PEC1 - Análisis de Datos Ómicos

Miguel Angel Rizzo Ignacio

2025-03-29

Contents

1 RESUMEN

Este informe presenta un análisis metabolómico basado en datos de pacientes con caquexia y controles. Se ha trabajado con el dataset ¨human_cachexia.csv¨ obtenido del repositoria GitHub de metabolomicsWorkbench (https://github.com/nutrimetabolomics/metaboData), transformandolo en un objeto de clase SummarizedExperiment para su análisis. Se realizó una exploración de los datos mediante visualizaciones y análisis estadísticos, incluyendo, boxplots, histogramas y análisis de componentes principales (PCA). Los resultados sugieren diferencias metabólicas clave entre los grupos, con una mayor heterogeneidad en los pacientes caquéxicos. Se discuten las implicaciones biológicas de estos hallazgos y las limitaciones del estudio. Todo el código y los resultados están disponibles en un repositorio de GitHub.

2 OBJETIVOS

Convertir el dataset en un formato adecuado para análisis bioinformático.

Realizar un análisis exploratorio para identificar patrones en los datos.

Aplicar ténicas estadísticas y visualizaciones para interpretar diferencias entre los grupos.

Evaluar la heterogeneidad metabólica en pacientes caquéxicos frente a controles.

3 MÉTODOS

Se utilizó el dataset ¨human_cachexia.csv¨, obtenido de una fuente pública de metabolómica. Se procesaron los datos en R, creando un objeto SummarizedExperiment para manejar la información de manera estructurada. Se aplicaron técnicas estadísticas como análisis de componentes principales (PCA) y pruebas ANOVA. Se usaron herramientas de visualización como ggplot2 y ggfortify.

4 RESULTADOS

```
library(SummarizedExperiment)
library(readr)
library(dplyr)
library(RColorBrewer)
library(ggplot2)
library(ggfortify)
```

```r
# Cargar datos del dataset
data <- read.csv("human_cachexia.csv")

# Extraer la matriz de datos numéricos (medición de metabolitos)
assay_data <- as.matrix(data[,3:ncol(data)])
assay_data <- apply(assay_data, 2, as.numeric)# Convertir a numérico.
assay_data_t <- t(assay_data) # Transponer para que las filas sean metabolitos y column
colnames(assay_data_t) <- data$`Patient ID`

# Crear los metadatos

# Metadatos de las muestras (colData)
colData <- data[, c("Patient.ID", "Muscle.loss")]
rownames(colData) <- data$`Patient.ID`
colData$`Muscle.loss` <- as.factor(colData$`Muscle.loss`)

# Metadatos de los metabolitos (rowData)
metabolite_names <- colnames(data)[3:ncol(data)]
rowData <- data.frame(metabolite = metabolite_names, row.names = metabolite_names)

# Crear el objeto SummarizedExperiment
se <- SummarizedExperiment(
  assays = list(counts = assay_data_t),
  colData = colData,
  rowData = rowData
)

# Resumen del Objeto creado
print(se)
```

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

Se puede apreciar que en dimensiones muestra los 63 metabolitos cuyos nombres se encuentran en rownames, los 77 pacientes, cuyas denominaciones se encuentran en colnames. Assays (1) indica que el objeto contiene una matriz de datos, donde cada celda tiene un dato

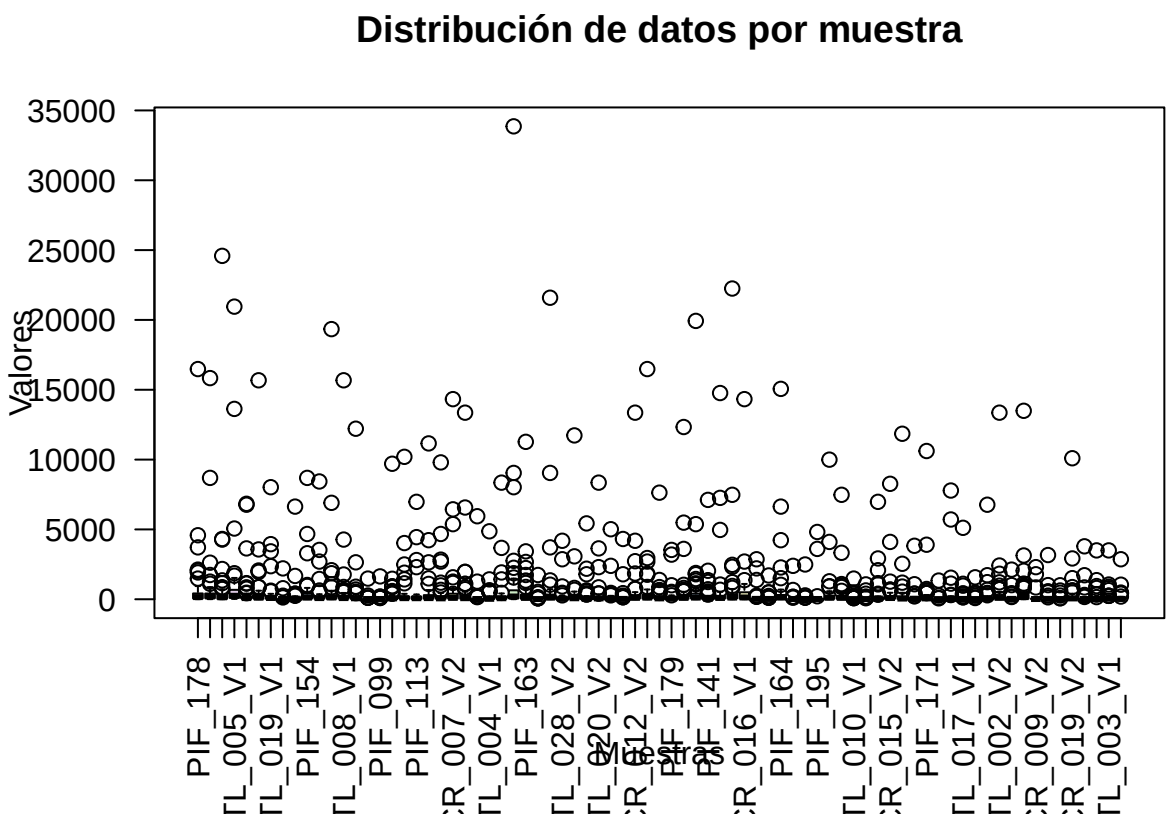
de un metabolito. En rowData names te indica que son metabolitos los datos grabados. Y en colData names hay 2 variables asociadas a las muestras.

```
# Resumen de los datos
summary(assay(se)[, 1:2])
```

```
##     PIF_178            PIF_087
##  Min.   :    5.58   Min.   :    7.69
##  1st Qu.:   52.72   1st Qu.:   78.66
##  Median :  154.47   Median :  208.51
##  Mean   :  699.86   Mean   :  708.30
##  3rd Qu.:  416.24   3rd Qu.:  412.10
##  Max.   :16481.60   Max.   :15835.35
```

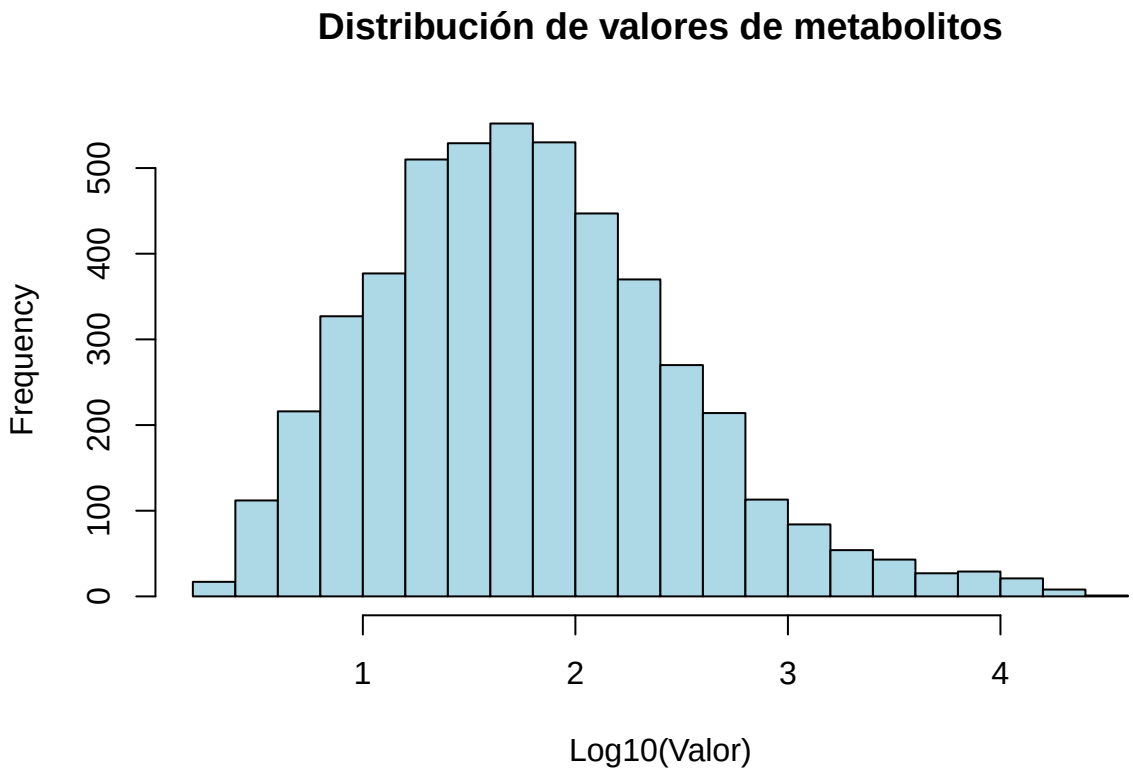
Resumen de los datos de cada paciente.

```
# Boxplot para visualizar la distribución de valores por muestra
boxplot(assay(se),
        main = "Distribución de datos por muestra",
        xlab = "Muestras",
        ylab = "Valores",
        las = 2, col = brewer.pal(8, "Pastel1"))
```



Dado que son muchos los pacientes, al querer visualizar todos los datos al mismo tiempo es difícil apreciar las cajas de Boxplot, pero se puede apreciar que la mayoria de estas cajas se encuentra casi al mismo nivel y que hay muchos residuos outliers, aunque la mayoria están debajo de 5000.

```
# Histograma con logaritmo para observar la distribucion
hist(log10(assay(se) + 1),
     breaks = 30,
     main = "Distribución de valores de metabolitos",
     xlab = "Log10(Valor)",
     col = "lightblue",
     border = "black")
```

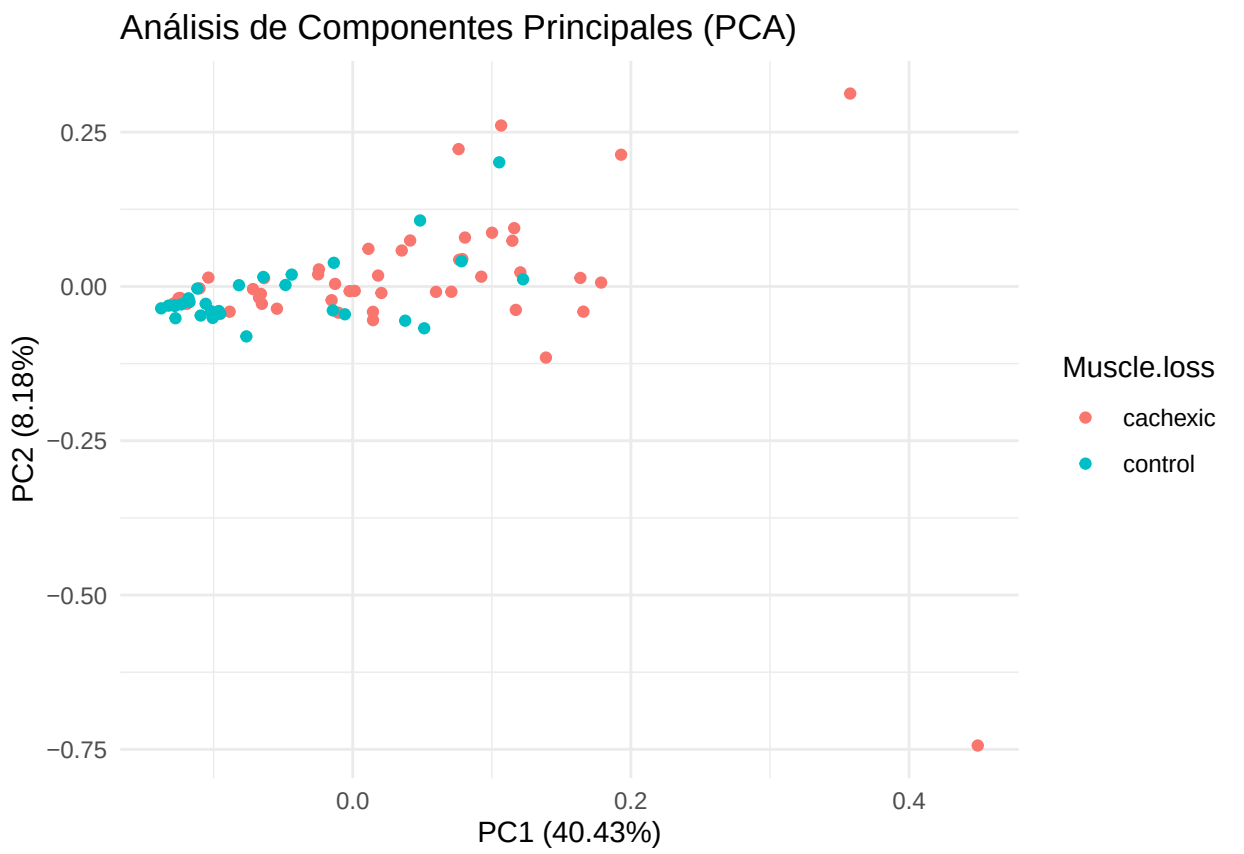


Se puede apreciar que la distribución de los valores tiene forma de campana por lo que indicaría que la distribución es normal.

```
# Analisis de Componentes Principales (PCA)
pca <- prcomp(t(assay(se)), scale = TRUE)

# Visualizar los datos
autoplot(pca, data = colData, colour = "Muscle.loss") +
```

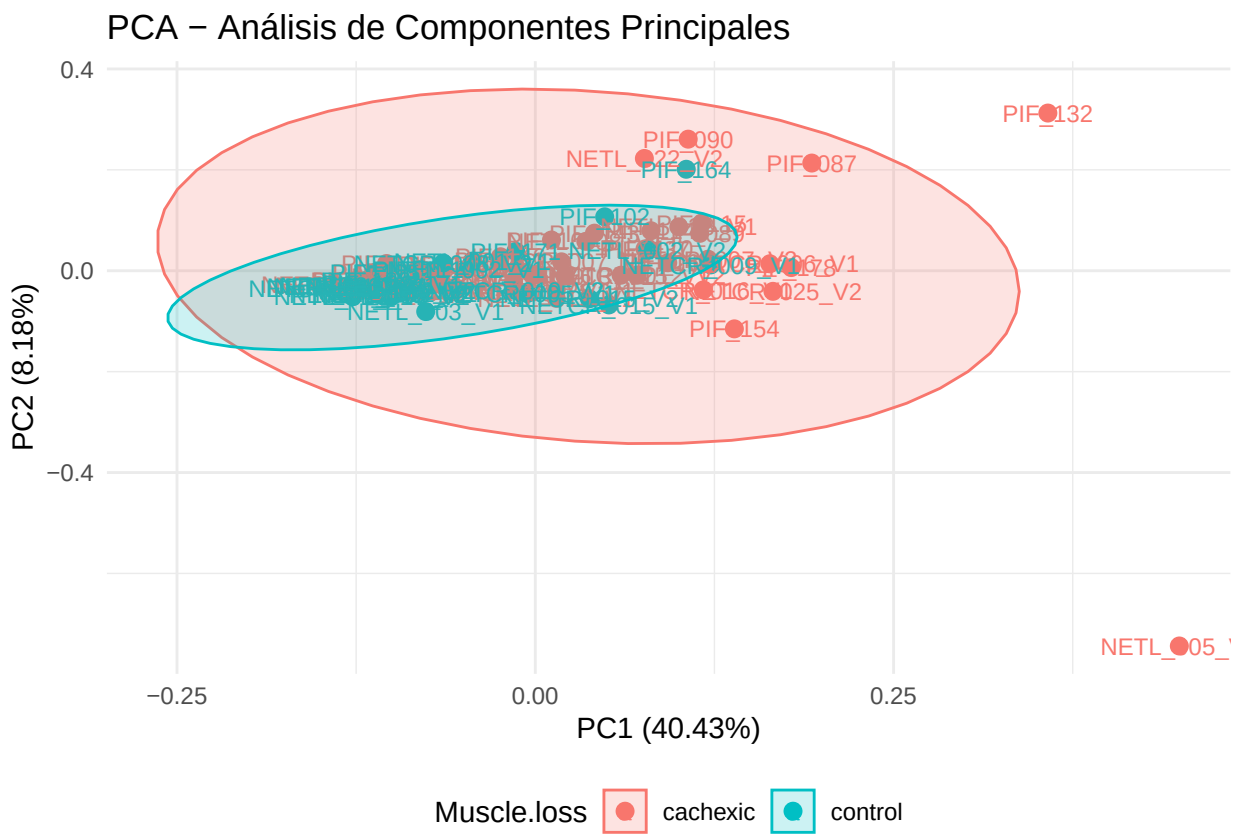
```
  ggtitle("Análisis de Componentes Principales (PCA)") +
  theme_minimal()
```



Se puede aprecia que los datos en su totalidad tiene forma de cono con base a la derecha. Los datos de los controles tienden a estar menos disperson comparado con lo de los caquéticos, lo que indicaría que la perdida muscular afecta de manera diferente a cada paciente, generando más variabilidad en sus perfiles metabólicos. Por lo que se puede concnluir que hay una mayor heterogeneidad en la respuesta de los pacientes enfermos.

```
# PCA plot
autoplot(pca,
         data = colData,
         colour = "Muscle.loss",
         shape = 16,
         size = 3,
         label = TRUE,
         label.size = 3,
         frame = TRUE,
         frame.type = 'norm')+
  ggtitle("PCA - Análisis de Componentes Principales") +
  theme_minimal() +
```

```
  theme(legend.position = "bottom")
```



Otra representación gráfica la variabilidad de los datos metabolómicos entre los controles y la caquéticos. Se podría decir que algunos pacientes caquéticos tienen actividad metabólica casi igual al de los controles y que algunos tienen una muy diferente.

```
# Análisis de ANOVA
anova_pca <- aov(pca$x[,1] ~ colData$Muscle.loss)
summary(anova_pca)
```

```
##                      Df Sum Sq Mean Sq F value   Pr(>F)
## colData$Muscle.loss   1    323   323.0   15.02 0.000226 ***
## Residuals            75   1613    21.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se puede aprecia que el valor p es de 0.000226, lo que significa que es estadisticamente significativo ya que es menor a 0.05. Por lo que hay diferencias entre los grupos (Control y Caquéxico) en el primer componente principal. Esto respalda a la interpretación de que los perfiles metabólicos entre los grupos son diferentes.

5 DISCUSIÓN

Los resultados sugieren que existen diferencias metabólicas entre los pacientes caquécos y controles. La mayor dispersión en el grupo caquéxico podría indicar diferentes subtipos o variabilidad en la respuesta metabólica a la enfermedad. Sin embargo, la separación entre los grupos en el PCA no es absoluta, lo que siguiere que otros factores pueden estar influyendo en los perfiles metabólicos. Algunas limitaciones del estudio incluyen el tamaño muestral y la ausencia de información adicional sobre los pacientes.

Es importante generar de manera visual los valores obtenidos para poder meditar sobre la posible asociación o diferencias entre los grupos de interes.

El empleo de R para estos tipos de estudio ayuda bastante en poder realizar este tratamiento de datos ya que nos ahorra tanto tiempo como dinero en tratar de manejar todos los datos de manera manual y con exactitud.

Es importante trabajar con SummarizedExperiment ya que es una herramienta clave en bioinformática y biología computacional para organizar, almacenar y analizar grandes volúmenes de datos biológicos complejos. Su estructua modular que integra datos cuantitativos, metadatos y arquitectura flexible permite realizar análisis avanzados, garantizar la reproducibilidad de los resultados y facilitar la integración de datos provenientes de diversas fuentes. Además, su compatibilidad con herramientas estadísticas y de visualización lo convierte en un componente esencial para la gestión y análisis de datos ómicos, y es una pieza central en el análisis de datos biológicos en entorno de investigación y profesionales.

6 CONLUSIONES

Se identificaron diferencias metabolómicas entre paceintes caquéxicos y controles.

El PCA reveló una separación parcial entre los grupos, con mayor variabilidad en los pacientes caquéxicos.

Se recomendaría complementar el análisis con modelos más avanzadoas para mejorar la clasificación de los grupos.

7 REFERENCIAS

Para más información les dejo el enlace de mi GitHub [Respositorio en GitHub] (https://github.com/MiguelARI/Rizzo-Ignacio-Miguel-Angel-PEC1)