ANÁLISIS DE DATOS ÓMICOS - PEC 1

Bruno Martin Farcic Melo

2024-11-04

Contents

1 INTRODUCCIÓN

Este estudio recopila datos de metabolitos obtenidos de muestras intestinales de seis pacientes a los que se les ha hecho un transplante. Este estudio pretende proporcionar una visión de los cambios bioquímicos que se llevan a cabo dentro del organismo tras pasar por este procedimiento.

2 OBJETIVOS

Mediante este análisis exploratorio, se busca identificar patrones y tendencias en los datos, con un enfoque particular en las posibles alteraciones antes y después del trasplante. Este análisis preliminar no solo pretende caracterizar el perfil metabolómico de las muestras, sino también establecer una base para estudios futuros que puedan profundizar en el impacto fisiológico del trasplante en los pacientes.

3 MATERIALES Y MÉTODOS

Los datos que se utilizarán en este estudio se encuentran disponibles en la página web de "Metabolomic Workbench". Se utilizarán estos datos para obtener un Summarized Experiment que contiene una matriz con los datos e información extra de los metabolitos de estudio.

Dentro de los metadatos podemos observar información adicional sobre el proyecto, como el nombre o la universidad donde se obtuvieron, así como de los métodos de análisis y maquinaria utilizada. El centro responsable de la preparación y análisis de estas muestras fue el laboratorio Feihn de la Universidad de California, Davis en el año 2013. Los métodos utilizados para el análisis fueron cromatografía de gases junto con espectrometría de masas.

3.1 Preparación del "SummarizedExperiment"

El archivo que contiene los datos es un .txt, que si lo abrimos y lo observamos nos delimita los metadatos, y nos los separa de los datos. Si nos fijamos las primeras 70 líneas aproximadamente son de metadatos, y a partir de "MS_METABOLITE_DATA_START" comienza la tabla de datos hasta "MS_METABOLITE_DATA_END", así como la información de los metabolitos, que se encuentra delimitada.

La construcción de la matriz tendrá una estructura donde cada fila representará un metabolito diferente, y cada columna las muestras (pre y post transplante). Los metadatos de las muestras como la condición se encuentra en la variable colData, mientras que el de los metabolitos en rowData.

Después de la preparación de las variables acabamos obteniendo el "SummarizedExperiment".

```
## class: SummarizedExperiment
## dim: 142 12
## metadata(0):
## assays(1): metabolomica
## rownames(142): 1-monoolein 1-monostearin ... xanthine xylose
## rowData names(9): metabolite_name moverz_quant ... other_id
##   other_id_type
## colnames(12): LabF_684508 LabF_684512 ... LabF_684499 LabF_684503
## colData names(1): Condition
```

4 ANÁLISIS

Lo primero que vamos a hacer en nuestro análisis exploratorio es tener una visión general de los datos con los que vamos a trabajar.

Table 1: Matriz de dato

	LabF_684508	LabF_684512	LabF_684516	LabF_684520	LabF_684524	LabF_68452
1-monoolein	6047.0000	2902.0000	1452.0000	3428.0000	2985.0000	16334.0000
1-monostearin	9771.0000	6521.0000	1302.0000	2781.0000	5789.0000	4338.0000
2-hydroxybutanoic acid	13238.0000	29774.0000	4134.0000	4419.0000	13334.0000	2115.0000
2-hydroxyglutaric acid	7160.0000	11501.0000	3202.0000	17238.0000	20376.0000	1109.0000
2-ketoisocaproic acid	812.0000	2011.0000	738.0000	2550.0000	871.0000	628.0000
2-monopalmitin	1511.0000	622.0000	883.0000	796.0000	623.0000	5716.0000

Table 2: Metadatos de las muestras (colData)

	Condition
LabF_684508	Transplantation:After transplantation
LabF_684512	Transplantation:After transplantation
LabF_684516	Transplantation:After transplantation
LabF_684520	Transplantation:After transplantation
LabF_684524	Transplantation:After transplantation
LabF_684528	Transplantation:After transplantation
LabF_684483	Transplantation:Before transplantation
LabF_684487	Transplantation:Before transplantation
LabF_684491	Transplantation:Before transplantation
LabF_684495	Transplantation:Before transplantation
LabF_684499	Transplantation:Before transplantation
LabF_684503	Transplantation:Before transplantation

Table 3: Inicio de los metadatos de los metabolitos (rowData)

	metabolite_name	moverz_quant	ri	ri_type	pubchem_id	inchi_key	kegg_
1-monoolein	1-monoolein	129	952993	Fiehn	5283468	NA	
1-monostearin	1-monostearin	399	959625	Fiehn	107036	NA	D019
2-hydroxybutanoic acid	2-hydroxybutanoic acid	131	258175	Fiehn	11266	NA	C059
2-hydroxyglutaric acid	2-hydroxyglutaric acid	129	506359	Fiehn	43	NA	C026
2-ketoisocaproic acid	2-ketoisocaproic acid	200	310629	Fiehn	70	NA	C002
2-monopalmitin	2-monopalmitin	218	889972	Fiehn	123409	NA	

Como podemos intuir al realizar un estudio sobre el metaboloma, varios metabolitos van a tener una expresión mucho más alta que otros, por lo que la variabilidad que obtendremos dentro de cada muestra será enorme. Para intentar palear esto, vamos a realizar unos ajustes de normalización y escalado.

Normalización por suma: Con este ajuste pretendemos llevar todos los valores de los metabolitos a la media de la suma, ya que asumimos que la suma de todos los metabolitos debería mantenerse constante entre muestras independientemente de las condiciones, así podremos encontrar esa variabilidad entre pre y post transplante que buscamos.

Escalado Pareto: Al hablar de metaboloma encontramos un problema, y es que los niveles de estos pueden variar entre ellos por varios órdenes de magnitud, por lo que aplicando el escalado de Pareto, centramos cada metabolito en su media y lo dividimos por la raiz cuadrada de su desviación estándar, para reducir este sesgo que se produce por los metabolitos que nos dan una intensidad mayor.

4.1 Densidad

Un gráfico de densidades nos permite ver la distribución de las intensidades de los metabolitos y las muestras. En nuestro caso podemos observar que se centran todos en valores bajos, y que las curvas son similares por lo que se nos invita a pensar que la variabilidad de las muestras es consistente, que si alguna se desviara veríamos un pico desplazado. Podemos decir, además, que el proceso de normalización ha sido efectivo ya que igualmente las curvas se ven de forma similar.

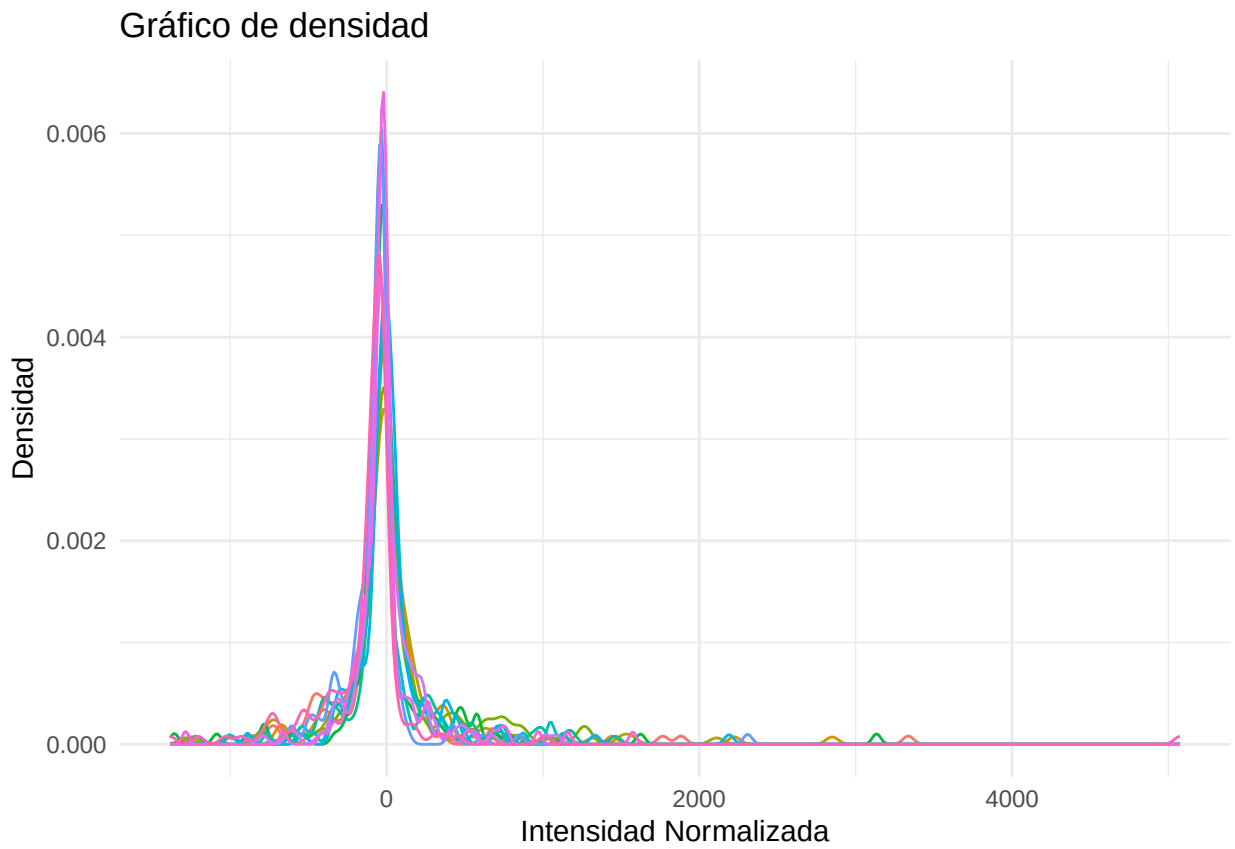

Gráfico 1: Gráfico de densidad de las intensidades después de la normalización y escalado.

4.2 Boxplot

Un boxplot tiene la misma finalidad que el gráfico de densidades, ver si hay algún metabolito que tenga una distribución alterada. Si tuviéramos una variabilidad elevada en alguno de estos, la caja torna un aspecto alto. Además, los puntos rojos nos representan valores atípicos, también conocidos como "outliers", lo cual de nuevo nos marca esta variabilidad.

Un posible análisis interesante sería buscar aquellos metabolitos que tengan una variación elevada entre los grupos de pre y post transplante, para posteriormente hacer un estudio de la significancia.

En nuestro gráfico podemos ver condensado todos los metabolitos y un gran número de muestras que tienen valores atípicos, y gracias a la normalización y escala podemos comparar la variabildiad entre ellos.

El boxplot de los metabolitos

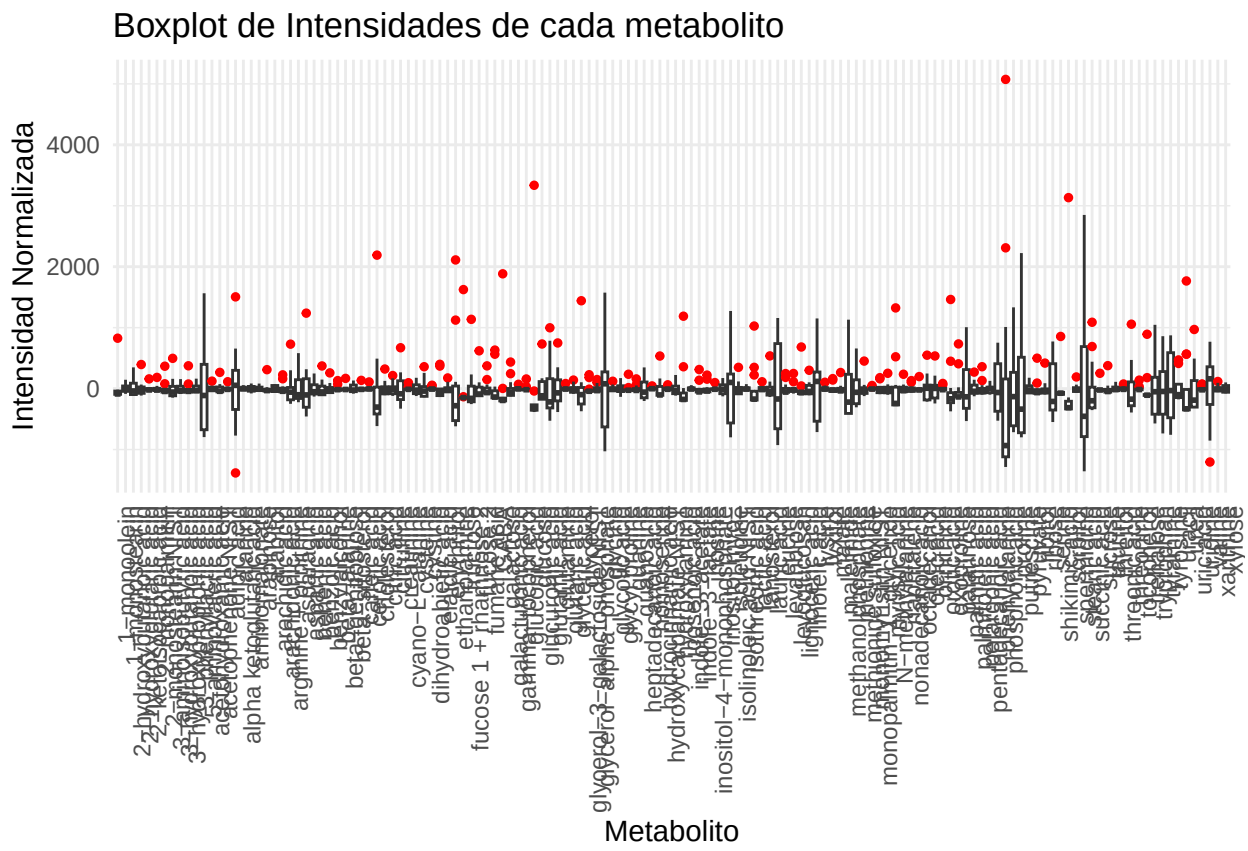

Gráfico 2: Diagrama de cajas de las intensidades de los metabolitos después de la normalización.

4.3 PCA

Un gráfico de PCA (Principal Component Analisis) nos permite evaluar la variabilidad proyectandola en un eje con los componentes principales. Este gráfico nos ayuda a ver si las muestras muestran algún tipo de agrupación según las condiciones de pre y post transplante.

Para elaborarlo, primero necesitamos conseguir los valores de los componentes, y vemos con el primero cubrimos cerca del 30% de la variabilidad y con el segundo alrededor del 17%

```
## Importance of components:
##                             PC1       PC2       PC3       PC4       PC5
## Standard deviation     2452.7106 1892.3348 1785.0804 1451.0104 1.227e+03
## Proportion of Variance    0.2972    0.1769    0.1574    0.1040 7.438e-02
## Cumulative Proportion     0.2972    0.4742    0.6316    0.7356 8.100e-01
##                             PC6       PC7       PC8       PC9      PC10
## Standard deviation     1.123e+03 980.40340 885.29670 691.92343 508.8906
## Proportion of Variance 6.235e-02   0.04749   0.03872   0.02365   0.0128
## Cumulative Proportion  8.723e-01   0.91984   0.95856   0.98222   0.9950
##                            PC11    PC12
## Standard deviation     317.75689 0.00388
## Proportion of Variance   0.00499 0.00000
## Cumulative Proportion    1.00000 1.00000
```

Tabla 4: Resultados del PCA

En el gráfico podemos observar que claramente existe una separación de los grupos pre y post transplante. Sería interesante ver qué metabolitos contribuyen al PC1 sobretodo ya que es el que mayor variabilidad presenta.

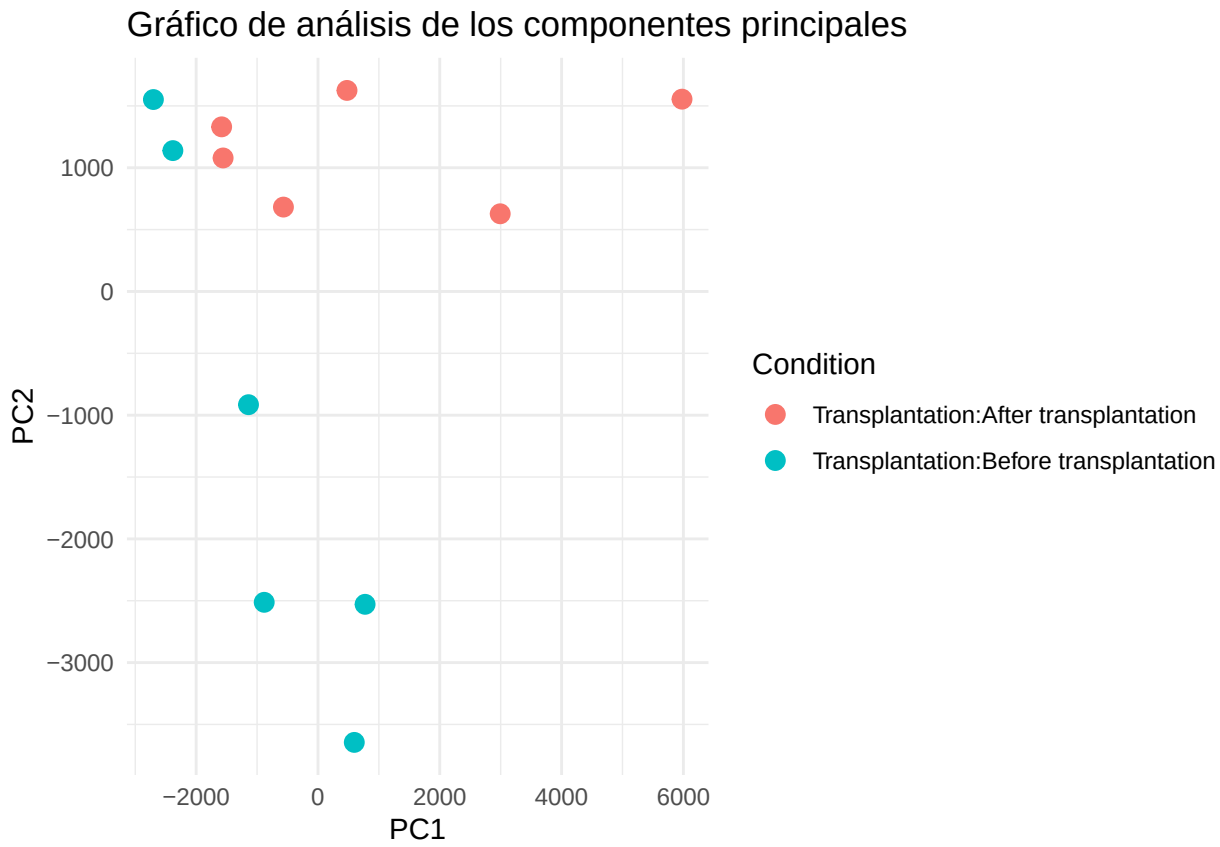

Gráfico 3: Gráfico PCA de las muestras.

4.4 Heatmap

El heatmap nos permite ver cómo varían los metabolitos entre las muestras y nos señala las intensidades de cada uno mediante el color, a más cálido más intenso.

Las muestras también se agrupan mediante un dendograma, que nos muestras las relaciones de similitud entre las muestras. Si nos fijamos, se ven los dos grupos separados exceptuando dos muestras que se ver alternadas, indicando posiblemente mayor variabilidad en esas muestras.

También podemos extraer de nuestro gráfico que hay diferencias en la expresión de varios metabolitos teniendo en cuenta el pre y post transplante, e identificar aquellos más relevantes podría ser un buen punto de inicio para siguientes estudios.

```
# Volvemos a llamar los datos y los manejamos para que sean más entendibles
datos_heatmap <- assay(sumex, "metabolomica")

anotaciones <- as.data.frame(colData(sumex)$Condition)
rownames(anotaciones) <- colnames(datos_heatmap)
colnames(anotaciones) <- "Condición"
```

```
# Generamos el heatmap
pheatmap(datos_heatmap, annotation_col = anotaciones, scale = "row", clustering_distance_rows = "euclid
```

```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

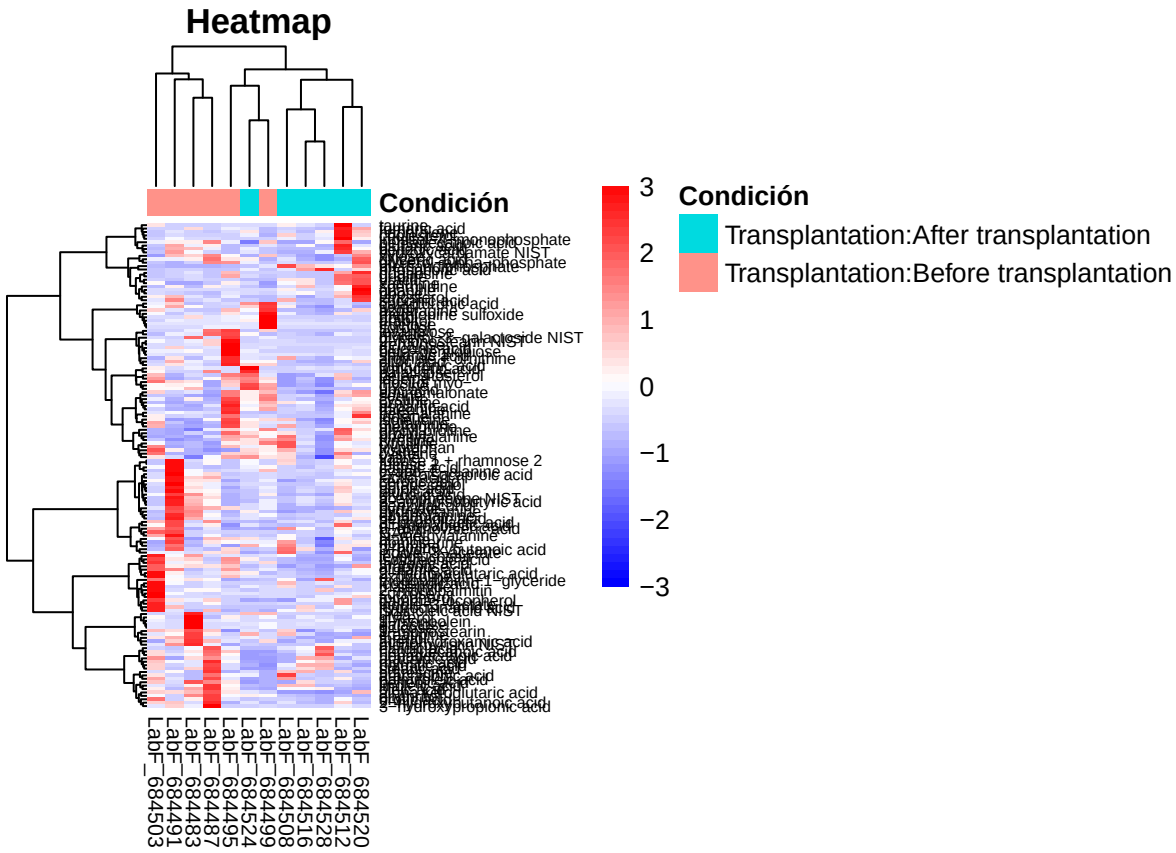

Gráfico 4: Mapa de calor de las muestras y metabolitos.

5 CONCLUSIONES

Como explicamos en los objetivos del trabajo, el proyecto pretendía hacer un análisis exploratorio preelimnar a cualquier análsis estadístico. La normalización y escalado de los datos era algo esencial para reducir todo lo que pudiera el sesgo que la diferencia de expresión entre metabolitos nos ocasionara.

Gracias a esto y los gráficos obtenido, podemos decir por la distribución que la normalización fue un éxito y al gráfico de cajas que hay una alta presencia de "outliers" en los metabolitos.

Mediante el PCA y heatmap pudimos observar diferencias de expresión tanto entre muestras como entre metabolitos, y que hay una agrupación de estas claramente marcada.

Los resultados nos sugieren que hay diferencias de expresión en el metaboloma entre muestras de pre y post transplante, indicando que posteriores estudios podrían esclarecer más información sobre cuáles son los principales metabolitos que varían.

Para terminar, y como futuros estudios, se podría realizar un análisis de Fold Change para observar aquellos metabolitos con mayor variabildad entre ambos grupos de transplante, así como un t-student para determinar si la diferencia entre ambos grupos es significativa.

6 APÉNDICE

Todos los archivos acerca de este trabajo, datos en formato binario, archivo markdown, y datos de estudio se encuentran en el repositorio de github: https://github.com/Bruno-FarMe/PEC-1-Analisis-de-Datos-Omicos.git

```
# Instalación de paquetes y carga de librerías
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager", repos = "https://cran.rstudio.com")

BiocManager::install()
BiocManager::install("SummarizedExperiment")


install.packages("SummarizedExperiment", repos = "https://cran.rstudio.com")
install.packages("readr", repos = "https://cran.rstudio.com")
install.packages("dplyr", repos = "https://cran.rstudio.com")
install.packages("DT", repos = "https://cran.rstudio.com")
install.packages("kableExtra", repos = "https://cran.rstudio.com")

library(SummarizedExperiment)
library(readr)
library(DT)
library(kableExtra)
library(ggplot2)
library(tidyr)

# Cargamos el archivo con los datos que usaremos en el análisis

archivo_datos <- "C:/Users/Bruno/Desktop/analisis de datos omicos/ST000002_AN000002.txt"

# Leeremos el archivo como lineas para facilitar su manejo a la hora de crear el SummarizedExperiment

datos_lineas <- readLines(archivo_datos)

# Comenzamos la extracción de datos. Previamente hemos abierto el .txt y vimos que los datos se contien

metabo_data_start <- which(datos_lineas == "MS_METABOLITE_DATA_START")+1
metabo_data_end <- which(datos_lineas == "MS_METABOLITE_DATA_END")-1
metabolito_start <- which(datos_lineas == "METABOLITES_START")+1
metabolito_end <- which(datos_lineas == "METABOLITES_END")-1

# Ahora pasaremos los datos de los metabolitos a una tabla

metabo_data_tabla <- read.table(text=datos_lineas[metabo_data_start:metabo_data_end], header=TRUE, sep=
row.names(metabo_data_tabla) <- metabo_data_tabla[,1]
metabo_data_tabla <- metabo_data_tabla[,-1]

# Pasamos también la información de los metabolitos

metabo_info_tabla <- read.table(text=datos_lineas[metabolito_start:metabolito_end], header=TRUE, sep="\

# Obtenemos los factores experimentales y la quitamos de la tabla original
```

```r
factores <- metabo_data_tabla[1,]
metabo_data_tabla <- metabo_data_tabla[-1,]

factores <- t(factores)
colnames(factores) <- "Condition"

# Generamos el contenedor SummarizedExperiment

sumex <- SummarizedExperiment(assays=list(metabolomica = as.matrix(metabo_data_tabla)), colData=factores

# Observamos que se haya generado bien

sumex

# Head de la matriz de datos
kable(as.data.frame(head(assay(sumex, "metabolomica"))), caption = "Matriz de datos") %>% kable_styling

# Tabla de colData
kable(as.data.frame(colData(sumex)), caption = "Metadatos de las muestras (colData)") %>% kable_styling

# Tabla de rowData
kable(as.data.frame(head(rowData(sumex))), caption = "Metadatos de los metabolitos (rowData)") %>% kabl

# Extraemos la matriz de datos y nos aseguramos de que todos los valores sean numéricos.

datos <- assay(sumex, "metabolomica")
datos <- apply(datos, 2, function(x) as.numeric(as.character(x)))

# Eliminamos las filas que contengan valores nulos después de buscar los numéricos
datos_finales <- na.omit(datos)

# Calculamos la suma total de cada muestra
suma_muestras <- colSums(datos_finales)

# Calculamos el promedio de las sumas de las muestras
promedio_suma <- mean(suma_muestras)

# Normalización por suma
datos_normalizados <- sweep(datos, 2, suma_muestras, FUN = "/") * promedio_suma

# Ahora vamos con el escalado, primero buscamos la media de cada metabolito
medias_metabolitos <- rowMeans(datos_normalizados)
datos_centrados <- sweep(datos, 1, medias_metabolitos, FUN = "-")

# Realizamos el escalado Pareto
sd_metabolitos <- apply(datos_centrados, 1, sd)
datos_escalados <- sweep(datos_centrados, 1, sqrt(sd_metabolitos), FUN = "/")

# Asignar los datos normalizados y escalados al SummarizedExperiment
assay(sumex, "metabolomica", withDimnames = FALSE) <- datos_escalados

# Verificamos y ajustamos los nombres de las filas y columnas si fuera necesario ya que eliminamos los
rownames(datos_escalados) <- rownames(assay(sumex, "metabolomica"))
```

```r
colnames(datos_escalados) <- colnames(assay(sumex, "metabolomica"))

# Para finalizar se asignan estos nuevos datos normalizados y escalados al summarizedExperiment
assay(sumex, "metabolomica") <- datos_escalados


# Convertimos la matriz de datos a un formato largo para ggplot2
datos_largos <- as.data.frame(assay(sumex, "metabolomica"))
datos_largos$Metabolito <- rownames(datos_largos)
datos_largos <- pivot_longer(datos_largos, -Metabolito, names_to = "Muestra", values_to = "Intensidad")

# Gráfico de densidad
ggplot(datos_largos, aes(x = Intensidad, color = Muestra)) + geom_density(alpha = 0.3) + labs(title = "


# Boxplot
ggplot(datos_largos, aes(x = Metabolito, y = Intensidad)) + geom_boxplot(outlier.color = "red", outlier

# Para realizar el PCA primero extraemos los datos y transponemos la matriz
datos_pca <- t(assay(sumex, "metabolomica"))

# Realizamos el PCA
pca_result <- prcomp(datos_pca, center = FALSE, scale. = FALSE)

summary(pca_result)

# Extraemos los scores de los dos primeros componentes principales
pca_data <- as.data.frame(pca_result$x)
pca_data$Condition <- colData(sumex)$Condition
pca_data$Muestra <- rownames(pca_data)

# Generamos el gráfico
ggplot(pca_data, aes(x = PC1, y = PC2, color = Condition, label = Muestra)) +
  geom_point(size = 3) +
  labs(title = "Gráfico de análisis de los componentes principales",
       x = "PC1", y = "PC2") +
  theme_minimal()

# Volvemos a llamar los datos y los manejamos para que sean más entendibles
datos_heatmap <- assay(sumex, "metabolomica")

anotaciones <- as.data.frame(colData(sumex)$Condition)
rownames(anotaciones) <- colnames(datos_heatmap)
colnames(anotaciones) <- "Condición"

# Generamos el heatmap
pheatmap(datos_heatmap, annotation_col = anotaciones, scale = "row", clustering_distance_rows = "euclid
```