Análisis de Datos Ómicos. PEC1.

María Jesús Acosta Silva

Tabla de contenidos.

1. Resumen.

En el presente informe se estudia el posible efecto de la metformina sobre el perfil metabolómico de ratones diabéticos. Utilizando la clase de objeto *SummarizedExperiment*, recogido en la librería *Bioconductor,* analizamos la distribución de abundancias de metabolitos, abordamos posibles transformaciones de datos y exploramos los perfiles metabólicos de los grupos de estudio mediante mapas de calor y análisis de componentes principales (PCA). Aunque muy preliminar, los resultados obtenidos apuntan a un efecto parcial de la metformina sobre el perfil metabolómico de ratones diabéticos que podrían justificar la realización de análisis más profundos.

2. Introducción y objetivos.

En el presente informe realizaremos un análisis de los datos obtenidos del **estudio ST000164** (ver aquí) extraído de ***Metabolomics Workbench***. El estudio realiza un análisis metabolómico no dirigido comparando muestras de médula ósea en ratones sanos (*wild type*) y ratones diabéticos en dos condiciones experimentales. La condición experimental consiste en la administración diaria vía intraperitoneal de 50 microlitros de PBS (condición control) o metformina (tratamiento de 200mg/kg peso) durante 14 días de duración. El estudio incluye, por tanto, cuatro grupos experimentales:

- Ratones control sin tratamiento (anotados como WTP).
- Ratones control tratados con metformina (WTM).
- Ratones diabéticos sin tratamiento (MKRP).
- Ratones diabéticos tratados con metformina (MKRM).

El análisis de los metabolitos presentes en las muestras de cada grupo se llevó a cabo mediante el espectómetro de masas de alta resolución Orbitrap, el cual permite la identificación precisa de metabolitos tanto conocidos como desconocidos. Se adquirieron datos en modo de ionización positiva (*ESI positive ion mode*) y negativa (*ESI negative ion mode*), lo que permitió detectar una mayor diversidad de compuestos, ya que ciertos metabolitos ionizan preferentemente en uno u otro modo.

El procesado y análisis de los datos obtenidos de este estudio se han realizado usando las librerías contenidas en *Bioconductor* mediante Rstudio. Para ello se ha creado un objeto *SummarizedExperiment*, el cual permite almacenar de manera estructurada datos de la matriz de abundancia, datos de anotación de muestras (metadatos) e información adicional de los metabolitos.

Los objetivos del estudio son:

- Descargar un estudio metabolómico de interés del *repositorio Metabolomics Workbench*.
- Crear un objeto de clase *SummarizedExperiment* a partir de estos datos.
- Realizar un análisis exploratorio de los datos.

3. Métodos.

3.1. Obtención de datos de Metabolic Workbench.

Accedemos a la página principal del estudio ST000164 de *Metabolic Workbench*. Descargamos el archivo .zip que contiene los datos del estudio. Tenemos:

- Un archivo .pdf con el protocolo del experimento.
- Dos archivos .excel que contienen los metadatos, uno de ellos ya procesado.
- Dos archivos .excel que contienen los resultados, uno de ellos ya procesado.

- Dos carpetas con datos crudos, que contienen los datos de iones positivos y negativos detectados por la técnica.

3.2. Creación del objeto *SummarizedExperiment*.

Como datos de partida, utilizaremos los archivos excel procesados de resultados y metadatos, los cuales hemos guardado en la carpeta del proyecto. Una consideración a tener en cuenta es que el programa por defecto utiliza la primera hoja del archivo excel. En el caso de los metadatos, el archivo incorpora varias hojas con información variada. En nuestro caso, nos interesa aquella que contenga información sobre los grupos experimentales (que aparezcan con la misma nomenclatura que la empleada en el archivo de resultados, ya que será lo que usemos de unión entre la parte del "*assay*" y la parte del "*metadata*" en nuestro objeto *SummarizedExperiment*), además de información sobre los tratamientos y el identificador de los sujetos de estudio, etc. Esa información la encontramos en la hoja "*Study Design*" del archivo procesado de metadatos.

```
library(SummarizedExperiment)
library(readxl)
library(tibble)

# Incorporamos los archivos excel de partida.
archivo_resultados <- "data/RESULTS_metformin_XinLi_Processed.xlsx"
archivo_metadata <- "data/METADATA_metformin_XinLi_Processed.xlsx"

# Resultados (prefiero "clonar" el archivo original por si tenemos que usarlo en adelante
o surgiera cualquier problema al manipular estos datos).
datos <- read_excel(archivo_resultados, .name_repair = "minimal")
head(datos)

## # A tibble: 6 × 48
##   `Metabolite Name`     METABOLITE_ID (will …¹ Units  MKRM1  MKRM1  MKRM1  MKRM2
##   <chr>                 <lgl>                 <chr>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 Metformin             NA                    peak… 2.63e6 2.82e6 3.01e6 2.72e6
## 2 Trigonellinamide      NA                    peak… 1.01e4 1.21e4 8.50e3 1.39e4
## 3 Ne-Methyl-L-lysine    NA                    peak… 2.71e3 3.57e3 3.95e3 5.06e3
## 4 Butyryl-L-carnitine   NA                    peak… 6.66e4 5.12e4 5.21e4 6.98e4
## 5 Palmitoyl-L-carnitine NA                    peak… 7.19e5 4.66e5 5.08e5 6.88e5
## 6 Succinic acid         NA                    peak… 7.94e3 5.59e3 5.09e3 1.50e3
## # ℹ abbreviated name: ¹`METABOLITE_ID (will be added)`
## # ℹ 41 more variables: MKRM2 <dbl>, MKRM2 <dbl>, MKRM3 <dbl>, MKRM3 <dbl>,
## #   MKRM3 <dbl>, MKRP1 <dbl>, MKRP1 <dbl>, MKRP1 <dbl>, MKRP2 <dbl>,
## #   MKRP2 <dbl>, MKRP2 <dbl>, MKRP3 <dbl>, MKRP3 <dbl>, MKRP3 <dbl>,
## #   MKRP4 <dbl>, MKRP4 <dbl>, MKRP4 <dbl>, WTM1 <dbl>, WTM1 <dbl>, WTM1 <dbl>,
## #   WTM2 <dbl>, WTM2 <dbl>, WTM2 <dbl>, WTM3 <dbl>, WTM3 <dbl>, WTM3 <dbl>,
## #   WTM4 <dbl>, WTM4 <dbl>, WTM4 <dbl>, WTP1 <dbl>, WTP1 <dbl>, WTP1 <dbl>, …

# Usamos el argumento "name_repair = minimal" para evitar que R añada sufijos como _1,
_2, etc. ya que las réplicas de cada muestra aparecen con nombres repetidos.

# Metadatos
metadata <- read_excel(archivo_metadata, sheet = "Study Design")
```

Necesitamos tres elementos imprescindibles para generar el objeto *SummarizedExperiment*:

- Los **datos de abundancia** ("***assay***"), que extraemos del *dataframe* que contiene los resultados ("datos"). Cada columna indica las abundancias de distintos metabolitos por muestra.

- Las **filas** ("***rowData***") con los **nombres de los metabolitos** encontrados, que también extraemos del dataframe "datos". Normalmente, se suele incluir además otra columna con una referencia (RefMet) de estos metabolitos, pero en nuestro caso, aunque aparece esa columna en el archivo, está vacía. Por tanto, no la incluimos.

- La **columna** ("***colData***") contenida en el **dataset de los metadatos** con los nombres anotados de cada sujeto experimental (MKRM1-3, MKRP1-4, WTM1-4, WTP1-4). Son cuatro muestras por grupo excepto en el caso de los ratones diabéticos tratados con metformina (MKRM), que son 3 muestras (MKRM1, MKRM2, MKRM3). Una particularidad a tener en cuenta es que se hicieron varias réplicas por muestra, lo cual hace que los nombres de las columnas de los datos de abundancia no coincidan con el de la columna de los metadatos.

Para hacer que coincidan exactamente, previamente vamos a efectuar el promedio de los datos de abundancia entre réplicas de la misma muestra para cada metabolito.

```
## 1. Nombre de los metabolitos ("rowData")
metabolite_names <- datos[, 1, drop=FALSE]

## 2. Datos de abundancia ("assay")

# Extraemos los datos de abundancia, eliminando las columnas con los nombres de los
metabolitos, la referencia y la unidad.
abundancia_raw <- as.matrix(datos[, -c(1, 2, 3)])

# Hacemos la media de las réplicas.
abundancia_promediada <- sapply(unique(colnames(abundancia_raw)), function(nombre) {
  cols <- which(colnames(abundancia_raw) == nombre)
  rowMeans(abundancia_raw[, cols, drop = FALSE])
})

# Convertimos el dataframe en una matriz
abundancia_matrix <- as.matrix(abundancia_promediada)

## 3. Metadatos de las muestras ("colData")

# Nos aseguramos de mantener solo las muestras para las que hay datos (tras promediar)
sample_info <- metadata[match(colnames(abundancia_matrix), metadata$Sample_Name), ]

# Verificamos que las muestras estén alineadas correctamente
stopifnot(all(sample_info$Sample_Name == colnames(abundancia_matrix)))

# Convertimos al formato requerido
sample_info <- DataFrame(sample_info)
rownames(sample_info) <- sample_info$Sample_Name

## Finalmente, construimos el objeto SummarizedExperiment

se <- SummarizedExperiment(
  assays = list(abundancia = abundancia_matrix),
  rowData = metabolite_names,
  colData = sample_info
)
```

Para acabar con este apartado, comentar las principales diferencias entre el objeto ***SummarizedExperiment* vs *ExpressionSet***. Aunque ambos son clases de objetos contenidos en *Bioconductor* y diseñados para almacenar datos ómicos y anotaciones, presentan varias diferencias. Los *ExpressionSet* se usan clásicamente para el análisis de microarrays, mientras que los objetos *SummarizedExperiment* se usan en RNAseq, metabolómica o proteómica. La principal diferencia reside en la cantidad de matrices que puede contener cada objeto. Mientras que los ExpressionSet solo soportan una matriz de datos, los objetos SummarizedExperiments pueden trabajar con múltiples matrices o "assays".

3.3. Análisis exploratorio.

En cuanto a los métodos estadísticos, se han considerado análisis univariantes (*box-plot*, histogramas) para estudiar la distribución general de los datos; así como análisis multivariante, mediante Análisis de Componentes Principales (PCA), para determinar si la variabilidad observada entre grupos puede ser explicada por los metabolitos considerados en el estudio.

4. Resultados.

4.1. Objeto *SummarizedExperiment*. Características básicas.

Nuestro primer resultado es el objeto *SummarizedExperiment*.

```
# Guardado de objeto SummarizedExperiment
saveRDS(se, file = "results/se.rds")
```

Realizaremos un primer análisis de la estructura del objeto como comprobación de que lo hemos creado correctamente.

```
# Dimensiones del objeto.
dim(se)

## [1] 228  15

# Resumen de los datos.
summary(assay(se))

##      MKRM1              MKRM2              MKRM3              MKRP1
##  Min.   :       0   Min.   :       0   Min.   :      48   Min.   :       0
##  1st Qu.:    3894   1st Qu.:    3976   1st Qu.:    4004   1st Qu.:    3263
##  Median :   10236   Median :   11854   Median :   15574   Median :    8883
##  Mean   :  372207   Mean   :  403002   Mean   :  563341   Mean   :  263124
##  3rd Qu.:   55144   3rd Qu.:   70900   3rd Qu.:   81462   3rd Qu.:   44868
##  Max.   :17635988   Max.   :18908069   Max.   :23996605   Max.   :14118003
##      MKRP2              MKRP3              MKRP4               WTM1
##  Min.   :      45   Min.   :      52   Min.   :      60   Min.   :       0
##  1st Qu.:    4365   1st Qu.:    4837   1st Qu.:    5222   1st Qu.:    3695
##  Median :   15838   Median :   15428   Median :   16969   Median :   11241
##  Mean   :  556135   Mean   :  719837   Mean   : 1074745   Mean   :  433251
##  3rd Qu.:   86228   3rd Qu.:  123216   3rd Qu.:  138152   3rd Qu.:   61641
##  Max.   :26023087   Max.   :33525201   Max.   :48123943   Max.   :19946761
##       WTM2               WTM3               WTM4               WTP1
##  Min.   :       0   Min.   :      50   Min.   :      44   Min.   :       0
##  1st Qu.:    4344   1st Qu.:    4152   1st Qu.:    4795   1st Qu.:    4932
##  Median :   17578   Median :   13140   Median :   14195   Median :   16580
##  Mean   :  763333   Mean   :  432421   Mean   :  657983   Mean   :  542598
##  3rd Qu.:  140347   3rd Qu.:   70637   3rd Qu.:  101927   3rd Qu.:  114126
```

```
##  Max.   :33164559   Max.   :21037646   Max.   :34208478   Max.   :23480105
##       WTP2              WTP3              WTP4
##  Min.   :      57   Min.   :       0   Min.   :       0
##  1st Qu.:    4739   1st Qu.:    4353   1st Qu.:    6138
##  Median :   15135   Median :   13624   Median :   19914
##  Mean   :  639700   Mean   :  481817   Mean   : 1075077
##  3rd Qu.:  127410   3rd Qu.:   96514   3rd Qu.:  170177
##  Max.   :32384441   Max.   :22302982   Max.   :47048878

# Vista rápida de los nombres de metabolitos y muestras.
head(rownames(rowData(se)))

## NULL

colnames(se)

##  [1] "MKRM1" "MKRM2" "MKRM3" "MKRP1" "MKRP2" "MKRP3" "MKRP4" "WTM1"  "WTM2"
## [10] "WTM3"  "WTM4"  "WTP1"  "WTP2"  "WTP3"  "WTP4"
```

4.2. Valores faltantes (NA).

Comprobemos la ausencia o no de datos en nuestra matriz de abundancias.

```
# Valores faltantes por metabolito (fila)
row_na_count <- rowSums(is.na(assay(se)))

# Valores faltantes por muestra (columna)
col_na_count <- colSums(is.na(assay(se)))
```

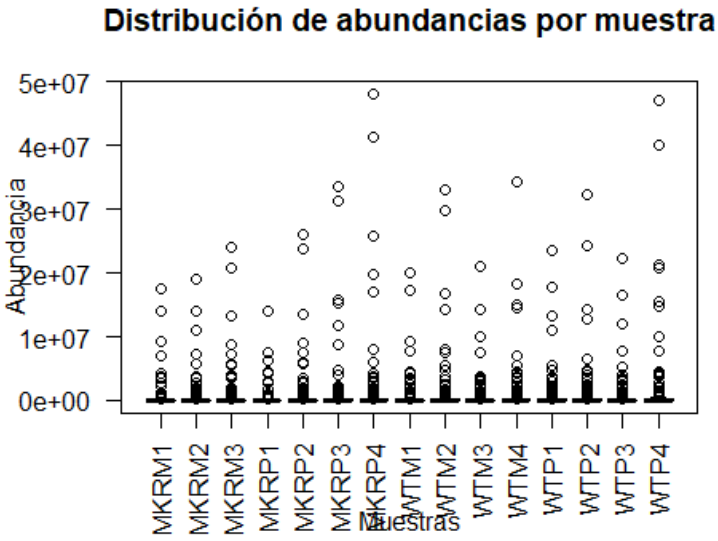
Ya que no hay ningún NA, cabe plantearse la presencia de ceros en los datos, ya que, en este caso concreto, **un valor 0 puede no indicar la ausencia real de un metabolito**, sino deberse a que el metabolito en cuestión se encuentre por debajo del umbral de detección de la técnica. En ese caso, habría que plantearse la imputación de los ceros como valores NA.

```
# Presencia de valor cero en matriz de abundancia.
sum(assay(se) == 0)

## [1] 10

# Valor cero por muestra (columna)
col_zero_count <- colSums(assay(se) == 0)
col_zero_count

## MKRM1 MKRM2 MKRM3 MKRP1 MKRP2 MKRP3 MKRP4  WTM1  WTM2  WTM3  WTM4  WTP1  WTP2
##     1     1     0     3     0     0     0     1     1     0     0     1     0
##  WTP3  WTP4
##     1     1
```

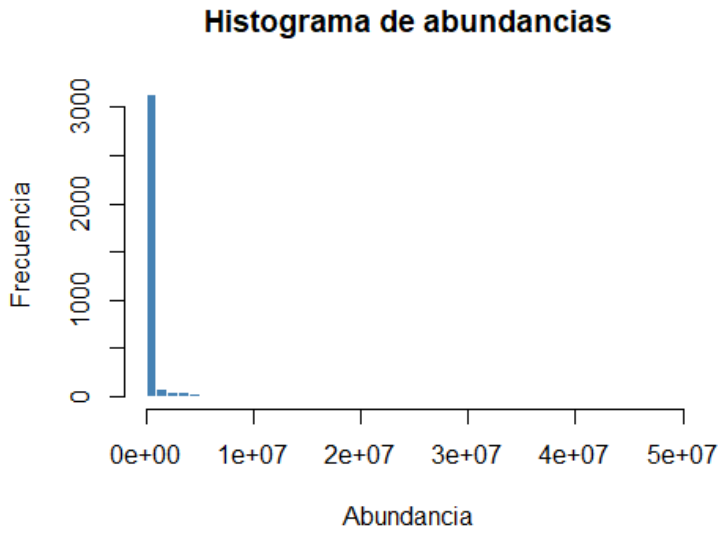
4.3. Distribución de abundancias.

Veamos primero la distribución abundancias por muestra.

```
## Box-plot de abundancia por muestra.
boxplot(assay(se),
        main = "Distribución de abundancias por muestra",
        xlab = "Muestras",
        ylab = "Abundancia",
        las = 2,          # para girar etiquetas del eje X
        col = "steelblue")
```

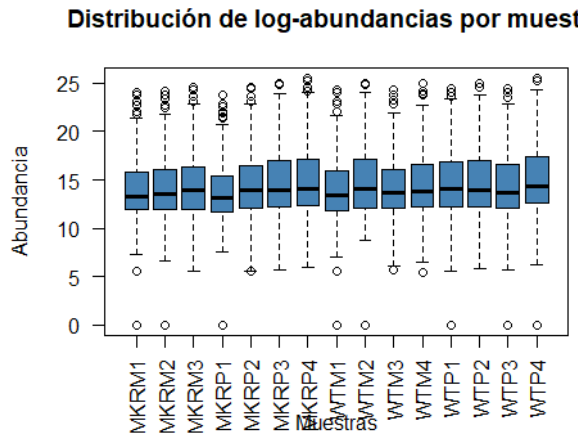



```r
## Histograma de abundancias globales
hist(as.vector(assay(se)), # Agrupamos todas las abundancias en un solo vector para
mostrar la distribución global.
     breaks = 50,
     main = "Histograma de abundancias",
     xlab = "Abundancia",
     ylab = "Frecuencia",
     col = "steelblue",
     border = "white")
```

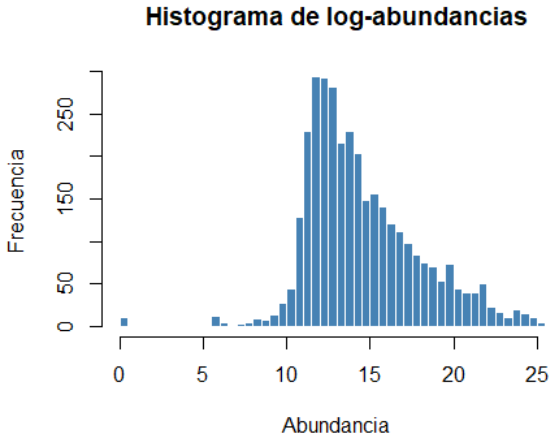


El *box-plot* **no parece mostrar asimetrías evidentes**, aunque la escala podría ser un problema, ya que tenemos diferentes metabolitos que podrían variar enormemente entre ellos. Por otro lado, el histograma de abundancias globales nos muestra una **asimetría positiva**, siendo recomendable la aplicación de una transformación logarítmica.

```r
## Transformación logarítmica de la matriz de abundancia.
log_assay <- log2(assay(se)+1) # Sumamos 1 para evitar problemas con los valoers cero.

## Box-plot de log-abundancia por muestra.
boxplot(log_assay,
        main = "Distribución de log-abundancias por muestra",
        xlab = "Muestras",
        ylab = "Abundancia",
        las = 2,          # para girar etiquetas del eje X
        col = "steelblue")
```




```r
## Histograma de log-abundancias globales.
hist(as.vector(log_assay), # Agrupamos todas las abundancias en un solo vector para
mostrar la distribución global.
     breaks = 50,
     main = "Histograma de log-abundancias",
     xlab = "Abundancia",
     ylab = "Frecuencia",
     col = "steelblue",
     border = "white")
```

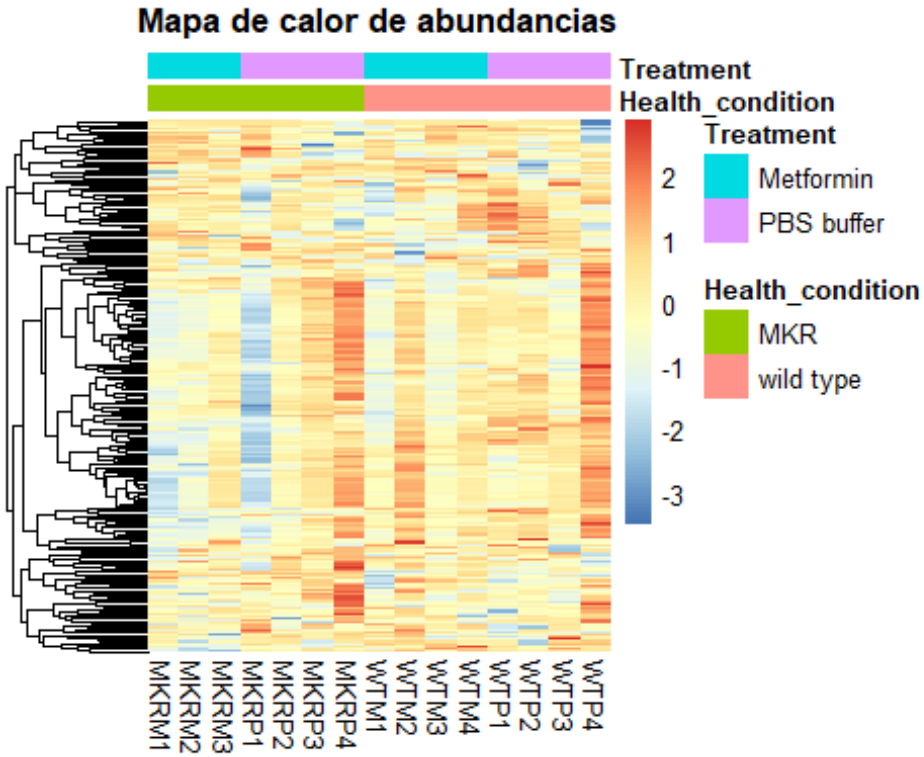


Comprobamos que, efectivamente, la transformación logarítmica mejora la distribución de los datos de abundancia.

4.4. Mapa de color.

A continuación, exploramos posibles patrones de abundancia por condición y grupo experimental usando mapas de calor. Dado que cada metabolito presenta diferentes rangos de abundancia y nuestro interés está en comparar patrones más que en las cantidades absolutas de metabolitos, el primer paso en este caso es realizar la normalización de los datos de cada metabolito (transformación *z-score* por filas, con media = 0, desviación típica = 1).

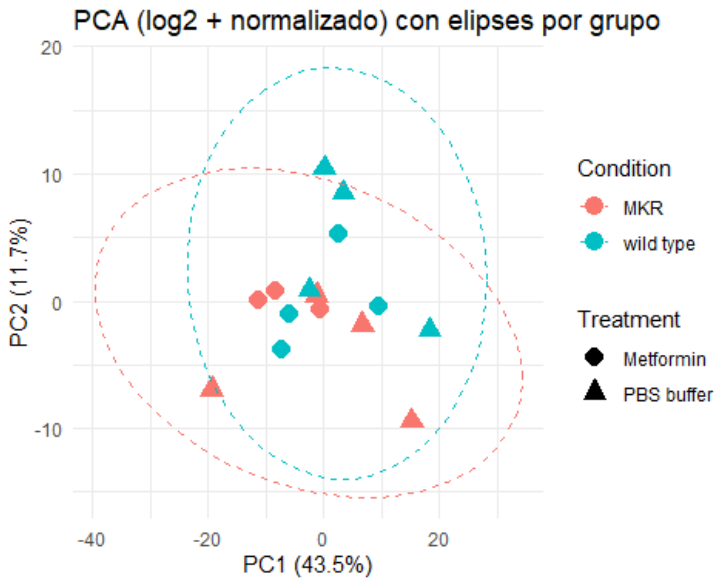
```
library(pheatmap)

# Normalización por filas de matriz de log-abundancias.
z_scaled <- t(scale(t(log_assay))) # Realizamos una doble trasposición ya que la función
de escalado por defecto normaliza columnas.

# Extraemos los metadatos de las columnas.
heat_col<- as.data.frame(colData(se)[, c("Health_condition", "Treatment")])

# Ordenamos las columnas por grupo experimental y tratamiento.
orden <- order(heat_col$Health_condition, heat_col$Treatment)
z_scaled_ordenado <- z_scaled[, orden]
heat_col_ordenado <- heat_col[orden, ]

# Mapa de calor por grupo experimental y tratamiento
pheatmap(z_scaled_ordenado,
         annotation_col = heat_col_ordenado,
         show_rownames = FALSE,
         cluster_cols = FALSE,
         main = "Mapa de calor de abundancias")
```



El mapa de calor revela algunas posibles tendencias. Por ejemplo, parece que hay una disminución en buena parte de los metabolitos testados en los ratones diabéticos tratados con metformina (MKRM) con respecto al resto de grupos. Lo más evidente es la **heterogeneidad interna dentro de los grupos** sin tratamiento (tanto MKRP como WTP), ya que en ambos grupos encontramos muestras con valores de abundancia muy por encima o por debajo de la tendencia del grupo. En general, da la impresión de que **el tratamiento con metformina reduce la producción de ciertos metabolitos tanto en el grupo control (WTM) como en los ratones diabéticos (MKRM)**.

4.2. Análisis de componentes principales (PCA).

```
# Trasposición de la matriz en escala logarítmica y normalizada.
z_scaled_t <- t(z_scaled)

# PCA
pca <- prcomp(z_scaled_t)

# Extraemos los metadatos.
coldata <- as.data.frame(colData(se))

# Creamos un nuevo dataframe que contenga los datos del PCA.
pca_df <- as.data.frame(pca$x)
pca_df$Condition <- coldata$Health_condition
pca_df$Treatment <- coldata$Treatment

# Visualización de muestras y elipses.
library(ggplot2)

ggplot(pca_df, aes(x = PC1, y = PC2, color = Condition, shape = Treatment)) +
  geom_point(size = 4) +
  stat_ellipse(aes(group = Condition), type = "norm", linetype = 2) +
  labs(
    title = "PCA (log2 + normalizado) con elipses por grupo",
    x = paste0("PC1 (", round(summary(pca)$importance[2, 1] * 100, 1), "%)"),
    y = paste0("PC2 (", round(summary(pca)$importance[2, 2] * 100, 1), "%)")
  ) +
  theme_minimal()
```



El Análisis de Componentes Principales (PCA) muestra cómo se distribuyen las muestras en función de su perfil metabolómico completo. Cada punto en el gráfico representa una muestra, cuyas coordenadas están determinadas por combinaciones lineales de los que metabolitos que explican la mayor parte de la variabilidad observada en los datos. En nuestro caso, ambas componentes **explican en torno a un 55% de toda la variabilidad observada**. Las elipses de confianza son una estimación visual que nos indican en este caso cómo se agrupan las muestras por su estado de salud (controles vs diabéticos). En este caso, aunque con un solapamiento considerable, la orientación de las elipses sugiere cierta separación entre los ratones diabéticos (MKR) y control (wild type o WT). La dispersión en la segunda componente indica variabilidad intragrupo, posiblemente derivada de los tratamientos (metformina vs PBS).

5. Discusión.

El análisis exploratorio de los datos transformados logarítmicamente y normalizados nos ha permitido obtener una primera visión global de los perfiles metabolómicos de las muestras según grupo experimental (ratones diabéticos vs. controles) y tratamiento (metformina vs PBS). Tanto el mapa de calor como el análisis de componentes principales (PCA) apuntan hacia una separación parcial de los grupos experimentales según su perfil metabólico. En conjunto, ambos métodos sugieren que el tratamiento con metformina tiene cierto impacto en el perfil metabolómico, especialmente en ratones diabéticos, aunque se requiere de un análisis estadístico más profundo para confirmar la significancia de estos patrones.

También cabe destacar la variabilidad interna dentro de algunos grupos, lo que plantea la posibilidad de que el promedio aplicado a las réplicas haya enmascarado parte de esa variación biológica. Además, la presencia de ceros en los datos podría justificar, en etapas posteriores, la aplicación de técnicas de imputación para mejorar la interpretación de los resultados.

De proseguir con este estudio, sería interesante identificar los metabolitos que exhiben mayor variabilidad entre grupos y/o condiciones experimentales; así como estudiar posibles cambios en la abundancia de metabolitos estrechamente relacionados con el metabolismo de la glucosa y los lípidos, los cuales, teóricamente, deberían ser los principales afectados la condición de diabetes y el tratamiento con metformina.

6. Conclusión.

El análisis exploratorio preliminar sugiere un efecto parcial de la metformina sobre el perfil metabólico de ratones diabéticos. No obstante, la variabilidad interna observada en algunos grupos y la posible necesidad de un mayor procesado de los datos, requieren de un análisis estadístico más profundo para confirmar la significancia de los patrones observados.

7. Referencia.

Consultar el siguiente repositorio en GitHub:

https://github.com/MJ-Acosta-Silva/Acosta_Silva_Mar-a_Jes-s_PEC1.git