Informe PEC1 Análisis de datos ómicos

Adrián Parrilla Mesas

Abstract

En este trabajo se ha realizado un análisis exploratorio de un dataset de metabolómica. El estudio escogido fue realizado en 2023 por investigadores de la Universidad de Texas y se centra en conocer mejor las adaptaciones metabólicas fetales al estrés cardiovascular en el síndrome de transfusión gemelo - gemelo (TTTS por sus siglas en inglés). Tras el análisis inicial del tipo, número y distribución de las variables, se realiza una heatmap y un análisis de componentes principales (PCA) para obtener una visión global de las muestran. Los resultados permiten distinguir una separación clara a nivel metabólico entre los grupos control y TTTS, identificando un grupo de metabolitos que podrían estar implicados.

Objetivos

Los objetivos de este trabajo son:

- Obtener y procesar los datos del estudio
- Realizar un análisis exploratorio de las muestras presentes
- Evaluar las diferencias generales entre los distintos grupos experimentales

Métodos

El dataset empleado fue obtenido a través del repositorio *Metabolics Workbench* mediante el ID del estudio ST002797. Las muestras proceden de líquido amniótico de mujeres embarazadas de gemelos monocoriónicos - diamnióticos que han sufrido una complicación TTTS. Tras el procesamiento de las muestras, los datos fueron obtenidos mediante LC-MS y aparecen pre-procesados en formato tabular mzML con los valores indicando la intensidad de los picos. Para su análisis se empleó el software Rstudio (4.3.0) con las librerías metabolomicsWorkbenchR (1.10.0) y SummarizedExperiment (1.30.2).

Resultados

Obtención de los datos

Una de las ventajas del repositorio *Metabolics Workbench* es que permite la descarga directa en R de los archivos de un estudio mediante la librería metabolomicsWorkbenchR. La elección del dataset se basa en su notable número de muestras y en el formato en el que se han depositado los resultados, lo que permite realizar un análisis comparativo sin necesidad de hacer un preprocesado de los datos de espectrometría de masas. Además, el tema del estudio es relativamente infrecuente, lo que lo hace más interesante y supone una motivación extra para indagar en este campo.

La librería metabolomicsWorkbenchR permite compilar los datos del estudio directamente en un objeto SummarizedExperiment simplemente indicando su ID. Cómo en el estudio aparecen dos análisis distintos y no se apreció diferencias significativas, se decide continuar solamente con el primero de ellos.

```
df <- do_query(context = 'study',
               input_item = 'study_id',
               input_value = 'ST002797',
               output_item = 'SummarizedExperiment')

df <- df[[1]]
```

Exploración de los metadatos

A continuación, se procede a explorar los metadatos del estudio y la información del número de muestras por grupo.

```
df@colData[1:6,]
```

```
DataFrame with 6 rows and 6 columns
        local_sample_id    study_id  sample_source mb_sample_id
            <character> <character>    <character>  <character>
C10_215         C10_215    ST002797 Amniotic fluid     SA300456
C1_160           C1_160    ST002797 Amniotic fluid     SA300454
C2_171           C2_171    ST002797 Amniotic fluid     SA300457
C3_183           C3_183    ST002797 Amniotic fluid     SA300462
C4_186           C4_186    ST002797 Amniotic fluid     SA300461
C5_189           C5_189    ST002797 Amniotic fluid     SA300463
                 raw_data    Group
```

```
              <character> <factor>
C10_215 C10_215.mzdata.xml  Control
C1_160   C1_160.mzdata.xml  Control
C2_171   C2_171.mzdata.xml  Control
C3_183   C3_183.mzdata.xml  Control
C4_186   C4_186.mzdata.xml  Control
C5_189   C5_189.mzdata.xml  Control
```

```
table(df@colData$Group)
```

```
Control    TTTS
     10      22
```

Se observa que hay más del doble de muestras del grupo TTTS que del control.

Por otro lado, es posible acceder a lista de los metabolitos identificados mediante:

```
rowData(df)[1:10,]
```

```
DataFrame with 10 rows and 3 columns
                metabolite_name metabolite_id            refmet_name
                    <character>   <character>            <character>
ME725430 10(E)-Heptadecenoic ..      ME725430
ME725432 "10(Z),13(Z)-Nonadec..      ME725432
ME725436 10(Z)-Heptadecenoic ..      ME725436 10Z-Heptadecenoic acid
ME725425     11-Dodecenoic Acid      ME725425
ME725435  11(E)-Eicosenoic Acid      ME725435     trans-Gondoic acid
ME725431 (11E)-Octadecenoic a..      ME725431    trans-Vaccenic acid
ME725428    12-Tridecenoic Acid      ME725428
ME725427 "12(Z),15(Z)-Heneico..      ME725427
ME725438 "1,3-Dipalmitoylglyc..      ME725438
ME725423       13 Retinoic Acid      ME725423
```

El número de assays del estudio se puede comprobar mediante:

```
df@assays
```

```
An object of class "SimpleAssays"
Slot "data":
List of length 1
```

Exploración de los datos

Como se ha mencionado, los datos aparecen ya procesados en formato tabular. Sin embargo, es necesaria la imputación de los valores nulos y su normalización para los posteriores análisis.

```
data <- as.data.frame(assay(df))

dim(data)
```

```
[1] 165  32
```

Los valores nulos se transforman en 1 para evitar posibles problemas a posteriori durante el análisis.

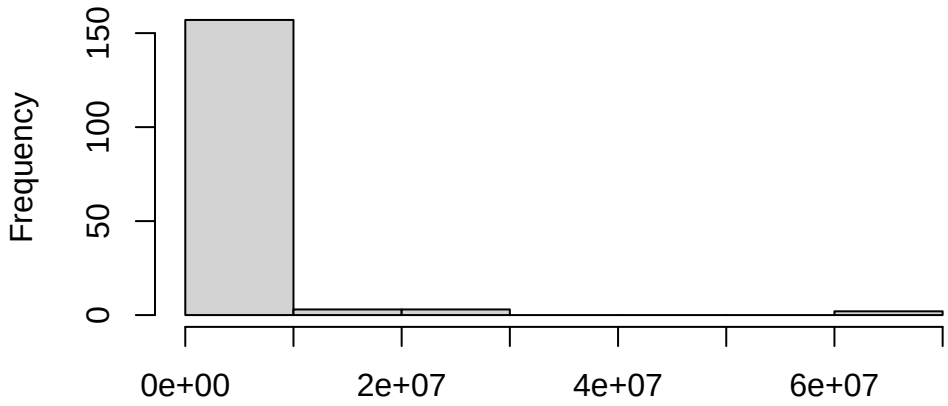
```
sum(is.na(data))
```

```
[1] 376
```

```
data[is.na(data)] <- 1
```

Con el fin de observar la distribución de los datos y de valorar si es necesaria su transformación, se visualiza una de las variables.

```
hist(data$C10_215, main = '', xlab = '')
```



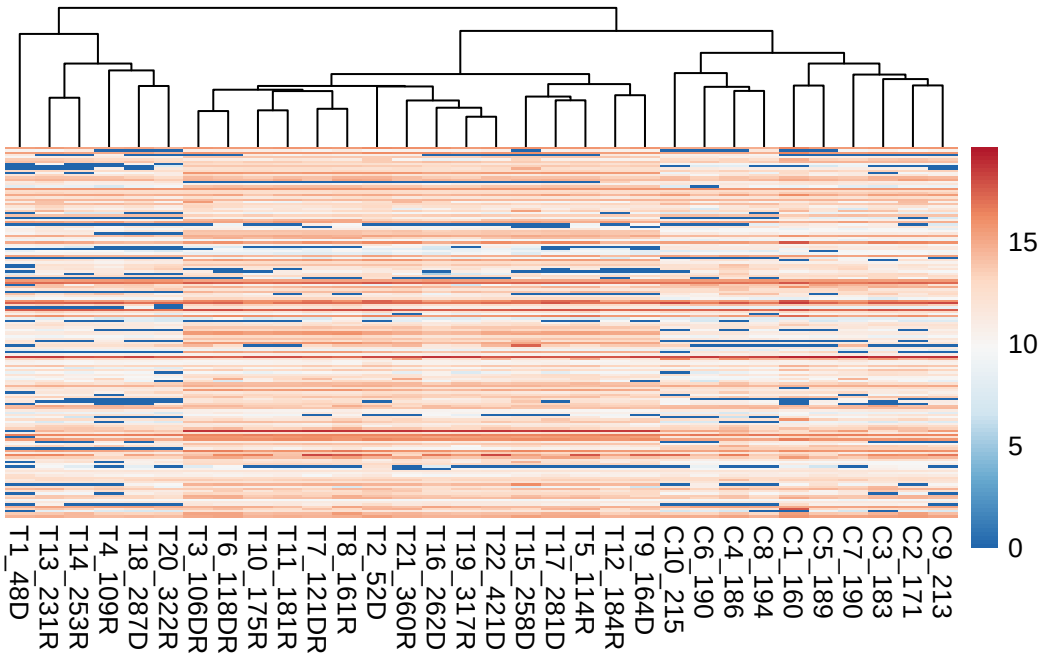
Al presentar un rango muy elevado, se procede a la normalización de los datos mediante el logaritmo natural.

```
data_norm <- log(data)
```

Heatmap y clustering

Para obtener una visión global de como se distribuyen las muestras e identificar si hay una sobrerepresentación aparente de algún metabolito, se procede a generar un heatmap. Además, se realiza un clustering jerárquico para comprobar que el agrupamiento de las muestras sea el esperado.

```
rownames(data_norm) <- rowData(df)$metabolite_name

pheatmap(data_norm,
         show_rownames = F,
         cluster_rows = F,
         cluster_cols = T,
         clustering_method = "complete",
         color = colorRampPalette(rev(brewer.pal(n = 7, name = "RdBu")))(100),
         )
```



Se observa dos clusteres diferenciados entre el grupo TTTS y el control, lo cual da indicios de las posibles diferencias subyacentes. Para confirmar este hallazgo, se realiza un análisis de componentes principales.

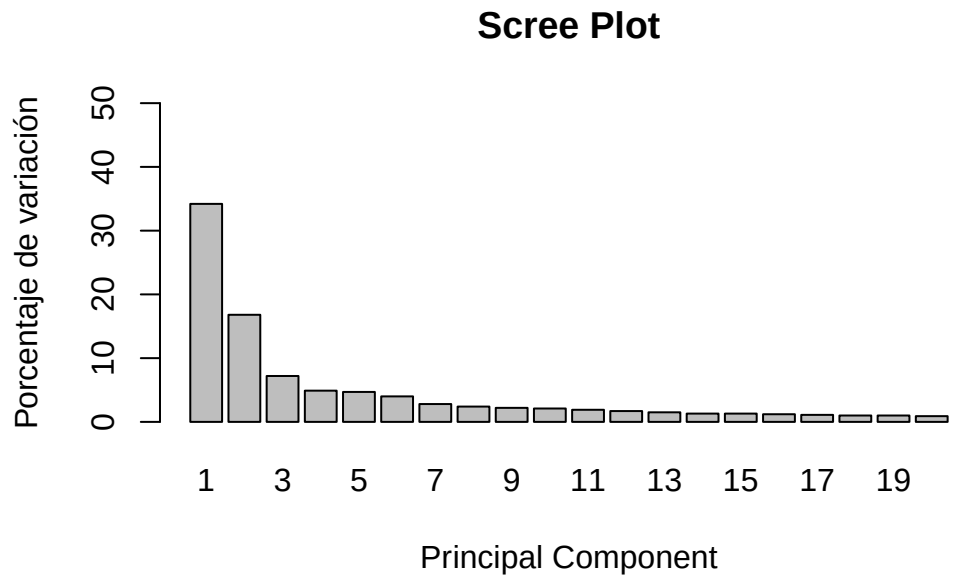
Principal component analysis (PCA)

```
data_norm_pca <- t(data_norm)

pca <- prcomp(data_norm_pca, scale = TRUE)
```

El porcentaje de varianza explicado por cada componente se ha calculado de la siguente manera:

```
pca.var <- pca$sdev^2
pca.var.per <- round(pca.var/sum(pca.var)*100, 1)
```

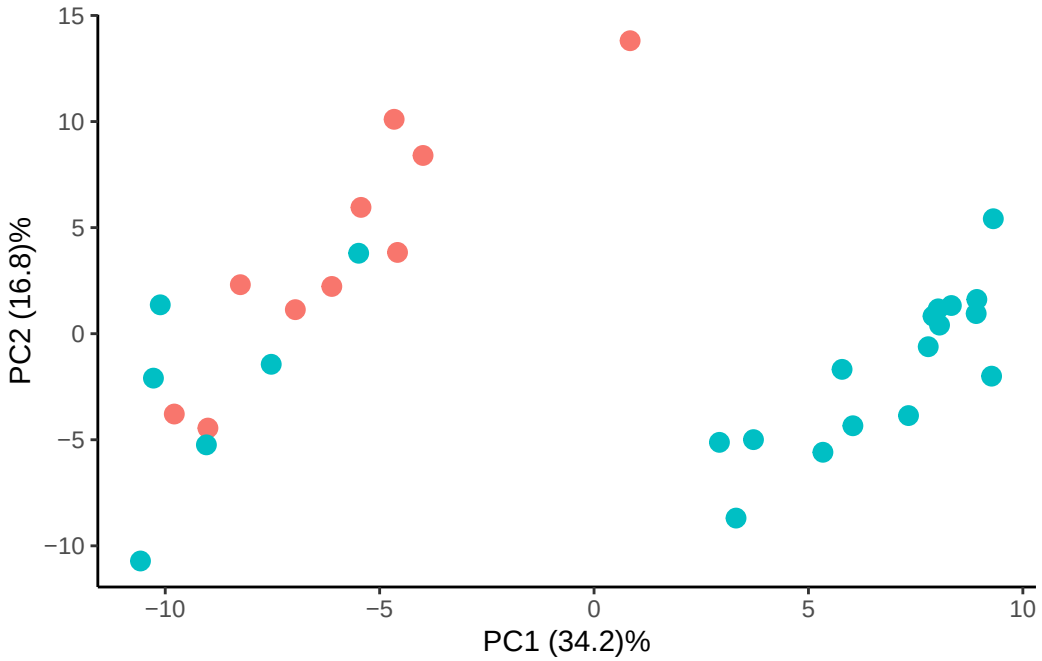
```
scree_plot <- barplot(pca.var.per[1:20], main="Scree Plot",
                      xlab="Principal Component",
                      ylab="Porcentaje de variación",
                      ylim = c(0,50),
                      names = seq(1:20))
```



Seguidamente, se preparan los datos en un dataframe y representan las 3 primeras componentes.

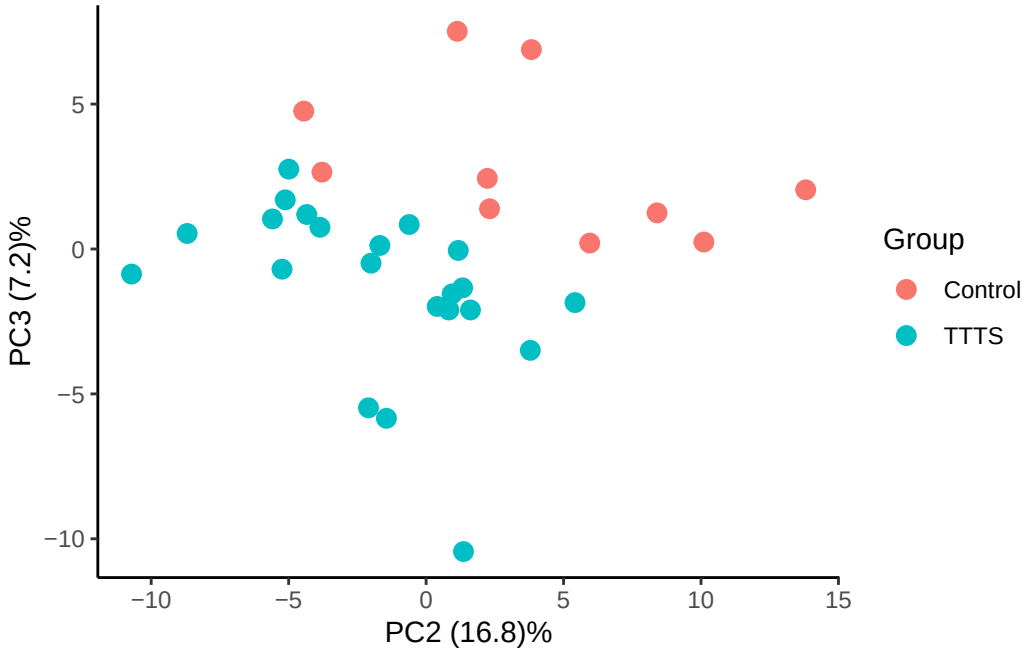
```
group <- c(rep(c("Control"), 10), rep("TTTS", 22))

pca.data <- data.frame(Sample=rownames(pca$x),
                       Group= group,
                       PC1=pca$x[,1], #PCA1
                       PC2=pca$x[,2], #PCA2
                       PC3=pca$x[,3]) #PCA3

pca.data$Group <- as.factor(pca.data$Group)
```

```
ggplot(data=pca.data, aes(x=PC1, y=PC2)) +
  geom_point(aes(color= Group), size = 3)+
  xlab(paste("PC1 (", pca.var.per[1],")%", sep="")) +
  ylab(paste("PC2 (", pca.var.per[2],")%", sep="")) +
  theme_classic() +
  theme(legend.position="none")
```




```
ggplot(data=pca.data, aes(x=PC2, y=PC3)) +
  geom_point(aes(color= Group), size = 3)+
```

```
  xlab(paste("PC2 (", pca.var.per[2],")%", sep="")) +
  ylab(paste("PC3 (", pca.var.per[3],")%", sep="")) +
  theme_classic()
```



Se observa una clara separación entre ambos grupos el la primera componente principal, donde todas las muestras control se encuentran agrupadas. A su vez, encontramos algunas muestras del grupo TTTS junto con las del control.

Por otro lado, se puede además obtener información acerca de qué metabolitos influyen más en la separación de las muestras.

```
loading_scores <- pca$rotation[,1]

metab_scores <- abs(loading_scores) # me quedo con las magnitudes
metab_scores_ranked <- sort(metab_scores, decreasing=TRUE)

top_10_metabs <- names(metab_scores_ranked[1:10])

pca$rotation[top_10_metabs,1]
```

```
4-Imidazoleacetic Acid              Hydrocinnamic Acid
            -0.1284947                       0.1271937
  Docosahexaenoic Acid                    Nervonic Acid
```

```
                0.1247929                       0.1247638
3-Aminoisobutyric Acid Eicosatrienoic Acid (20:3 N-3)
              -0.1246159                       0.1245886
Dihomo-Gamma-Linolenic Acid           Mead Acid (20:3 N-9)
                0.1245746                       0.1245331
          Retinyl Palmitate   "5,8,11-Eicosatriynoic Acid"
                0.1244252                       0.1241605
```

Se observa que el 4-Imidazoleacetic Acid es el que guía la separación del grupo control hacia la izquierda, aunque no hay una gran diferencia respecto al resto de moléculas.

Discusión

En el presente trabajo se ha realizado un análisis exploratorio de una dataset público de metablómica, obteniendo una visión general de la distribución de las muestras bastante prometedora de cara a un análisis futuro más detallado. El heatmap junto con el clustering y el PCA han permitido observar que existe una diferencia a nivel metabólico entre las muestras del grupo control y TTTS, con algunas de este último siendo similares a las del control. Sería necesario pues obtener más información acerca de las muestras clínicas para estudiar un posible *batch effect*. Aprovechando el buen número de replicados biológicos del estudio, se podría investigar si existen diferencias significativas en la presencia de algún metabolito mediante un análisis de expresión diferencial. De esta forma se podrían obtener metabolitos candidatos sobre los que centrar una futura hipótesis.

Conclusiones

Las conclusiones de este trabajo son las siguientes:

- Las muestras del grupo TTTS parecen metabolicamente distintas a las del grupo control.
- El paquete metabolomicsWorkbenchR permite el análisis de datos publicos de metabolómica de una forma sencilla y fluida.

Referencias

1. Metabolomic Workbench ID: ST002797
2. Github repository