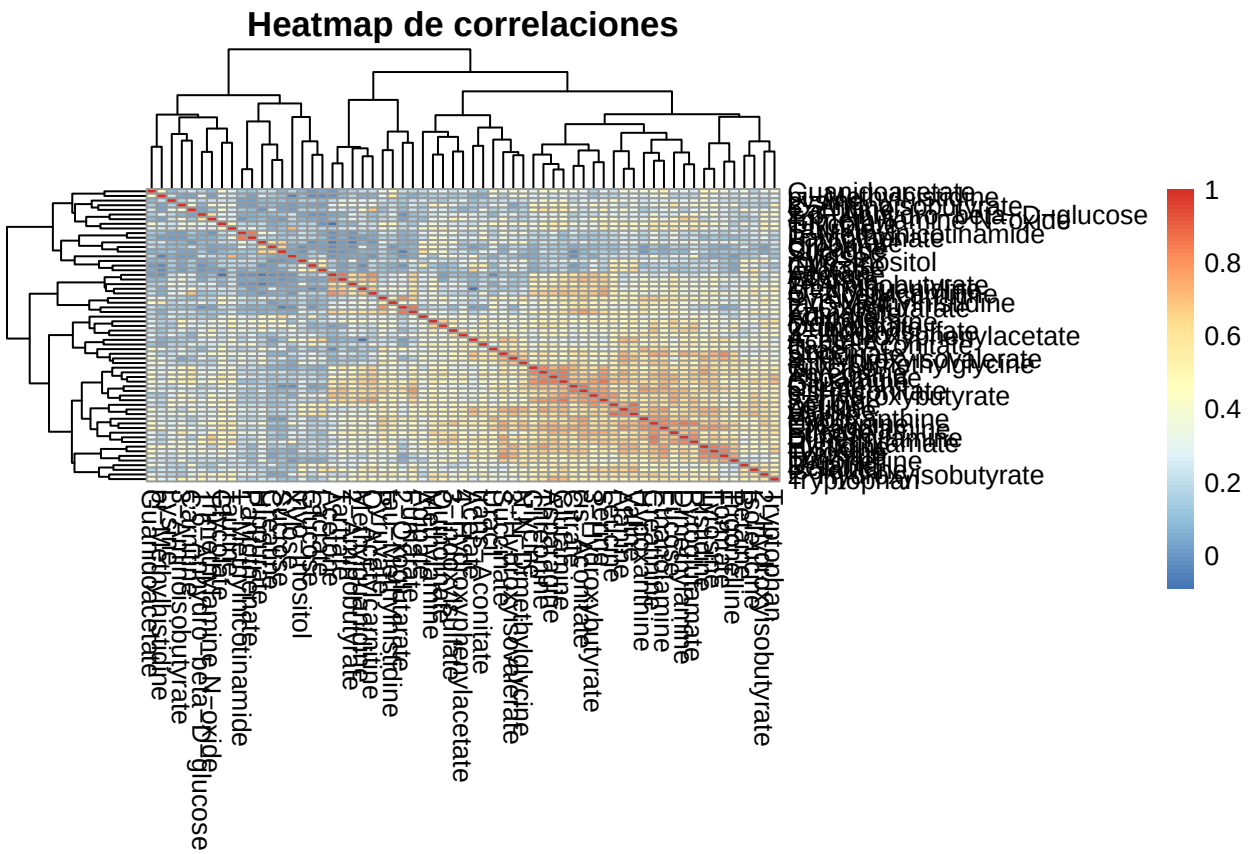
Para hacer un análisis exploratorio de los datos podemos realizar en primer lugar un par de graficaciones para ver de manera general si hay diferencias en las variables entre los dos grupos a testar ("Cachexia" y "Control"). Para ello podemos hacer un análisis de reducción dimensional para ver de manera fácil y rápida en un gráfico si existe una diferencia clara entre ambos grupos; en este caso vamos a realizar un análisis de componente principal (PCA) como método de reducción dimensional.

Por otra parte, también se puede realizar un análisis de correlación de muestras mediante un heatmap que nos aporte un dendrograma correlativo. Si en este dendrograma se distingue una separación entre los dos grupos podremos deducir que existe una diferencia metabólica entre los dos grupos (aunque no sabríamos si es significativo, ya que para ello habría que hacer los correpondientes tests estadísticos)

```
## Análisis de componente principal(PCA)

# Generamos el objeto PCA al cual añadimos primero los datos y luego
# los metadatos del set
pca <- prcomp(matriz, scale. = TRUE)
pca <- data.frame(
  PC1 = pca$x[,1],
  PC2 = pca$x[,2],
  Group = Cachexia[,2],
  PatientID = Cachexia[,1]
)

# Graficamos el PCA
library(ggplot2)
ggplot(pca, aes(x = PC1, y = PC2, color = Group, label = PatientID)) +
  geom_point(size = 2.5) +
  geom_text(vjust = -0.5, hjust = 1, size = 2.5) +
  labs(title = "Análisis de componente principal", x = "PC1", y = "PC2") +
  theme_minimal()
```




```
## Análisis de correlación de muestras (mediante un Heatmap)

library(pheatmap)
# Calculamos la matriz de correlación y gaficamos con "pheatmap"
correlaciones <- cor(matriz)
pheatmap(correlaciones,
         main = "Heatmap de correlaciones",
         clustering_distance_rows = "euclidean",
         clustering_distance_cols = "euclidean")
```




A juzgar por el PCA, existe una diferencia metabólica entre ambos grupos pero esta no es del todo clara. El primer componente principal se para en cierta medida ambos grupos (el grupo de cachexia parece tender a mayores valores); sin embargo, hay un rango relativamente amplio de valores donde se mezclan muestras de ambos grupos. Además hay que reseñar la presencia de dos outliers del grupo de cachexia muy llamativos que se salen de cualquier agrupación y que amplian el rango de valores del segundo componente principal.

En cuanto al Heatmap, se distingue la presencia de dos grupos con grandes correlaciones metabólicas entre muestras, y el dendrograma así lo atestigua. Para averigüar si esta bifurcación se corresponde con los grupos habría que echar un ojo en mayor profundidad para asignar las muestras a las ramas del dendrograma.

## Búsqueda de potenciales biomarcadores

Una cuestión de interés biomédico que podría surgir a partir de los datos es la posibilidad de utilizar alguno de los metabolitos analizados como biomarcador de la enfermedad. Para ello podríamos analizar la distribución de valores de un par de variables candidatas entre ambos grupos. Para escoger a dichas candidatas podemos calcular qué dos variables son las que más diferencia poseen entre los grupos *cachexia* y *control* y a continuación graficarlas con plots de dispersión como son el diagrama de cajas y el diagrama de violin.

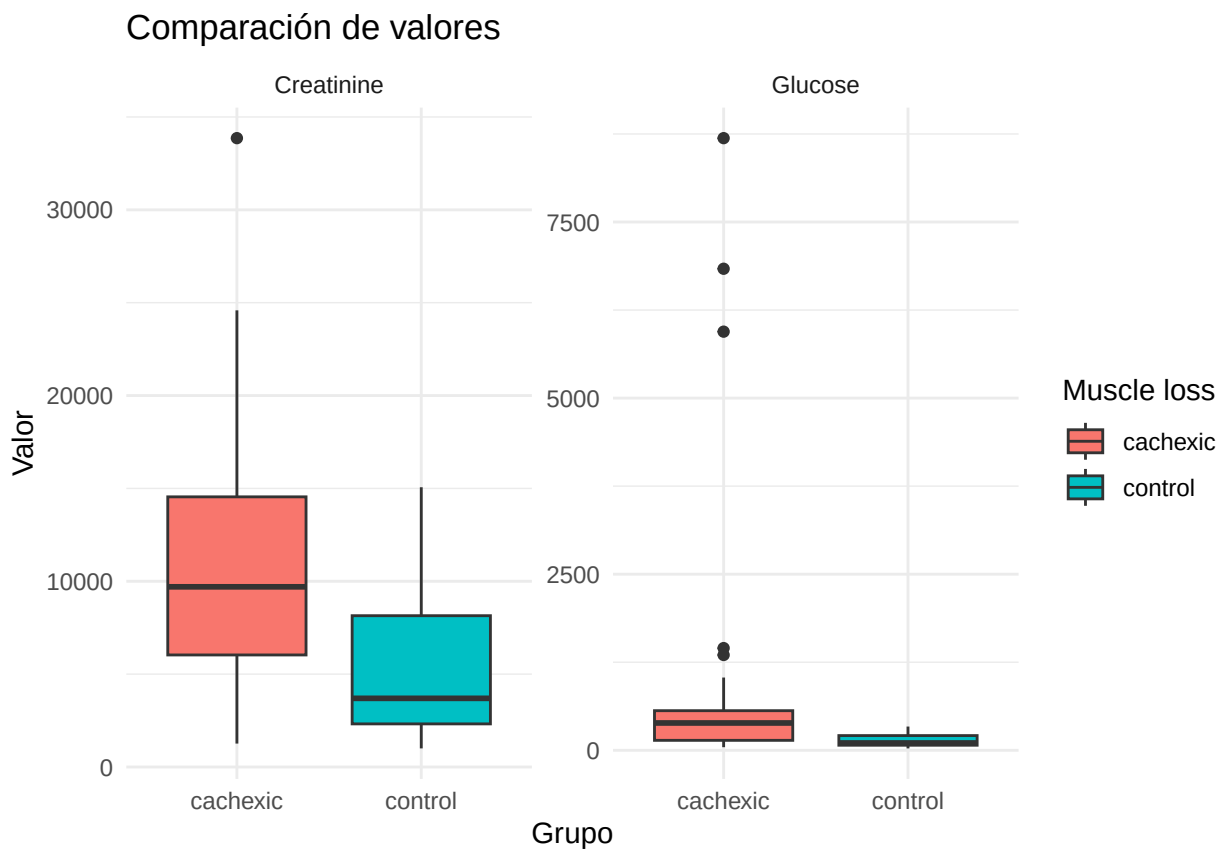
```
# Calculamos qué metabolitos son más variables de manera intergrupal
# entre las muestras de cachexia y control

library(dplyr)
library(tidyr)

top2 <- Cachexia %>%
```

```
  pivot_longer(
    cols = -c(`Patient ID`, `Muscle loss`),
    names_to = "Metabolite",
    values_to = "Value"
  ) %>%
  mutate(Value = as.numeric(Value)) %>%
  group_by(`Muscle loss`, Metabolite) %>%
  summarise(sd = sd(Value, na.rm = TRUE), .groups = "drop") %>%
  pivot_wider(names_from = `Muscle loss`, values_from = sd) %>%
  mutate(diff = abs(cachexic - control)) %>%
  arrange(desc(diff)) %>%
  slice(1:2)

top2
```

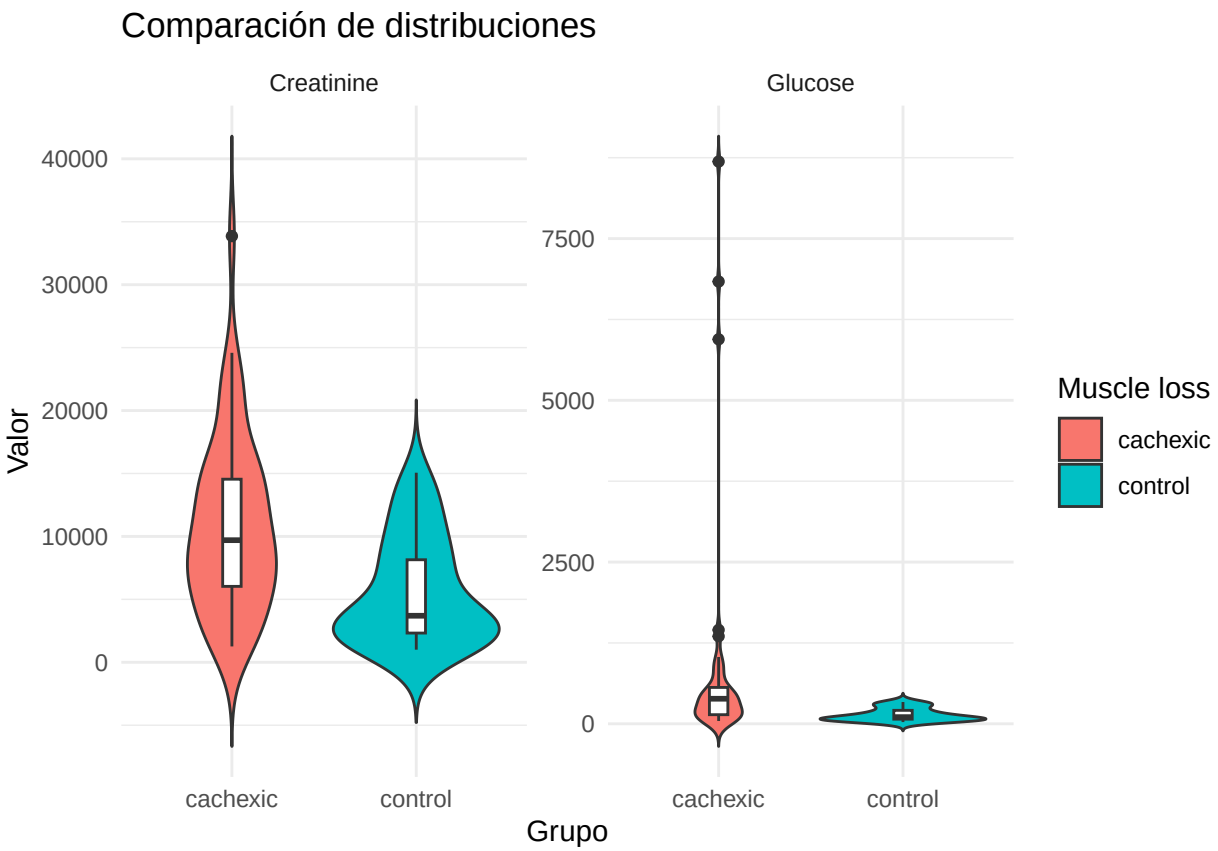
```
## # A tibble: 2 x 4
##   Metabolite cachexic control  diff
##   <chr>         <dbl>   <dbl> <dbl>
## 1 Creatinine    6905.  4229.  2677.
## 2 Glucose       1728.    99.2 1628.
```

Hemos obtenido que las variables *Creatinine* y *Glucose* son las que más varían entre los grupos *cachexia* y *control*

```
## Graficamos las diferencias de las variables "Creatinine" y "Glucose"
## entre los dos grupos

metabolitos_diferenciales <- c("Creatinine", "Glucose")

library(reshape2)
data_long <- melt(Cachexia,
                  id.vars = c("Patient ID", "Muscle loss"),
                  measure.vars = metabolitos_diferenciales,
                  variable.name = "Metabolite",
                  value.name = "Value")
data_long$Value <- as.numeric(as.character(data_long$Value))

# Graficamos un diagrama de cajas
ggplot(data_long, aes(x = `Muscle loss`, y = Value, fill = `Muscle loss`)) +
  geom_boxplot() +
  facet_wrap(~Metabolite, scales = "free_y") +
  labs(title = "Comparación de valores",
       x = "Grupo", y = "Valor") +
  theme_minimal()
```




```
# Graficamos un diagrama de violín
ggplot(data_long, aes(x = `Muscle loss`, y = Value, fill = `Muscle loss`)) +
  geom_violin(trim = FALSE) +
  geom_boxplot(width = 0.1, fill = "white") +
  facet_wrap(~Metabolite, scales = "free_y") +
  labs(title = "Comparación de distribuciones",
       x = "Grupo", y = "Valor") +
  theme_minimal()
```




En los gráficos podemos observar que existe una clara diferencia en los niveles de ambos metabolitos de manera intergrupal, tanto en la media de ambos (líneas negras de las cajas) como en la distribución general de valores. Resalta que en el caso de la glucosa existen varios outliers muy evidentes en el grupo de *cachexia* que altera mucho la distribución. Sea como fuere, la distribución de valores de glucosa es mayor en *cachexia* que en *control*, y los outliers son a niveles más altos, por lo que un nivel alto de glucosa podría servir como biomarcardor de caquexia. Lo mismo sucede, y de manera más robusta e independiente de tantos outliers, con la creatinina.