Análisis de Datos Ómicos PEC1

Jesús Cea García

2 de abril, 2025

Contents

1. Abstract

En este trabajo se construyó un objeto *SummarizedExperiment* a partir de datos de metabolómica en orina de pacientes caquéxicos y controles. Se exploraron las concentraciones de metabolitos por grupo, observando diferencias claras entre ambos. Un análisis de componentes principales (PCA) evidenció una separación más o menos marcada entre caquéxicos y controles, sin mostrar *batch effect* asociado al tipo de cáncer. Además, un dendrograma confirmó esta separación a nivel jerárquico. Finalmente, se realizó un análisis de expresión diferencial y un volcano plot y se identificaron metabolitos con diferencias significativas de concentración entre los grupos, los cuales pueden ser potencialmente relevantes como biomarcadores de caquexia.

2. Objetivos

Los objetivos de este trabajo se enumeran a continuación:

1. Integrar los datos metabolómicos de un archivo .csv, separando la información de concentraciones, metadatos y características de los metabolitos, para construir un objeto de clase *SummarizedExperiment*.
2. Explorar las diferencias globales en las concentraciones de metabolitos entre pacientes caquéxicos y controles.
3. Discriminar entre ambos grupos mediante técnicas de análisis multivariante.
4. Detectar y caracterizar posibles efectos de batch relacionados con el tipo de cáncer.
5. Identificar metabolitos diferencialmente expresados que puedan actuar como biomarcadores asociados a la caquexia.

3. Métodos

3.1. Origen de los datos

Los datos utilizados provienen del repositorio habilitado para la PEC de la asignatura, los cuales parecen provenir del paquete de R specmine.datasets. Este dataset contiene datos provenientes del estudio de Eisner et al. (2010), en el que se analizaron perfiles de 63 metabolitos en muestras de orina (concentración en μM) de pacientes oncológicos mediante espectroscopía de resonancia magnética nuclear (1H-NMR). El estudio tuvo como objetivo identificar metabolitos asociados a la pérdida de masa muscular en pacientes con caquexia. El conjunto de datos estaba compuesto por 77 pacientes, distribuidos en 47 caquéxicos y 30 controles. Los IDs de las muestras incluían siglas como PIF, NETCR y NETL, que se consideraron como tipos de cáncer y se incorporaron al análisis para aportar una dimensión adicional al análisis posterior.

3.2. Construcción del objeto *SummarizedExperiment*

Se cargaron los datos de metabolómica desde un archivo .csv y se realizó una preprocesamiento inicial, que incluyó la creación de una columna adicional con el tipo de cáncer y la renombración de las muestras para distinguir claramente entre pacientes caquéxicos y controles en el ID. Posteriormente, se extrajeron los metadatos de las muestras y la matriz de concentración de metabolitos. Finalmente, se construyó el objeto de clase *SummarizedExperiment*, integrando la matriz de expresión y los metadatos de las muestras con la función *SummarizedExperiment()* siguiendo el tutorial de Morgan et al. 2024.

3.3. Normalización de los datos

Para estabilizar la varianza y reducir la influencia de valores extremos, las concentraciones de metabolitos fueron transformadas mediante logaritmo en base 2. Esta normalización se aplicó directamente a la matriz de expresión (concentración) contenida en el objeto *SummarizedExperiment* con la función log2().

3.4 Análisis univariante

Como primer análisis exploratorio, se calculó la mediana de las concentraciones normalizadas de metabolitos para cada muestra. . Posteriormente, se representaron las medianas mediante un boxplot por grupo, mostrando las diferencias en las distribuciones de las concentraciones globales de metabolitos entre ambos conjuntos de pacientes.

3.4. Análisis multivariante y Batch Effect

Se realizó un análisis de componentes principales (PCA) utilizando las concentraciones normalizadas de metabolitos. Las muestras fueron representadas diferenciando tanto el grupo de estudio (caquéxicos y controles) como el tipo de cáncer (PIF, NETCR, NETL) mediante codificación de colores, con el objetivo de explorar la variabilidad y evaluar posibles efectos de agrupación o batch effect. Además, se realizó un análisis de clustering jerárquico mediante distancias euclídeas, representando los resultados en un dendrograma con separación visual en dos grupos para ver si existe separación entre caquéxicos y controles.

3.5. Análisis de expresión diferencial

Se realizó un análisis de expresión diferencial para comparar las concentraciones de metabolitos entre pacientes caquéxicos y controles. Para cada metabolito, se aplicó un test t de Student, calculando además el log2 fold change entre ambos grupos. Los resultados se representaron mediante un volcano plot con la librería *EnhancedVolcano()*, resaltando aquellos metabolitos con cambios significativos en concentración.

4. Resultados

4.1 Caracterización inicial de los datos

Tras la construcción del *SummarizedExperiment* se realizó una primera exploración del objeto donde se visualizaron los primeros registros de metadatos de las muestras (Tabla 1), la distribución de los pacientes en función de la condición (Tabla 2) y los primeros registros de la matriz de expresión (Tabla 3). Se observa que los metadatos se crearon correctmente, con los IDs de los pacientes, el tipo de cáncer y la condición de pérdida de masa muscular (47 caquéxicos y 30 controles). También se observa que la matriz de expresión se genéro, con los 63 metabolitos en las filas y sus concentraciones por paciente (columnas).

Table 1: Metadatos de las muestras

	Muscle.loss	Cancer_Type
CA-1	cachexic	PIF
CA-2	cachexic	PIF
CA-3	cachexic	PIF
CA-4	cachexic	NETL
CA-5	cachexic	PIF
CA-6	cachexic	PIF

Table 2: Distribución de pacientes por grupo

Condición	Frecuencia
cachexic	47
control	30

Table 3: Matriz de expresión

	CA-1	CA-2	CA-3	CA-4	CA-5	CA-6	CA-7	CA-8	CA-9	CA-10
1,6-Anhydro-beta-D-glucose	40.85	62.18	270.43	154.47	22.20	212.72	151.41	31.50	51.42	117.92
1-Methylnicotinamide	65.37	340.36	64.72	52.98	73.70	31.82	36.60	6.82	30.27	52.46
2-Aminobutyrate	18.73	24.29	12.18	172.43	15.64	18.36	8.67	4.18	7.54	19.49
2-Hydroxyisobutyrate	26.05	41.68	65.37	74.44	83.93	80.64	42.52	12.94	34.81	72.24
2-Oxoglutarate	71.52	67.36	23.81	1199.91	33.12	47.94	223.63	25.03	80.64	73.70
3-Aminoisobutyrate	1480.30	116.75	14.30	555.57	29.67	17.46	56.26	8.67	17.99	57.97

4.2 Análisis univariante de la concentración de metabolitos por condición.

En primer lugar, se realizó una exploración preliminar de las concentraciones de metabolitos sin normalizar. En la Figura 1 se representa la distribución de los valores de concentración para cada muestra, diferenciando visualmente a los pacientes caquéxicos y controles. Se observa una elevada dispersión y la presencia de valores extremos, con algunas muestras mostrando metabolitos con concentraciones notablemente superiores al resto. Debido a esto, se procedió al normalizado de los datos con una transformación logarítmica en base 2.

Tras el proceso de normalización, se representa en la Figura 2 la distribución de las concentraciones de metabolitos para cada muestra. Se observa una notable reducción de la dispersión y de los valores extremos respecto a las concentraciones sin normalizar. Además, las distribuciones de ambos grupos (caquéxicos y controles) presentan ahora rangos de concentración más homogéneos.

A continuación, se compararon las medianas de concentración de metabolitos a nivel de muestra entre pacientes caquéxicos y controles. La Figura 3 muestra los boxplot resultantes, donde se observa que las muestras caquéxicas presentan, en general, valores superiores a las de los controles.

4.3 Análisis multivariante: PCA y Batch Effect

El objetivo del PCA fue explorar si existe separación entre los individuos caquéxicos y los controles en función de las concentraciones de metabolitos en orina, reduciendo así la dimensionalidad de los datos originales (63 metabolitos por muestra) a un espacio bidimensional formado por las dos primeras componentes principales, que explican la mayor parte de la varianza. En la Figura 4 se observa una cierta separación entre los individuos control (en azul) y los caquéxicos (en rojo) principalmente a lo largo de la PC1, que es la que captura la mayor variabilidad de los datos, lo cual sugiere que existen diferencias en el perfil metabolómico entre ambos grupos.

Posteriormente y con el objetivo de comprobar la presencia de un posible batch effect asociado al tipo de cáncer, se representaron las muestras en el espacio de las dos primeras componentes principales diferenciando los tipos de los mismos (PIF, NETCR y NETCL). En la Figura 5 no se aprecia una separación clara de las muestras según su tipo de cáncer, indicando que no parece haber un batch effect por esta variable.

Por último, en la Figura 6, se muestra el dendrograma resultante del análisis de agrupamiento jerárquico, construido a partir de las distancias euclídeas entre las muestras. Este dendrograma permite observar la proximidad entre las muestras en función de sus perfiles metabolómicos. Aunque no se aprecia una separación totalmente definida, se puede identificar cierta tendencia a la agrupación de los individuos según su condición (caquéxicos y controles), lo que es consistente con las observaciones obtenidas previamente mediante el PCA.

4.4 Análisis de expresión diferencial

Se realizó un análisis de expresión diferencial para identificar metabolitos con diferencias significativas entre pacientes caquéxicos y controles. Los resultados se representan en la Figura 7 mediante un volcano plot, destacando aquellos metabolitos que cumplen simultáneamente los criterios de significación estadística ($p < 0.05$) y de magnitud de cambio ($|\log2FC| > 0.5$) en rojo. Entre ellos, se identificaron el adipato, el 3-hidroxibutirato y el fumarato. Se aprecia, además, que todos los metabolitos presentan un valor de log2 fold change positivo, lo cual es coherente, ya que previamente se había observado que las concentraciones globales de metabolitos son, en general, superiores en las muestras de pacientes caquéxicos en comparación con las de los controles.

5. Discusión

La caquexia es un síndrome asociado al cáncer caracterizado por una pérdida continua de masa muscular y peso corporal, acompañado de alteraciones metabólicas e inflamación sistémica (Ni & Zhang, 2020) . Los resultados obtenidos en este trabajo muestran diferencias claras en el perfil metabolómico de pacientes caquéxicos frente a controles, destacando una tendencia general a mayores concentraciones de metabolitos en orina en los pacientes con caquexia, lo cual es esperable, ya que la degradación muscular característica de la caquexia provoca la liberación de metabolitos al torrente sanguíneo, que posteriormente son eliminados a través de la orina. Este comportamiento se ha reflejado tanto en los análisis multivariantes como en el análisis de expresión diferencial, donde metabolitos como el adipato, 3-hidroxibutirato y fumarato mostraron diferencias significativas entre ambos grupos. De hecho, el 3-hidroxibutirato, es un cuerpo cetónico que ha sido previamente relacionado con estados de caquexia y pérdida de peso severa en pacientes oncológicos, por lo que su presencia elevada en pacientes caquéxicos podría actuar como un marcador temprano de deterioro nutricional (Boguszewicz et al., 2019). En cuanto a las limitaciones del análisis, al hacer el análisis de expresión diferencial hubiese sido conveniente realizar alguna corrección por comparaciones múltiples para evitar los falsos positivos al realizar tantos tests simultáneamente.

6. Conclusiones

El análisis realizado ha permitido identificar diferencias en la concentración de metabolitos entre controles y caquéxicos utilizando técnicas de análisis univariante y multivariante, identificando además metabolitos que podrían ser utilizados como marcadores de este síndrome.

7. Figuras

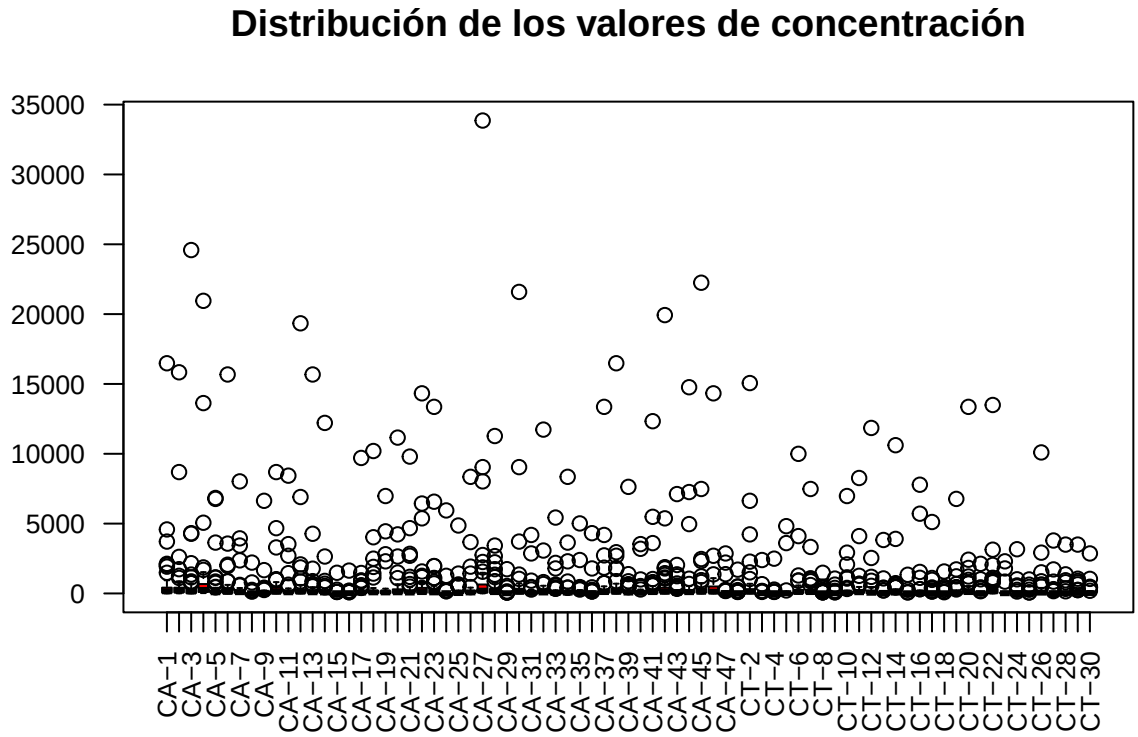

Figure 1: Distribución de los valores de concentración sin normalizar.

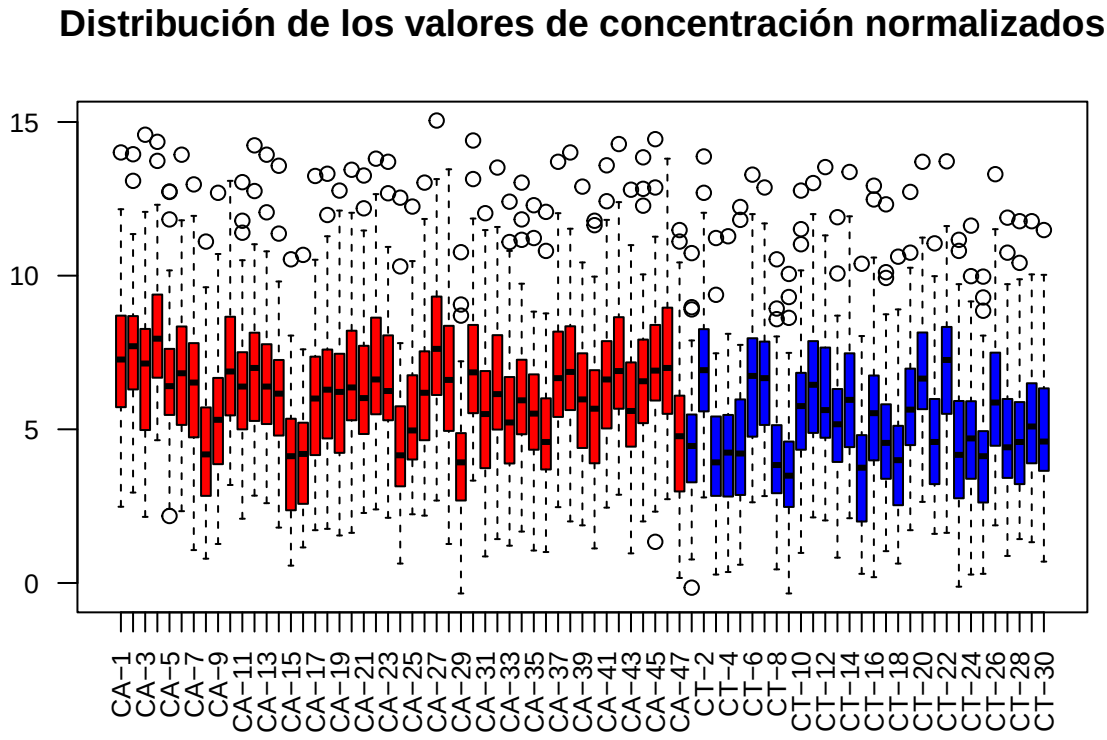

Figure 2: Distribución de los valores de concentración normalizados.

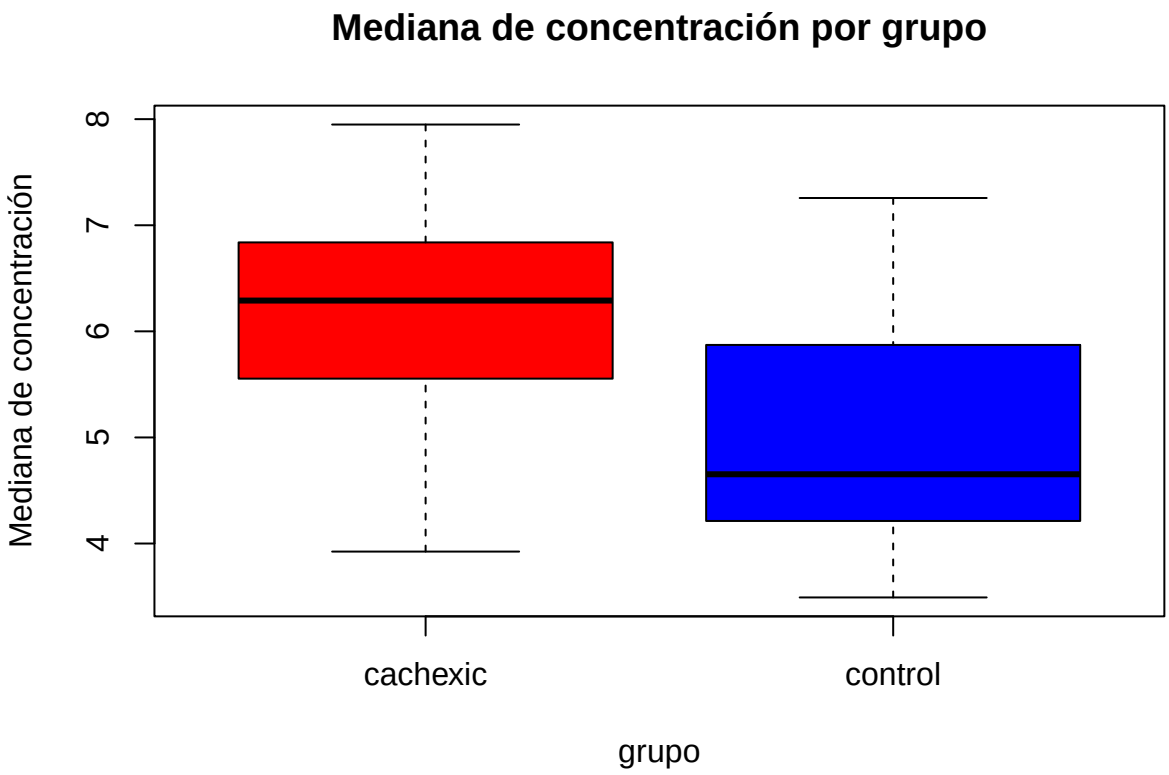

Figure 3: Comparación de la mediana de concentración entre caquéxicos y controles.

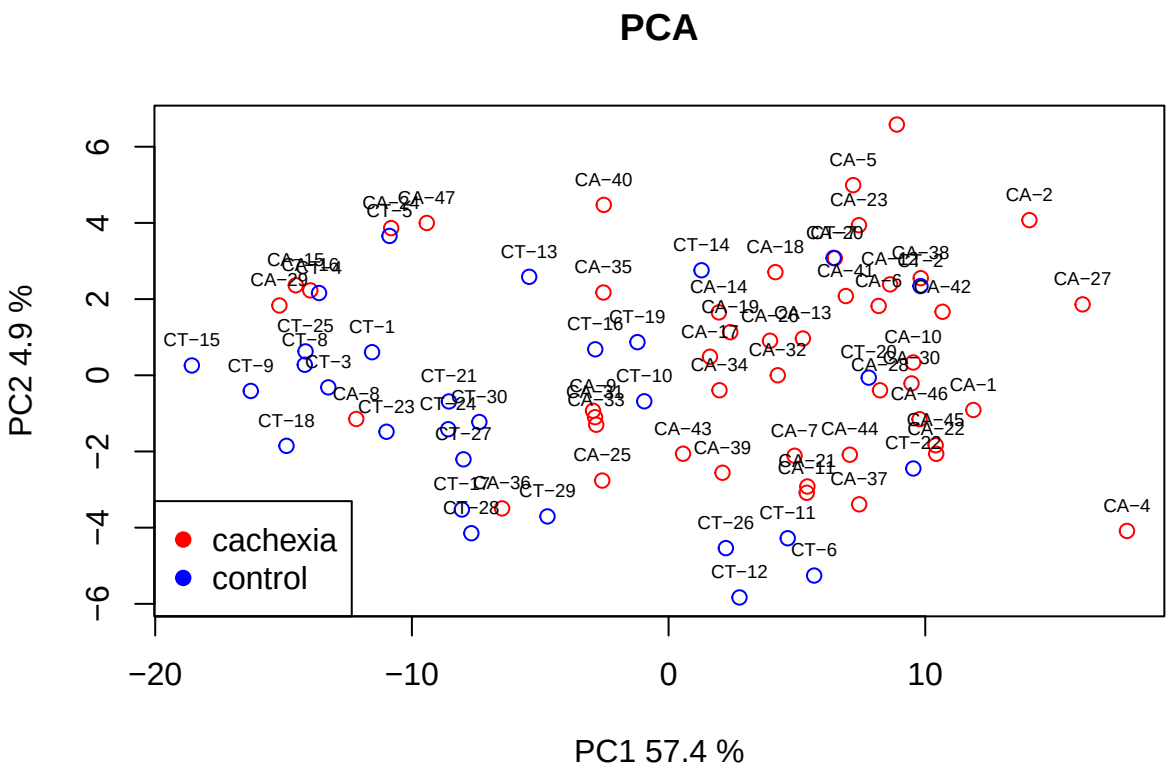

Figure 4: Análisis PCA de las concentraciones de metabolitos por condición

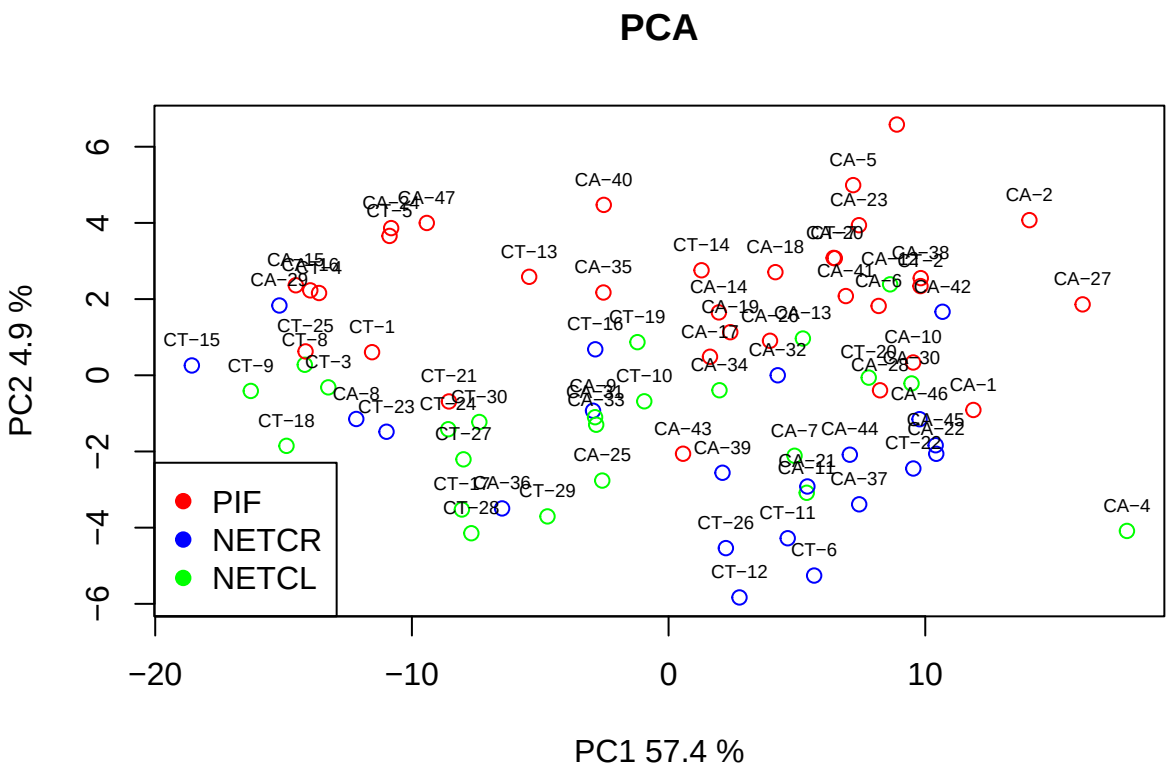

Figure 5: Análisis PCA de las concentraciones de metabolitos por tipo de cáncer

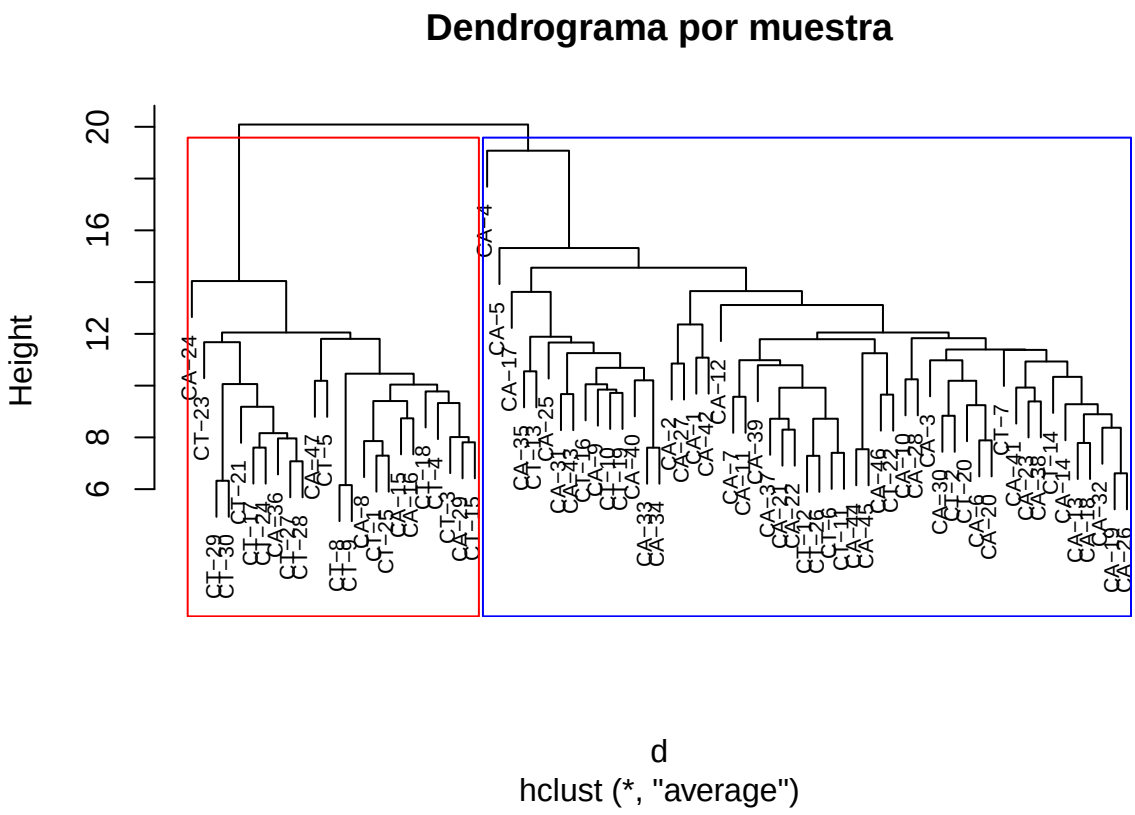

Figure 6: Agrupamiento jerárquico de las muestras basado en las concentraciones de metabolitos.

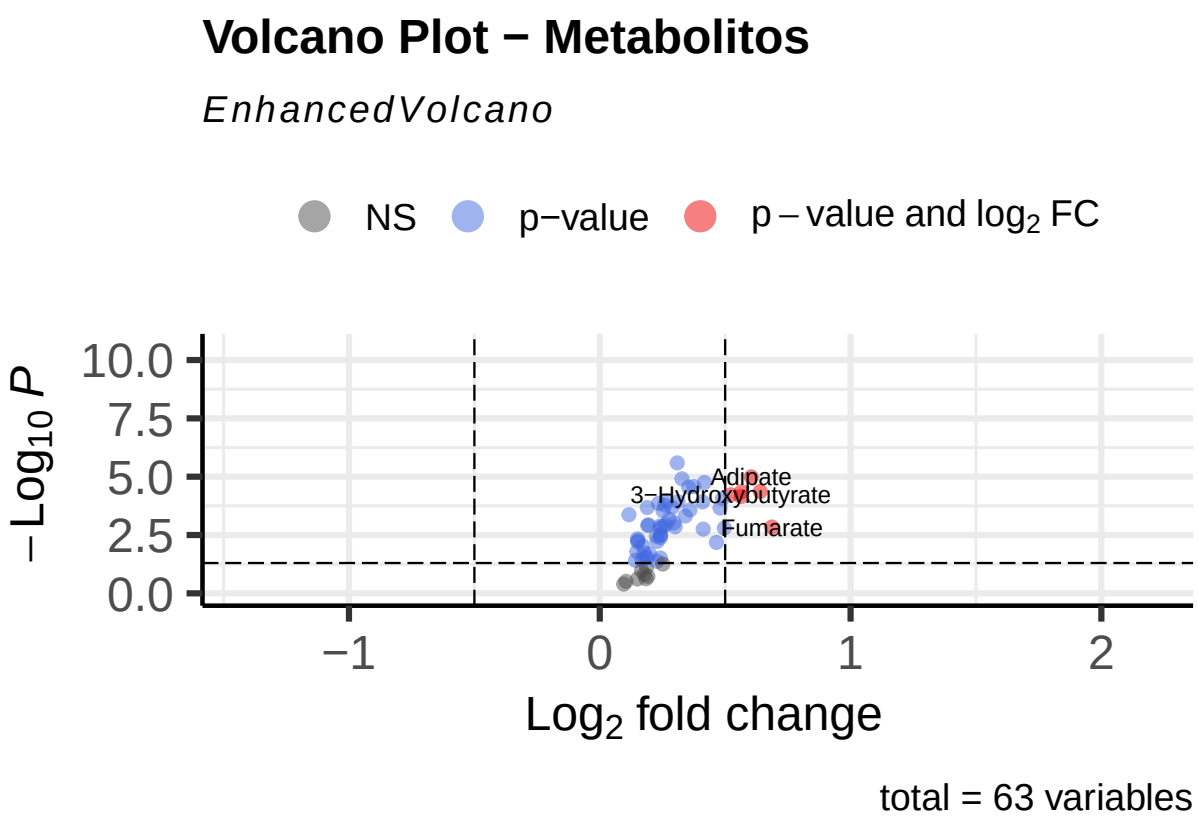

Figure 7: Volcano plot de la expresión diferencial de metabolitos entre caquéxicos y controles.

8. Referencias

Repositorio GitHub: https://github.com/jcea97/Cea_Garcia_Jesus_PEC1.git

Eisner, R., Stretch, C., Eastman, T., Xia, J., Hau, D., Damaraju, S., Greiner, R., Wishart, D. S., & Baracos, V. E. (2011). Learning to predict cancer-associated skeletal muscle wasting from 1H-NMR profiles of urinary metabolites. Metabolomics, 7(1), 25-34. https://doi.org/10.1007/s11306-010-0232-9

Morgan, M., Obenchain, V., Hester, J., & Pagès, H. (2024). SummarizedExperiment: A container (S4 class) for matrix-like assays. R package version 1.36.0, https://bioconductor.org/packages/SummarizedExperiment.

Ni, J., & Zhang, L. (2020). Cancer Cachexia: Definition, Staging, and Emerging Treatments. Cancer Management and Research, 12, 5597-5605. https://doi.org/10.2147/CMAR.S261585

Boguszewicz, Ł., Bieleń, A., Mrochem-Kwarciak, J., Skorupa, A., Ciszek, M., Heyda, A., Wygoda, A., Kotylak, A., Składowski, K., & Sokół, M. (2019). NMR-based metabolomics in real-time monitoring of treatment induced toxicity and cachexia in head and neck cancer: a method for early detection of high risk patients. Metabolomics : Official journal of the Metabolomic Society, 15(8), 110. https://doi.org/10.1007/s11306-019-1576-4

ANEXOS

Carga y preparación de los datos

```
# Cargo el csv
cachexia <- read.csv("human_cachexia.csv", check.names = FALSE)

# Creo la columna Cancer_Type
cachexia$Cancer_Type <- sub("_.*", "", cachexia$`Patient ID`)

# Reordeno columnas
cachexia <- cachexia %>% relocate(Cancer_Type, .after = 2)

# Renombramos pacientes
cachexia$`Patient ID`[cachexia$`Muscle loss` == "cachexic"] <- paste0("CA-", 1:sum(cachexia$`Muscle loss
cachexia$`Patient ID`[cachexia$`Muscle loss` == "control"] <- paste0("CT-", 1:sum(cachexia$`Muscle loss

# Preparo los metadatos
sample_metadata <- cachexia[, 1:3]
rownames(sample_metadata) <- cachexia$`Patient ID`

# Preparo la matriz de expresión
expr_matrix <- as.matrix(cachexia[, -(1:3)])
rownames(expr_matrix) <- sample_metadata$`Patient ID`
expr_matrix <- t(expr_matrix)

# Creo el SummarizedExperiment
se <- SummarizedExperiment(
  assays = list(counts = expr_matrix),
  colData = sample_metadata
)
```

Generación de tablas de los metadatos y matriz de expresión

```r
knitr::kable(head(as.data.frame(colData(se))[ , -1]), caption = "Metadatos de las muestras")

tabla_grupo <- as.data.frame(table(cachexia$`Muscle loss`))
colnames(tabla_grupo) <- c("Condición", "Frecuencia")

knitr::kable(tabla_grupo, caption = "Distribución de pacientes por grupo")
knitr::kable(head(expr_matrix[, 1:10]), caption = "Matriz de expresión")
```

Gráficos de distribución de los valores de concentración

```r
# Accedo a la matriz de expresión
expr_unn <- assay(se)

# Defino los grupos y sus colores
grupos <- colData(se)$`Muscle loss`
colores <- ifelse(grupos == "cachexic", "red", "blue")

# Creo el boxplot
boxplot(expr_unn,
        las = 2, col = colores,
        cex.axis = 0.8,
        main = "Distribución de los valores de concentración")
```

```r
#Normalización
expr_norm <- log2(assay(se))
# Crear boxplot
boxplot(expr_norm,
        las = 2,
        col = colores,
        cex.axis = 0.8,
        main = "Distribución de los valores de concentración")
```

```r
# Obtener expresión y grupos
grupo <- colData(se)$`Muscle loss`
# Calcular medianas por muestra
medianas <- apply(expr_norm, 2, median)

# Crear boxplot por grupo
boxplot(medianas ~ grupo,
        col = c("red", "blue"),
        main = "Mediana de concentración por grupo",
        ylab = "Mediana de concentración")
```

Análisis de componentes principales (PCA)

```r
pc <-prcomp(t(expr_norm), scale=FALSE) #PCA sobre los datos normalizados
loads <- round(pc$sdev^2 / sum(pc$sdev^2) * 100, 1) #Varianza explicada por PC
```

```r
# Representación de las muestras en el espacio definido por las dos primeras componentes principales (P
# diferenciando por condición

pacientes_ID <- cachexia[ ,1]
```

```
xlab<-c(paste("PC1",loads[1],"%"))
ylab<-c(paste("PC2",loads[2],"%"))

plot(pc$x[,1:2],xlab=xlab,ylab=ylab, col=colores,
     main ="PCA")

legend("bottomleft",
       legend=c("cachexia", "control"),
       col=c("red", "blue"),
       pch=19, xpd=TRUE)

text(pc$x[,1],pc$x[,2],pacientes_ID, pos=3, cex=.6)
```

```
grupo_batch <- colData(se)$Cancer_Type

colores_batch <- ifelse(grupo_batch == "PIF", "red",
                  ifelse(grupo_batch == "NETCR", "blue",
                         "green"))

plot(pc$x[,1:2],xlab=xlab,ylab=ylab, col=colores_batch,
     main ="PCA")

legend("bottomleft",
       legend=c("PIF", "NETCR", "NETCL"),
       col=c("red", "blue", "green"),
       pch=19, xpd=TRUE)

text(pc$x[,1],pc$x[,2],pacientes_ID, pos=3, cex=.6)
```

Agrupamiento jerárquico

```
# Distancia y clustering jerárquico
d <- dist(t(expr_norm))
hc <- hclust(d, method = "average")

# Dibujo el dendrograma
hc$labels <- gsub("_.*", "", hc$labels)
plot(hc, main = "Dendrograma por muestra", cex = 0.7)
rect.hclust(hc, k = 2, border = c("red", "blue"))
```

Análisis de expresión diferencial

```
ttest <- function(x) {
  tt = t.test(x[1:47], x[48:77]) # Aplico un test t entre los 47 caquéxicos y los 30 controles
  return(c(
    t = tt$statistic,
    p.value = tt$p.value,
    log2FC = log2(mean(x[1:47]) / mean(x[48:77])) # Calculo el log2 Fold Change entre las medias de caq
  ))
}

ans <- apply(expr_norm, 1, ttest) # Aplico la función ttest() a cada fila de la matriz expr_norm (cada
```

```r
# Convertimos la matriz a data.frame
ttest_df <- as.data.frame(t(ans))

# Añadir nombres de los metabolitos
ttest_df$metabolitos <- rownames(ttest_df)
```

```r
EnhancedVolcano(ttest_df,
                lab = ttest_df$metabolitos,
                x = 'log2FC',
                y = 'p.value',
                pCutoff = 0.05,
                FCcutoff = 0.5,
                pointSize = 2,
                labSize = 3,
                title = 'Volcano Plot - Metabolitos')
```