

PEC 1 - Análisis de Datos Ómicos

María José Beltrán

2025-03-31

ABSTRACT

Este estudio analiza un conjunto de datos clínicos y de metabolitos de pacientes sometidos a dos tipos de cirugía: bypass y tubular. Se realizó un análisis descriptivo de las variables clínicas (edad, género, tipo de cirugía y grupo de pertenencia), encontrando una muestra predominantemente femenina (69.2%) y con una edad media de 40.8 años. El análisis mostró que la mayoría de los participantes fueron sometidos a bypass (66.7%), y que no hubo valores faltantes en las variables clínicas. Sin embargo, los datos de metabolitos presentaron un 12.6% de valores faltantes, que fueron imputados con el valor cero. El análisis de Componentes Principales (PCA) reveló que el tipo de cirugía fue el principal factor que explicó la variabilidad en los metabolitos, mientras que el grupo de pertenencia no mostró una influencia significativa. Adicionalmente, el análisis de similitud también indicó que las muestras se agrupan principalmente según el tipo de cirugía. Finalmente, se propone una reorganización de los datos de expresión metabólica, sugiriendo un enfoque donde el tiempo se maneje como una variable independiente, lo que facilitaría la observación de cambios temporales en las concentraciones postquirúrgicas.

OBJETIVOS

El objetivo principal es realizar un análisis exploratorio de datos metabólicos obtenidos a través de un repositorio público. Se pretende extraer algún patrón acerca de la distribución de los mismos que tenga explicación biológica.

MATERIALES Y MÉTODOS

Dataset

Se ha seleccionado, a partir del repositorio nutrimetabolomics/metaboData el dataset “2018-MetabotypingPaper”. En concreto, se ha elegido este dataset debido a que dispone de la matriz de datos y los metadatos, que es lo que se necesitará para hacer la PEC, en un formato cómodo como csv.

```
metadatos <- read_delim("DataInfo_S013.csv")
datos <- read_csv("DataValues_S013.csv")
```

Según el archivo “datos”, existen 39 muestras o individuos y 695 variables, pero 4 de ellas (las primeras columnas) son “datos clínicos” o información de los individuos y el resto corresponde a la expresión de metabolitos. Cada columna que corresponde a los metabolitos, aparece el nombre del metabolito y T0, T2, T4 o T5, lo que indica que son los mismos metabolitos medidos en 4 puntos en el tiempo. En el archivo “metadatos” se encuentra la información de cada columna de “datos” como de qué tipo es esa columna y donde se puede encontrar más información. Está completo ya que el número de filas de los metadatos coincide con el número de columnas de los datos.

```
# Vamos a guardar en otro dataframe los datos clínicos
datos_clinicos <- datos[, 1:5]

# Eliminamos esas columnas/filas de ambos archivos
```

```
datos <- datos[, -c(1:5)]
metadatos <- metadatos[-c(1:5), ]

# Ahora hay 690 variables o medidas de metabolitos.
```

Se realizaron ciertas modificaciones a los archivos antes de almacenarlos en el SummarizedExperiment. En concreto, las columnas Group, GENDER y SURGERY de los datos clínicos se convirtieron en factores y se asignó el mismo nombre de fila a los datos clínicos y a los datos de expresión.

```
datos_clinicos <- datos_clinicos %>%
  mutate(across(c(SURGERY, GENDER, Group), as.factor))

# Vamos a asignarle nombres a las filas que sean informativos para poder
# representar las etiquetas de cada muestra en los análisis posteriores.
# Solo vamos a añadir la palabra Sample_ para que no sean números únicamente.
rownames(datos_clinicos) <- paste("Sample_", rownames(datos_clinicos), sep = "")
rownames(datos) <- rownames(datos_clinicos)
```

SummarizedExperiment

El análisis se llevó a cabo a partir de un SummarizedExperiment. Existen diferencias entre SummarizedExperiment y ExpressionSet. Esta última, está más orientada a los datos de expresión, mientras que SummarizedExperiment puede manejar varios tipos de datos más allá de los de expresión génica. ExpressionSet es más comúnmente utilizado en el análisis de datos de microarrays, aunque no es tan utilizado para RNA-Seq u otros tipos de datos más recientes, al contrario que el SummarizedExperiment. Además, en SummarizedExperiment rowData y colData pueden almacenar una variedad más amplia de información.

Se utilizó el paquete SummarizedExperiment para la creación del objeto.

```
# Creamos el objeto SummarizedExperiment
se <- SummarizedExperiment(
  assays = list(counts = as.matrix(datos)), # Es necesario que sea una matriz
  rowData = datos_clinicos,
  colData = metadatos
)

# Mostramos sus características
show(se)

## class: SummarizedExperiment
## dim: 39 690
## metadata(0):
## assays(1): counts
## rownames(39): Sample_1 Sample_2 ... Sample_38 Sample_39
## rowData names(5): SUBJECTS SURGERY AGE GENDER Group
## colnames(690): MEDDM_T0 MEDCOL_T0 ... SM.C24.0_T5 SM.C24.1_T5
## colData names(3): VarName varTpe Description
```

Una vez creado el objeto, se procedió con el análisis exploratorio de los datos, incluyendo la comprobación de valores faltantes. Para ello se utilizó el paquete Summarytools de R. La matriz assay fue normalizada mediante logaritmos y posteriormente, se llevó a cabo un análisis de componentes principales (PCA) con el fin de visualizar patrones en la distribución de los datos usando el paquete stats. Para su visualización se utilizó el paquete ggplot2 y RColorBrewer.

Adicionalmente, usando el paquete factorextra, se realizó un análisis de similitud entre muestras donde se llevó a cabo el cálculo de la distancia euclidiana entre las muestras de la matriz assay y se representó en un heatmap. Una matriz de distancias es una representación matemática que mide las diferencias o similitudes entre muestras en un espacio multidimensional. De manera que, valores pequeños indican que las muestras son similares en términos de sus perfiles de


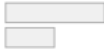
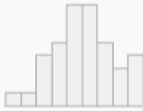
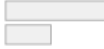
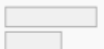
expresión génica, mientras que, valores grandes indican diferencias importantes. Con este mismo objetivo se ha realizado un dendrograma adicional al heatmap, utilizando el método “ward.D2”. En este árbol jerárquico se agrupan las muestras según su similitud, indicando, las ramas más cercanas, muestras con perfiles de expresión génica similares, mientras que las más alejadas representan diferencias mayores. Se ha utilizado el paquete factoextra.

RESULTADOS

Análisis Exploratorio

```
# Extraemos los datos clínicos y comprobamos si se han almacenado bien.
datos_clinicos <- as.data.frame(se@elementMetadata@listData)

# Usando el paquete summarytools hacemos una descriptiva básica y rápida de
# las variables. (Muestro una imagen y no la salida porque es propia de html.
# print(dfSummary(datos_clinicos, method = "render"))
```

| Variable | Stats / Values | Freqs (% of Valid) | Graph | Valid |
|--------------------|---|--------------------------|---|-------------|
| SUBJECTS [numeric] | Mean (sd) : 20 (11.4) min ≤ med ≤ max: 1 ≤ 20 ≤ 39 Q1 - Q3 : 10 - 30 | 39 distinct values |  | 39 (100.0%) |
| SURGERY [factor] | 1. by pass 2. tubular | 26 (66.7%) 13 (33.3%) |  | 39 (100.0%) |
| AGE [numeric] | Mean (sd) : 40.8 (9.9) min ≤ med ≤ max: 19 ≤ 41 ≤ 59 Q1 - Q3 : 35 - 46 | 22 distinct values |  | 39 (100.0%) |
| GENDER [factor] | 1. F 2. M | 27 (69.2%) 12 (30.8%) |  | 39 (100.0%) |
| Group [factor] | 1. 1 2. 2 | 24 (61.5%) 15 (38.5%) |  | 39 (100.0%) |

El conjunto de datos analizado incluye información sobre sujetos, edad, género, tipo de cirugía y grupo de pertenencia. El análisis descriptivo del conjunto de datos clínicos se representó en una tabla. Esta muestra varias columnas: “Variable” que indica el nombre de la variable y tipo de dato (numérica o factor), “Stats / Values” que muestra los estadísticos descriptivos para variables numéricas (media, desviación estándar, mínimo, máximo, mediana, cuartiles) o categorías disponibles (variables de tipo factor) y “Freqs (% of Valid)” que indica el número de valores únicos en variables numéricas o la distribución de frecuencias y porcentaje para variables categóricas. Además, en la columna “graph” se representa la distribución de los datos en un gráfico de barras y en “Valid” la cantidad total de observaciones sin valores perdidos y su porcentaje sobre el total.

En cuanto al tipo de cirugía (SURGERY), los procedimientos se dividen en dos categorías: bypass y tubular. La mayoría de los sujetos (66.7%) han sido sometidos a un bypass, mientras que el 33.3% restante ha recibido cirugía tubular. La edad (AGE) de los participantes se encuentra entre los 19 y 59 años, con una media de 40.8 años y una mediana de 41. La dispersión en los valores es moderada, con la mitad de los datos concentrados entre los 35 y 46 años. El género (GENDER) de los sujetos está compuesto mayoritariamente por mujeres, quienes representan el 69.2% de la muestra, mientras que los hombres constituyen el 30.8%. Por último, en cuanto a la clasificación por grupos (Group), se identifican dos categorías. La mayor parte de los sujetos (61.5%) pertenece al Grupo 1, mientras que el 38.5% corresponde al Grupo 2.

Ninguna de las variable clínicas presentan valores faltantes.

```
# Ahora vamos a comprobar los valores faltantes en la matriz de expresión.
# Extraemos la matriz del objeto.
assays <- se@assays@data@listData[["counts"]]
cat("Hay", sum(is.na(assays)), "valores faltantes. \n")
```

```
## Hay 3390 valores faltantes.
```

```
# Calculamos el porcentaje de valores faltantes.
```

```
cat("El porcentaje de valores faltantes es", (sum(is.na(assays))/length(assays))*100, "\n")
```

```
## El porcentaje de valores faltantes es 12.59755
```

```
# En este caso, vamos a completar los valores faltantes porque a la hora de hacer
```

```
# los análisis pueden dar problemas. Añadimos 0 ya que no sabemos
```

```
# porque no están esos valores.
```

```
assays[is.na(assays)] <- 0
```

```
cat("Hay", sum(is.na(assays)), "valores faltantes. \n")
```

```
## Hay 0 valores faltantes.
```

```
# Introducimos la matriz sin valores faltantes al objeto
```

```
assays(se)$counts_sinNA <- assays
```

```
show(se)
```

```
## class: SummarizedExperiment
```

```
## dim: 39 690
```

```
## metadata(0):
```

```
## assays(2): counts counts_sinNA
```

```
## rownames(39): Sample_1 Sample_2 ... Sample_38 Sample_39
```

```
## rowData names(5): SUBJECTS SURGERY AGE GENDER Group
```

```
## colnames(690): MEDDM_TO MEDCOL_TO ... SM.C24.0_T5 SM.C24.1_T5
```

```
## colData names(3): VarName varTpe Description
```

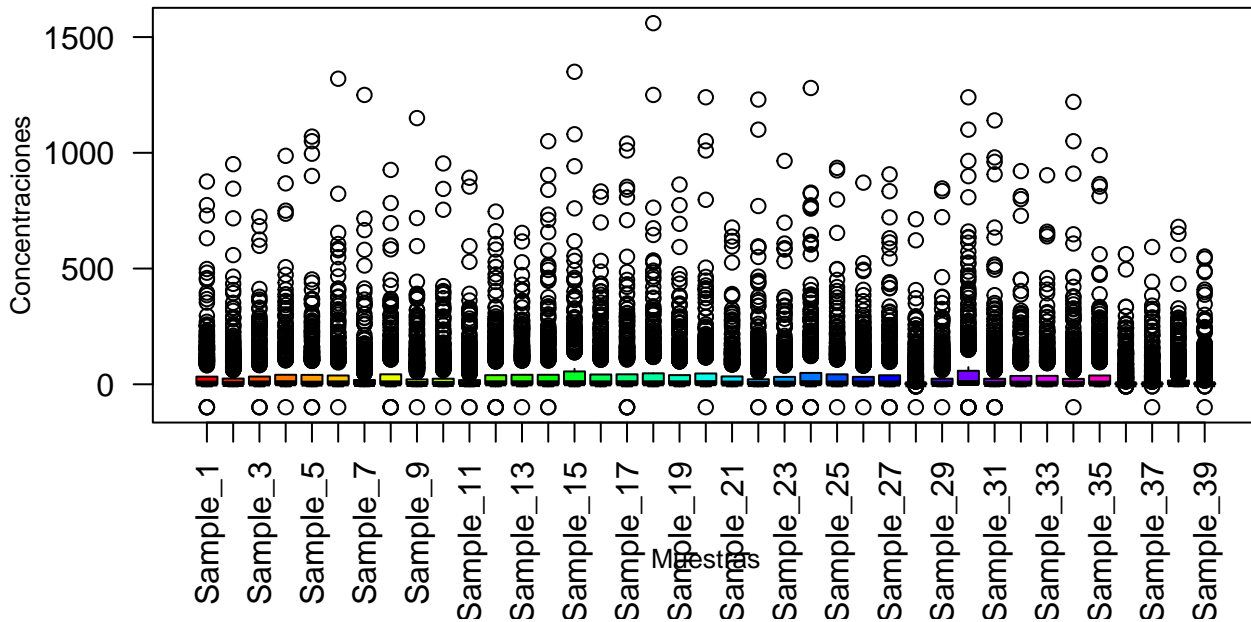
Por otro lado, los datos de metabolitos, mostraron un 12.60% de valores faltantes, los cuales fueron completados con el valor 0.

```
assays <- t(as.data.frame(se@assays@data@listData[["counts_sinNA"]]))
```

```
# Creamos un gráfico de barras con boxplot
```

```
boxplot(assays,  
  xlab = "Muestras",  
  ylab = "Concentraciones",  
  cex.lab = 0.8,  
  horizontal = FALSE,  
  las = 2,  
  main = "Distribución de la concentración de metabolitos por muestra",  
  cex.main = 0.8,  
  col = rainbow(ncol(assays)))
```

Distribución de la concentración de metabolitos por muestra



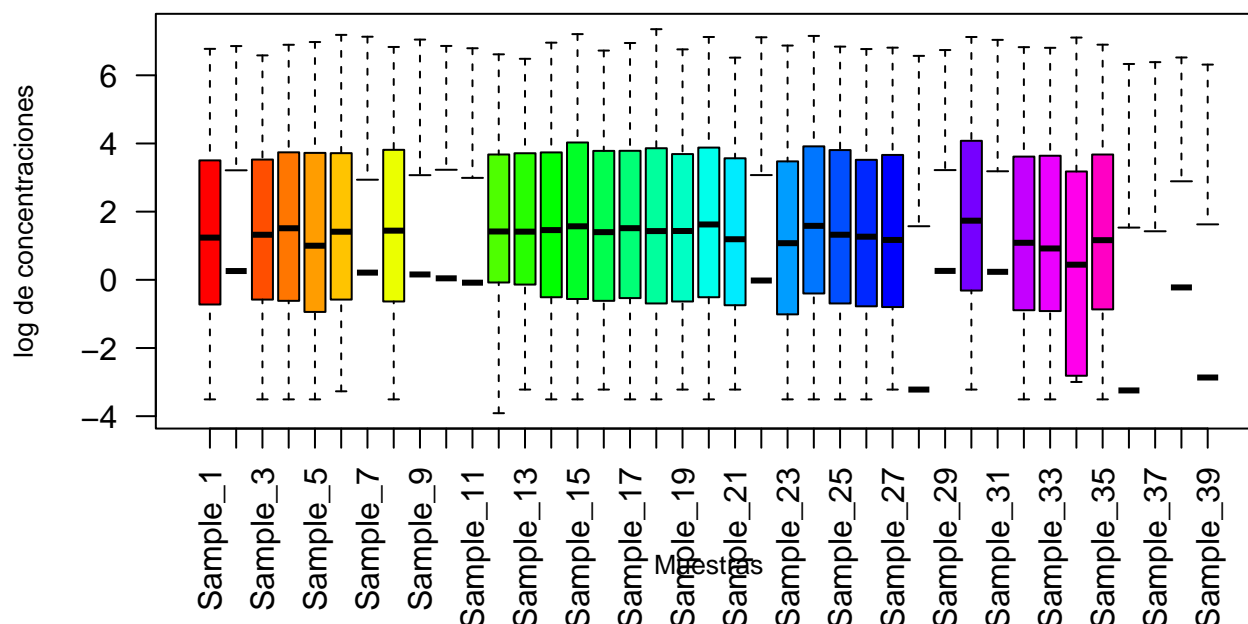
Existió mucha variabilidad entre las concentraciones de los metabolitos por cada muestra, lo que justificó la necesidad de normalizar los datos mediante un logaritmo.

```
assays <- se@assays@data@listData[["counts_sinNA"]]

# Introducimos el logaritmo de la matriz al objeto
assays(se)$counts_log <- log(assays)
assays <- t(se@assays@data@listData[["counts_log"]])

# Volvemos a crear el gráfico de barras
boxplot(assays,
        xlab = "Muestras",
        ylab = "log de concentraciones",
        cex.lab = 0.8,
        horizontal = FALSE,
        las = 2,
        main = "Distribución de la concentración de metabolitos por muestra",
        cex.main = 0.8,
        col = rainbow(ncol(assays)))
```

Distribución de la concentración de metabolitos por muestra



Análisis de Componentes Principales

```
# Para hacer el PCA usamos la matriz sin normalizar debido a que al haber
# sustituido los valores faltantes por 0 y aplicar el logaritmo, aparecen nuevamente
# NAs o valores infinitos.
```

```
assays <- se@assays@data@listData[["counts_sinNA"]]
```

```
# Realizamos el PCA
pca_obj <- prcomp(assays)
```

```
# Obtenemos la varianza explicada por cada componente
var_exp <- (pca_obj$sdev^2) / sum(pca_obj$sdev^2) * 100
pca_data <- as.data.frame(pca_obj$x)
```

```
# Cargamos los datos clínicos para usarlo en la leyenda
datos_clinicos <- as.data.frame(se@elementMetadata@listData)
pca_data$Group <- datos_clinicos$Group
pca_data$GENDER <- datos_clinicos$GENDER
pca_data$SURGERY <- datos_clinicos$SURGERY
```

```
# Definimos las etiquetas con la varianza explicada y los colores
x_lab <- paste0("PC1 (", round(var_exp[1], 2), "%)")
y_lab <- paste0("PC2 (", round(var_exp[2], 2), "%)")
```

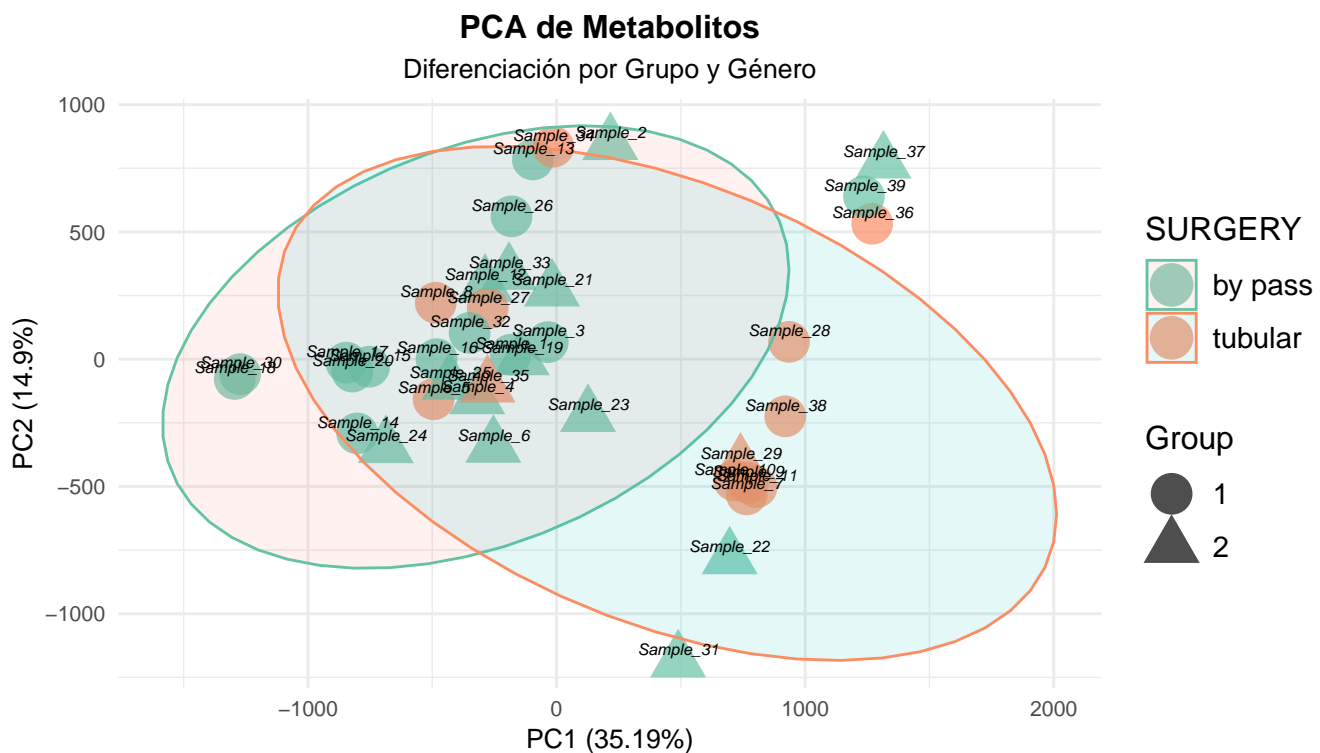
```
n_colors <- length(unique(pca_data$Group))
colors <- brewer.pal(n_colors, "Set2")
```

```
# Dibujamos el PCA con elipses del 95% de confianza para cada grupo
ggplot(pca_data, aes(x = PC1, y = PC2, color = SURGERY, shape = Group)) +
  geom_point(size = 7, alpha = 0.7) +
```

```

stat_ellipse(aes(group = SURGERY, fill = SURGERY),
             level = 0.95, geom = "polygon", alpha = 0.1) +
scale_color_manual(values = colors) +
labs(title = "PCA de Metabolitos",
     subtitle = "Diferenciación por Grupo y Género",
     x = x_lab, y = y_lab) +
theme_minimal() +
theme(
  text = element_text(size = 10),
  legend.title = element_text(size = 12),
  legend.text = element_text(size = 11),
  plot.title = element_text(hjust = 0.5, size = 12, face = "bold"),
  plot.subtitle = element_text(hjust = 0.5, size = 10)
) +
geom_text(aes(label = rownames(pca_data)), hjust = 0.5, vjust = -0.5,
          size = 2, fontface = "italic", color = "black")

```



El análisis de Componentes Principales (PCA) indica que la mayoría de la variabilidad entre las muestras es explicada por el tipo de cirugía, aunque no todas las muestras quedan separadas. La otra variable representada (grupo) no muestra ninguna influencia en la variabilidad de las muestras.

```

# Para calcular el peso de las variables en cada componente, lo extraemos de
# la columna rotation
loadings <- as.data.frame(pca_obj$rotation)
loadings$Variable <- rownames(loadings)

# Asignamos un umbral para que filtre las variables que tienen una carga absoluta
# superior a 0.2 en PC1 o PC2.
threshold <- 0.2
top_variables <- loadings %>%
  filter(abs(PC1) > threshold)

top_variables2 <- loadings %>%

```

```

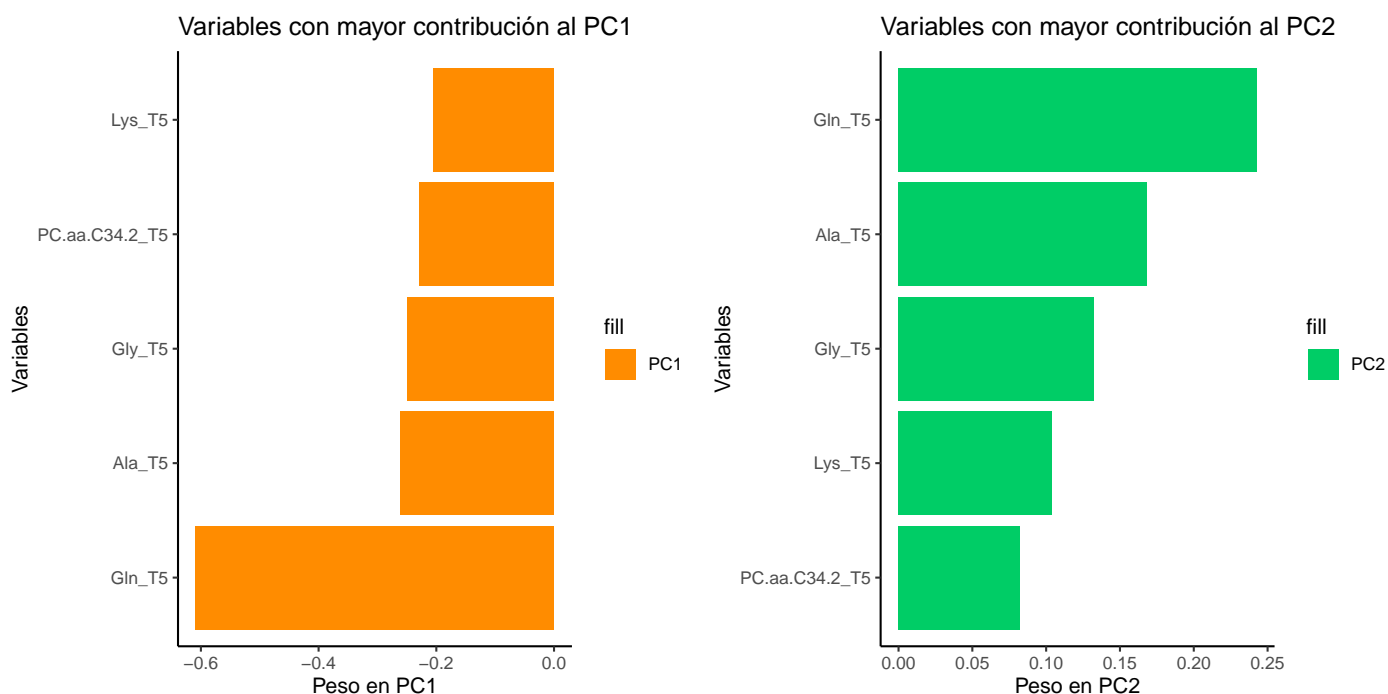
filter(abs(PC1) > threshold | abs(PC2) > threshold)

# Representamos las variables que más contribuyen en los componentes en un gráfico
g1 <- ggplot(top_variables, aes(x = reorder(Variable, PC1), y = PC1, fill = "PC1")) + geom_bar(stat = "identity")
coord_flip() +
scale_fill_manual(values = c("PC1" = "darkorange")) +
labs(title = "Variables con mayor contribución al PC1", x = "Variables",
      y = "Peso en PC1") +
theme_classic()

g2 <- ggplot(top_variables, aes(x = reorder(Variable, PC2), y = PC2, fill = "PC2")) +
geom_bar(stat = "identity", position = "dodge") +
coord_flip() +
scale_fill_manual(values = c("PC2" = "springgreen3")) +
labs(title = "Variables con mayor contribución al PC2", x = "Variables",
      y = "Peso en PC2") +
theme_classic()

grid.arrange(g1, g2, ncol = 2)

```



Además, observamos que los metabolitos asociados al componente 1 son Ala_T5, Gly_T5, Gln_T5, Lys_T5 y PC.aa.C34.2_T5, teniendo todos cargas negativas, lo que significa que cuando el valor de los metabolitos disminuye, el valor del PC aumenta. El que todos tengan cargas negativas podría sugerir que se “mueven” en conjunto. Mientras que los metabolitos que más contribuyen al 2 son Gln_T0, Gln_T2, Ala_T4, Gln_T4, PC.aa.C34.2_T4 y Gln_T5, todos con carga negativa a excepción de Gln_T5.

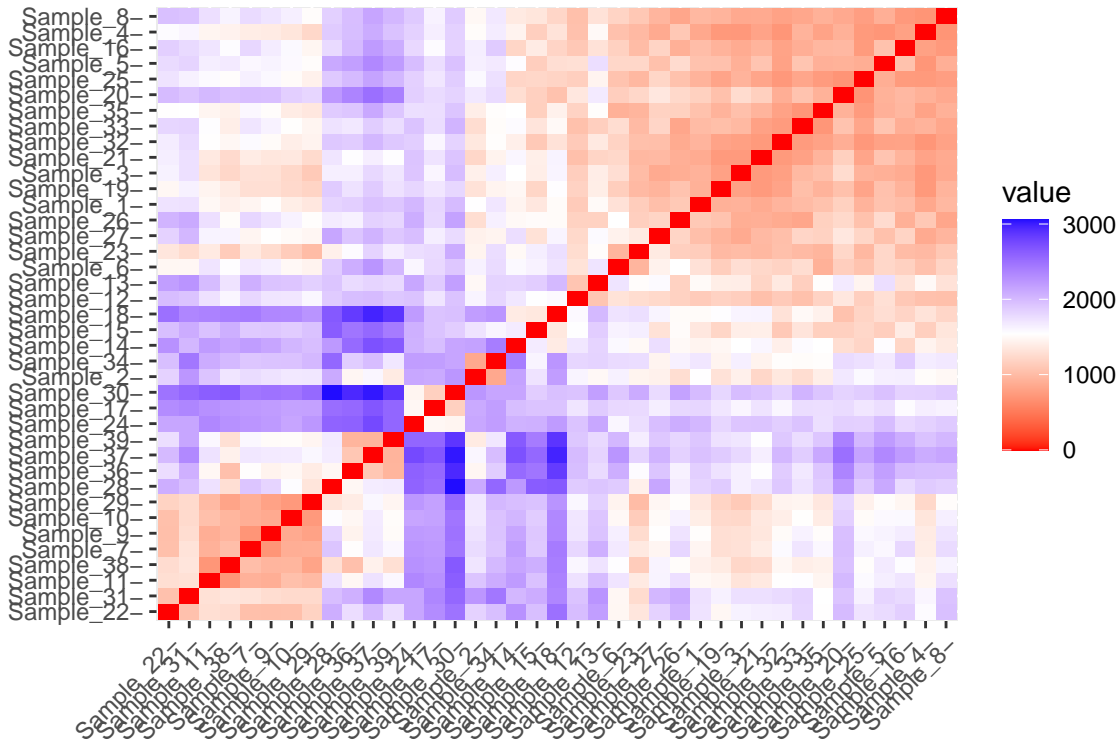
Análisis de similitud entre muestras

```

# Calculamos la distancia euclidiana de la matriz de datos
dist_matrix <- dist(assays, method = "euclidean")

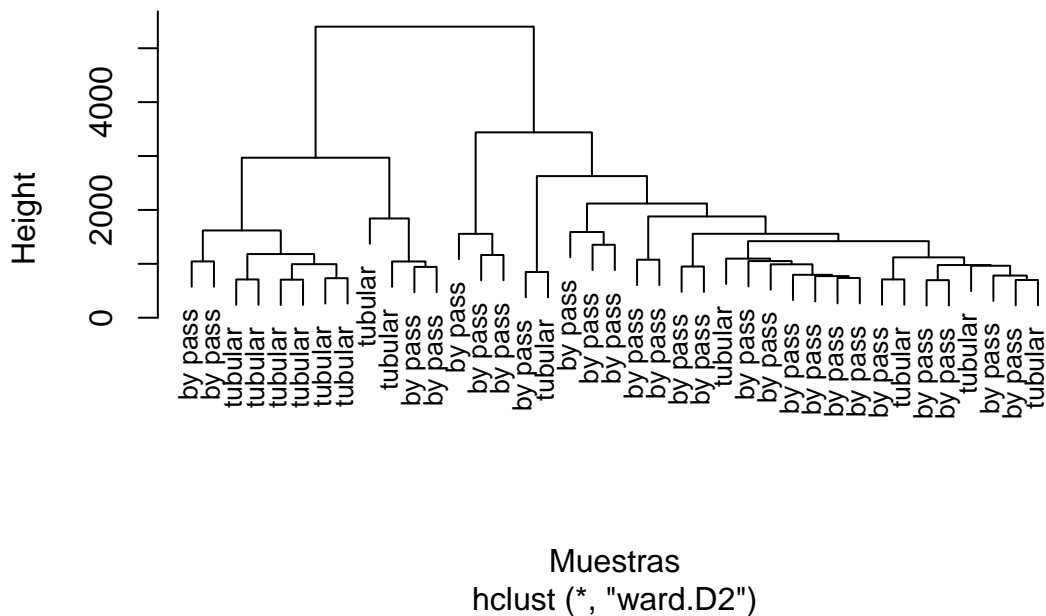
# Representamos las distancias en el heatmap
fviz_dist(dist_matrix)

```

```
# Dibujamos el dendrograma de las distancias
plot(hclust(dist_matrix, method = "ward.D2"), labels = datos_clinicos$SURGERY,
     main = "Dendrograma de distancias entre muestras", xlab = "Muestras",
     cex=0.8)
```

Dendrograma de distancias entre muestras



El dendrograma y heatmap basado en la distancia euclidiana entre muestras indican la tendencia a agruparse los datos según su tipo de cirugía. En la rama principal izquierda del dendrograma, mayoritariamente se agrupan las muestras con un tipo de cirugía “tubular”. Al contrario que la rama derecha, donde se agrupan mayoritariamente las del tipo de cirugía “by pass”.

DISCUSIÓN

En conclusión, los resultados sugieren que el tipo de cirugía es un factor determinante en las diferencias observadas en las concentraciones de metabolitos, lo que podría tener implicaciones importantes para entender cómo los procedimientos quirúrgicos afectan el perfil metabólico de los pacientes. Tanto el análisis de compoinentes principales como el análisis de similitud entre muestras indican que el tipo de cirugía tiene un padel potencial, pero también abre la puerta a explorar otras variables que podrían interactuar con los metabolitos y aportar más claridad en el análisis.

La distribución de la edad mostró una muestra predominantemente adulta, con una media cercana a los 41 años, lo que sugiere que los participantes son en su mayoría adultos jóvenes. El análisis del género reveló una mayor representación de mujeres, por lo que es necesario más representación masculina.

El análisis reveló que no había valores faltantes en las variables clínicas, pero sí en los datos de expresión de metabolitos. En los datos de expresión de metabolitos, sí se encontraron valores faltantes, lo cual es relevante al momento de interpretar los resultados, ya que estos valores fueron imputados con el valor cero, lo que podría afectar la precisión del análisis.

Además, considero que el hecho de que las mediciones de los metabolitos estén representadas en función del tiempo resulta un tanto confuso y podría mejorarse. Sería más adecuado organizar los datos de manera que se asignen múltiples columnas de muestras para cada tiempo, en lugar de tener una de cada metabolito con los tiempos. El formato ideal sería una fila por cada metabolito y tantas columnas como muestras y tiempo. De esta forma, el tiempo podría ser tratado como una variable independiente contenida en los datos clínicos y sería posible observar de manera más clara cómo cambia la concentración de los metabolitos a lo largo del tiempo, dependiendo de la evolución postoperatoria y del tipo de cirugía.

CONCLUSIONES

Las conclusiones extraídas de este análisis son:

- El tipo de cirugía es un factor determinante en las concentraciones de metabolitos medidos después de la intervención.
- Otras variables podrían estar implicadas en la variabilidad de la concentración de metabolitos entre las muestras, adicional e independientemente del tipo de cirugía.
- Es necesario un diseño de los datos óptimo para optimizar al máximo los resultados.

REFERENCIAS

<https://github.com/Mjbeltranrod/Beltran-Rodriguez-Maria-Jose-PEC1>