Análisis de datos ómicos (M0-157). Primera prueba de evaluación continua

María Torés España

2025-03-31

Contents

1 Abstract.

En este estudio, se analizó un conjunto de datos de metabolómica para investigar los metabolitos asociados a la respuesta a la cirugía bariátrica, independientemente de la magnitud de la pérdida de peso. Además, se incluyen otros factores factores clínicos como la edad o el tipo de cirugía.

Se utilizó la clase SummarizedExperiment para organizar los datos y los metadatos, permitiendo un análisis más estructurado y accesible. A través de un análisis exploratorio, incluyendo la visualización de valores faltantes y un análisis de componentes principales (PCA), se identificaron patrones metabólicos clave que podrían estar vinculados a la respuesta clínica a la intervención. A pesar de las limitaciones del dataset, como la presencia de valores faltantes, los resultados sugieren que factores metabólicos más allá de la pérdida de peso pueden influir en la respuesta al tratamiento.

2 Objetivos.

El objetivo principal de este trabajo es realizar un análisis de un conjunto de datos de metabolómica mediante la creación de un objeto de clase `SummarizedExperiment` que contenga tanto los datos experimentales como los metadatos asociados. A partir de este objeto, se llevará a cabo un análisis exploratorio para obtener una visión general de los datos, lo que permitirá identificar patrones, relaciones y posibles áreas de interés en el contexto biológico.

El trabajo se complementará con la creación de un repositorio de GitHub que contendrá el informe, el objeto `SummarizedExperiment` en formato binario, el código R debidamente comentado, los datos en formato texto y los metadatos acompañados de una breve descripción en un archivo markdown.

3 Métodos.

En este trabajo, se ha trabajado con un conjunto de datos de metabolómica, que incluye tanto las concentraciones de metabolitos en las muestras como los metadatos asociados a las mismas. Los datos fueron obtenidos a partir de archivos CSV que contienen dos componentes principales: los valores de las mediciones de metabolitos y otros datos clínicos (`DataValues_S013.csv`) y los metadatos asociados a las muestras (`DataInfo_S013.csv`). Se selecciona este conjunto de datos porque incluye el archivo de metadatos, lo que facilita la interpretación y análisis de las muestras.

3.1 Procesamiento de datos.

Los pasos iniciales incluyeron la carga de los datos en R. Posteriormente, se eliminaron las columnas innecesarias, como la columna de número de línea en el caso de los datos y los encabezados adicionales en los metadatos. Se identificaron las primeras cinco columnas del conjunto de datos como la información de las muestras, que fue separada del conjunto de datos principal para su uso posterior en la creación del objeto `SummarizedExperiment`.

3.2 Objeto `SummarizedExperiment`.

El objeto de clase `SummarizedExperiment` fue creado para estructurar adecuadamente los datos y los metadatos. En esta clase, los datos numéricos de metabolómica son almacenados en una matriz, mientras que los metadatos (información sobre las muestras) son almacenados en un data frame. Los datos de las muestras fueron transformados en un formato adecuado, con la creación de un identificador único para cada muestra que se utilizó como nombre de las filas en el objeto `SummarizedExperiment`. Las variables categóricas fueron convertidas en factores.

3.3 Análisis exploratorio.

El análisis exploratorio de los datos se llevó a cabo en varias etapas. En primer lugar, se realizó un análisis descriptivo básico de las variables categóricas mediante la creación de tablas de frecuencia para cada una

de las variables del `rowData`. Además, se calcularon las proporciones relativas para obtener una visión más clara de la distribución de las categorías.

Para las variables numéricas, se aplicaron estadísticas descriptivas. Se observó que los datos presentaban una distribución asimétrica, por lo que se optó por la transformación logarítmica para estabilizar la varianza y aproximar la distribución a una forma más simétrica.

Se realizó un análisis de los valores faltantes en los datos, visualizándolos mediante la función gg_miss_upset(), que permite identificar patrones de datos faltantes a lo largo de las muestras. Además, se asignaron valores de 1 a los valores faltantes en los datos para poder realizar un análisis sin perder información crucial.

Uno de los enfoques principales del análisis exploratorio fue la reducción de dimensionalidad mediante el análisis de componentes principales (PCA). Este método permitió visualizar la variabilidad global de los datos y detectar posibles patrones en las muestras. Los resultados del PCA fueron visualizados mediante gráficos como el `pairsplot()`, que mostró las relaciones entre los componentes principales, y el `eigencorplot()`, que visualizó las correlaciones de los componentes principales con las variables de metadatos (como el tipo de cirugía). También se generó un gráfico `biplot` para visualizar tanto las muestras como las variables en el espacio de los componentes principales.

Además de PCA, se generó un `boxplot` para evaluar la distribución de las concentraciones en las muestras, lo que permitió obtener una visión rápida de las variaciones entre los sujetos y las concentraciones medidas.

Para el análisis estadístico, se utilizó la corrección de *Benjamini-Hochberg* para controlar el error de tipo I en las correlaciones múltiples, lo cual fue esencial para manejar las pruebas realizadas entre los componentes principales y las variables de metadatos.

4 Resultados.

La principal diferencia entre `SummarizedExperiment` y `ExpressionSet` radica en los argumentos `assays()` y `exprs()` respectivamente. Cuando trabajamos con `exprs()` estamos enfocados en lo que seria datos de transcriptómica (fila=gen y columna=sample). `assays()` es más global (más general), almacena como lista (lo que permite acceder a una matriz específica dentro de una lista de assays: raw, normalised…) y cualquier tipo de datos (metabolómica, proteómica…)

4.1 Análisis exploratorio de datos.

Este análisis descriptivo están orientados a obtener una comprensión más completa de los datos y sus relaciones. Algunos estadísticos relevantes podrían ser:

4.1.1 Tabla de frecuencias de variables categóricas.

Importante para ver si alguna categoría domina nuestro conjunto de datos y tener una visión clara de la composición de las variables.

```
info_samples %>%
  ##Esto crea una lista donde cada elemento es la tabla
  map(~ table(.))
```

```
## $SUBJECTS
## .
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
##  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
## 27 28 29 30 31 32 33 34 35 36 37 38 39
##  1  1  1  1  1  1  1  1  1  1  1  1  1
##
## $SURGERY
```

```
## .
## by pass tubular
##      26      13
##
## $AGE
## .
## 19 24 26 27 29 33 35 37 38 39 40 41 42 44 45 46 47 55 56 57 58 59
##  1  1  1  2  1  3  2  3  1  3  1  3  3  1  1  4  1  3  1  1  1  1
##
## $GENDER
## .
##  F  M
## 27 12
##
## $Group
## .
##  1  2
## 24 15
```

Podemos además visualizar la frecuencia relativa porcentual.

```
info_samples[c(2:5)] %>%
    map(~ prop.table(table(.))*100)
```

```
## $SURGERY
## .
##  by pass  tubular
## 66.66667 33.33333
##
## $AGE
## .
##        19        24        26        27        29        33        35        37
##  2.564103  2.564103  2.564103  5.128205  2.564103  7.692308  5.128205  7.692308
##        38        39        40        41        42        44        45        46
##  2.564103  7.692308  2.564103  7.692308  7.692308  2.564103  2.564103 10.256410
##        47        55        56        57        58        59
##  2.564103  7.692308  2.564103  2.564103  2.564103  2.564103
##
## $GENDER
## .
##        F        M
## 69.23077 30.76923
##
## $Group
## .
##        1        2
## 61.53846 38.46154
```

4.1.2 Un resumen estadístico de las variables numéricas:

```
##Como tenemos tantas variables, para mostrar el output en el documento,
##se seleccionan 5 variables
summary(datos[6:11])
```

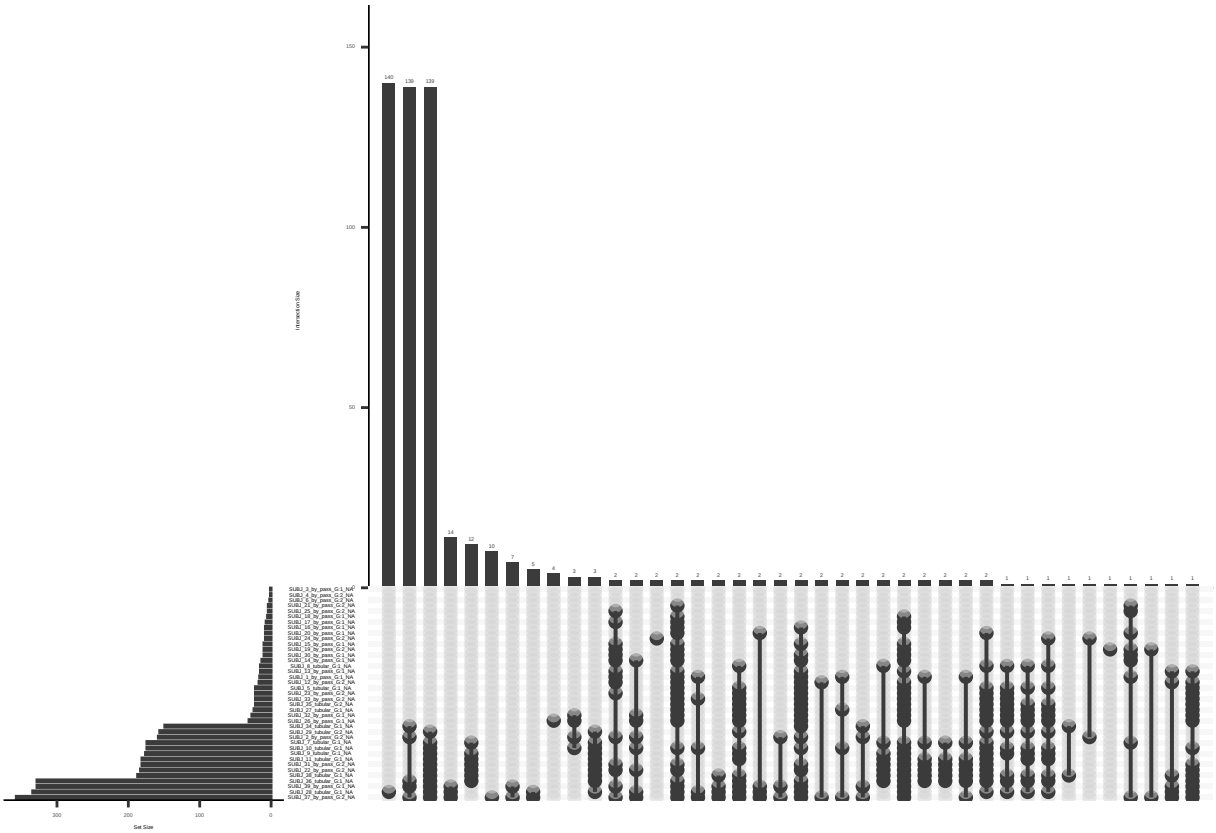
```
##      INS_T0          HOMA_T0          HBA1C_T0     HBA1C.mmol.mol_T0
```

```
##  Min.   : 3.40   Min.   : 0.760   Min.   :5.100   Min.   :32.23
##  1st Qu.:11.85   1st Qu.: 2.435   1st Qu.:5.400   1st Qu.:35.51
##  Median :16.00   Median : 4.210   Median :5.600   Median :37.69
##  Mean   :17.60   Mean   : 4.890   Mean   :5.592   Mean   :37.60
##  3rd Qu.:21.15   3rd Qu.: 5.720   3rd Qu.:5.800   3rd Qu.:39.88
##  Max.   :43.00   Max.   :13.600   Max.   :6.400   Max.   :46.44
##                                   NA's   :15      NA's   :15
##     PESO_T0          bmi_T0
##  Min.   : 84.0   Min.   :29.80
##  1st Qu.:119.5   1st Qu.:44.40
##  Median :135.0   Median :48.80
##  Mean   :140.0   Mean   :50.52
##  3rd Qu.:155.0   3rd Qu.:55.35
##  Max.   :200.0   Max.   :68.60
##
```

De aquí podemos observar que los datos están bastante asimétricos, por lo que la mejor opción es aplicar logaritmos para volverlos simetricos, pero antes de esto debemos analizar los valores faltantes.

4.1.3 Visualización de valores faltantes.

Los datasets a veces tienen valores `NA`, estos son fáciles de visualizar:

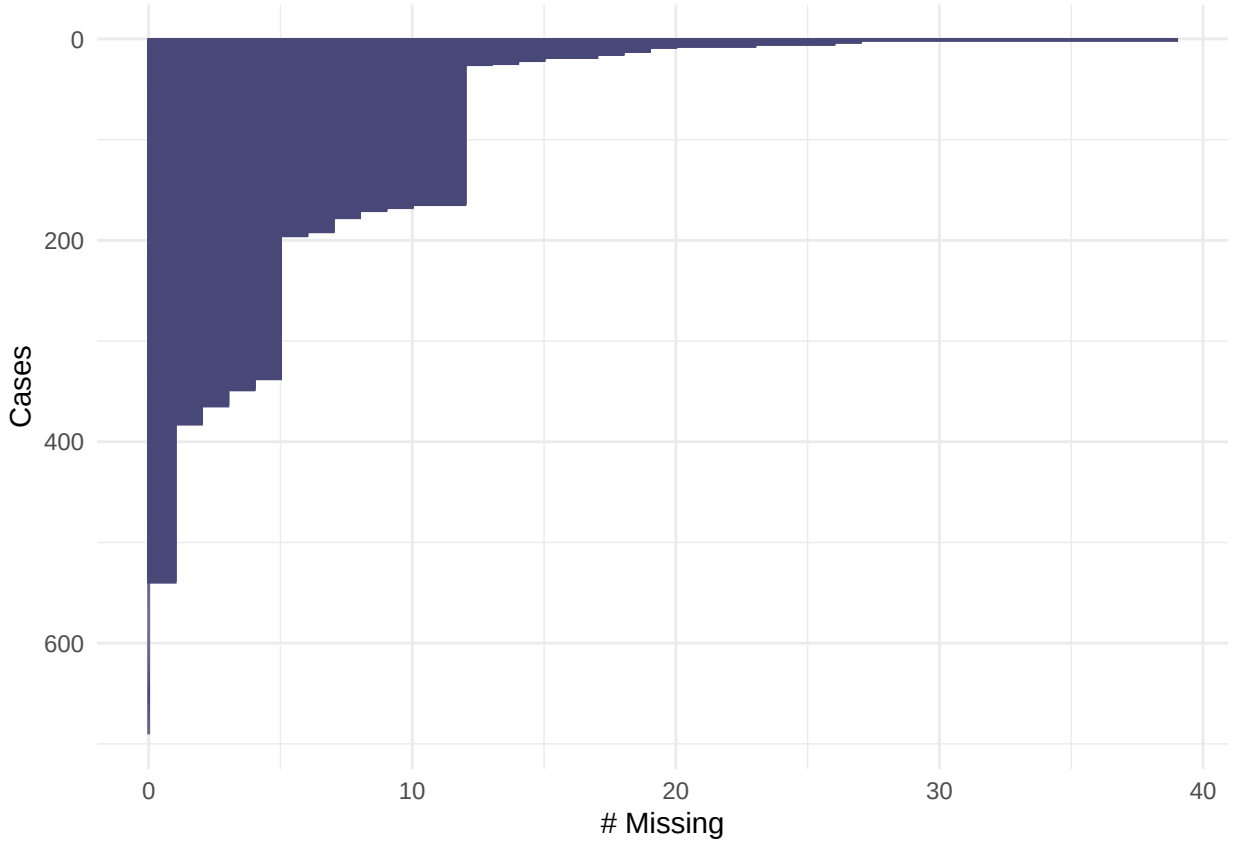


Podemos observar las sujetos que contienen datos faltantes, además de los casos en los que estos tienen datos faltantes en común (interaciones).

Si quisieramos saber con exactitud cuando sujetos tienen datos faltantes:

```
## [1] 39
```

Además también es fácil visualizar el número de valores faltantes en cada caso:



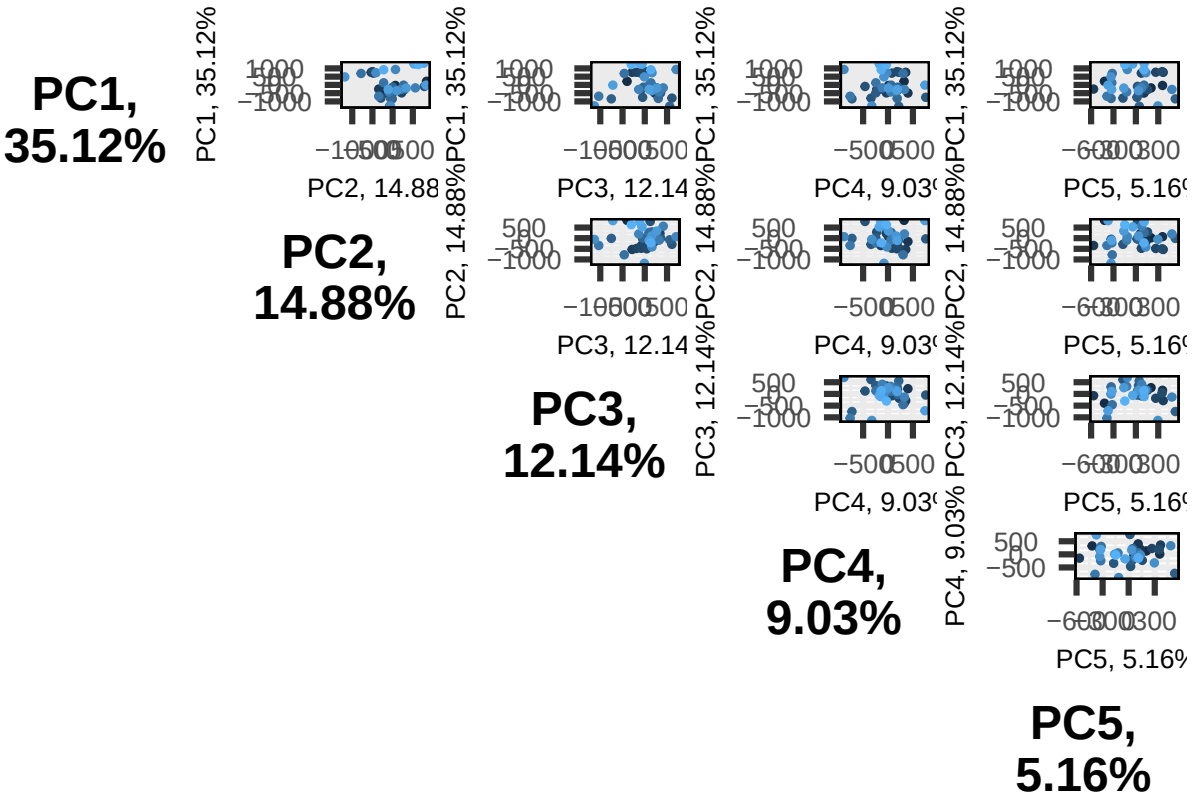
Como todos tienen en algún momento valores faltantes, la opción de quedarnos solo con aquellos que estén completos se descarta. Por tanto entonces lo que hacemos es **designar 1 a los valores faltantes**.

Una vez hemos tratados los valores faltantes, podemos aplicar logaritmos para su correcta normalización y guardar nuestros datos en formato binario.

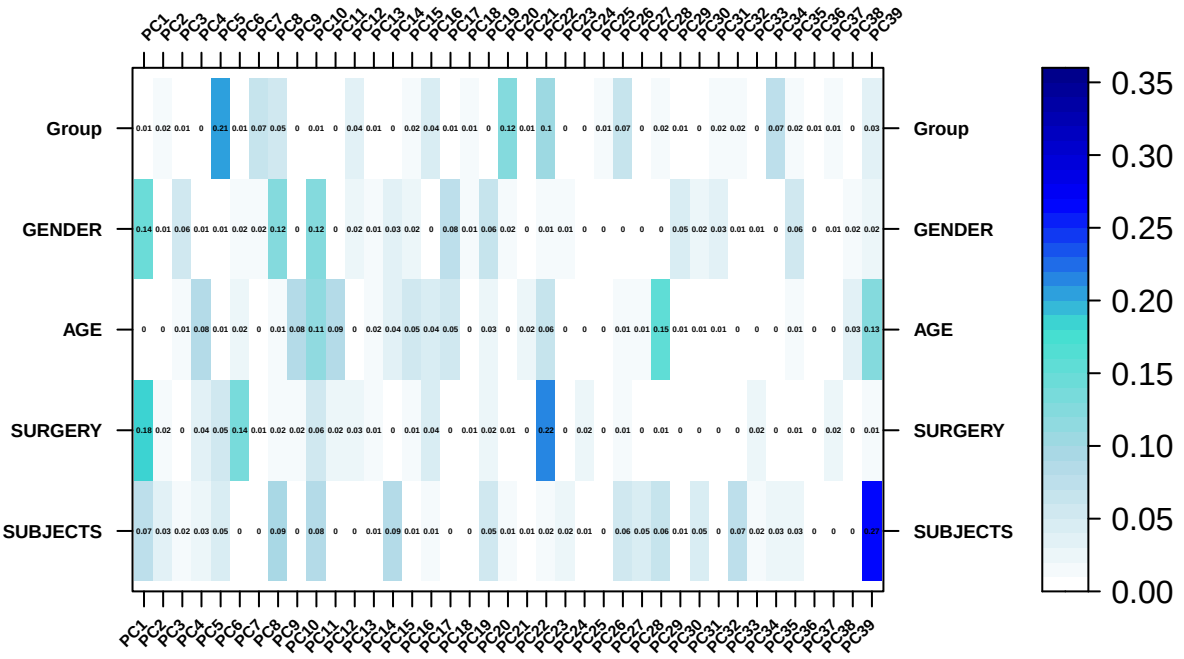
4.1.4 Exploración gráfica.

4.1.4.1 Análisis de Componentes Principales (PCA). Nos permite reducir la dimensionalidad de nuestros sin perder información clave, de forma que captura la mayor parte de la variabilidad y permitiéndonos ver patrones globales.

Para ver las relaciones entre los diferentes componentes principales. Cada gráfico representa la correlación entre dos componentes principales.

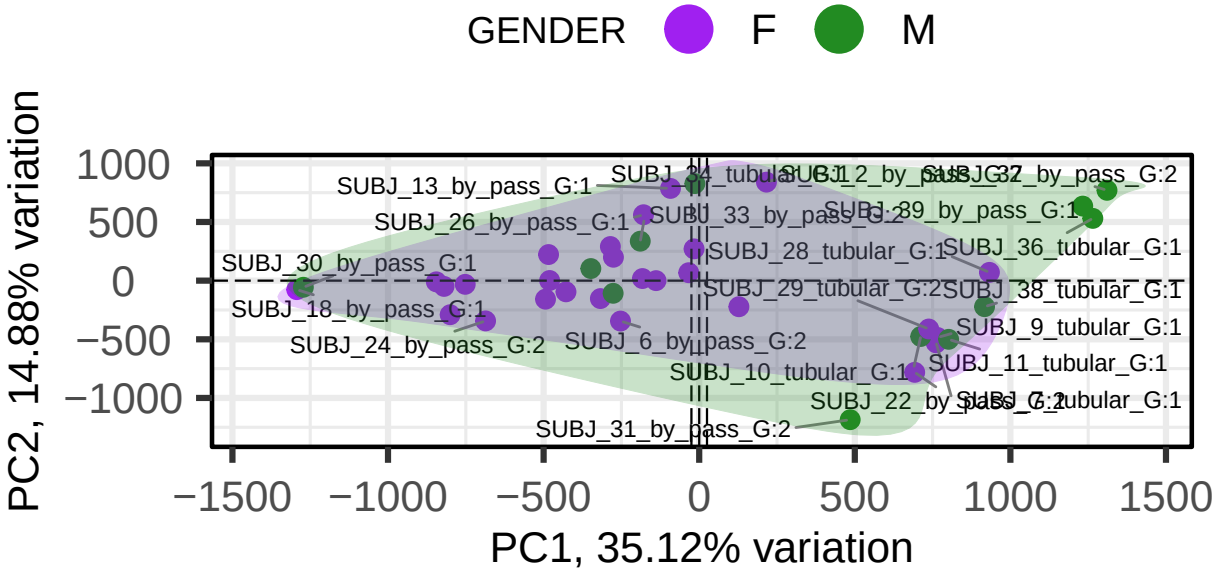


Se genera también un gráfico de correlación de los componentes principales con las metavariables de las muestras. El gráfico mostrará entonces la relación entre cada componente y las variables de nuestras metavariables. Se calcula la correlación de Pearson, con corrección por pruebas múltiples (método de *Benjamini-Hochberg*).



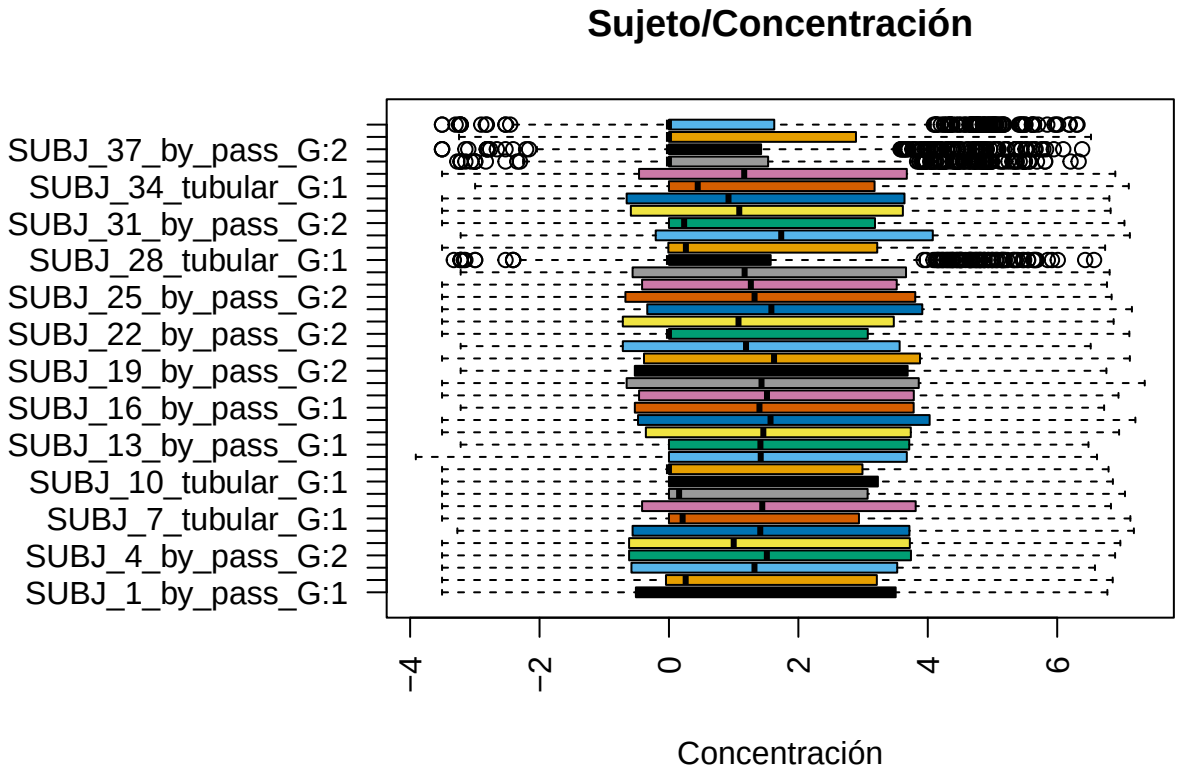
Los números de las celdas representan los valores r^2, indicando cuanta varianza explica cada componente principal en relación con factores Grupo, Género, Edad, Cirugía y Sujetos.

También con un gráfico biplot se visualizan las muestras y las variables en el espacio de los componentes principales. Nuestros sujetos se colorean en función del sexp pero podría hacer con cualquier otra variable (surgery, group, etc). En caso de que hubiera una clara separación entre grupos, significaría que los perfiles metabólicos son diferentes según tipo de cirugía.

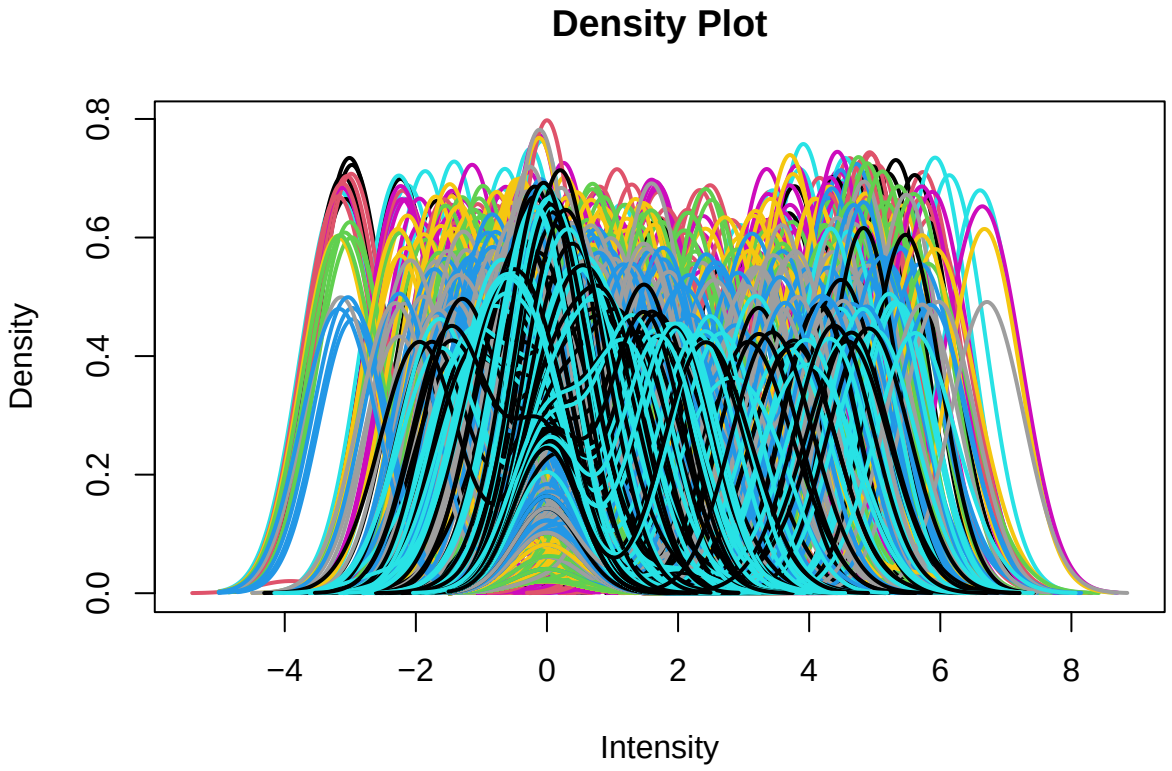


Los dos componentes principales (PC1 y PC2) tienen aproximadamente casi el 50% de la varianza total de nuestros datos. Como nuestro clustering fue basado en el sexo, esto nos indicaría que esto es un factor importante que está influyendo en la variación de nuestro perfil metabólico.

4.1.4.2 Boxplot y Gráfico de densidad. Generamos un boxplot de las concentraciones en cada sujeto, permitiendo evaluar la distribución de los datos en nuestro objeto `sum_ex` y un gráfico de densidad para visualizar la distribución de los metabolitos. Esperamos que si nuestra transformación logarítmica fue efectiva, la distribución sea más simétrica ya que picos múltiples podrían sugerir grupos de muestras con perfiles metabólicos distintos.



Los outliers de nuestro diagrama de caja representan valores atípicos, lo que sugiere que hay sujetos con metabolitos que presentan concentraciones fuera del rango típico, quizás influido por los valores faltantes en esos sujetos.



También vemos que nuestra normalización fue efectiva, en el gráfico de densidad hay una gran superposición

de curvas, es decir distribuciones similares (esto también se ve en el diagrama de cajas). Sin embargo, el hecho de haber múltiples picos puede que nos indique subgrupo de metabolitos.

5 Discusión.

En este estudio se ha abordado la exploración de metabolitos asociados a la respuesta a la cirugía bariátrica. Una de las principales limitaciones del estudio es la cantidad de los datos disponibles. Aunque el conjunto de datos seleccionado incluye información relevante sobre las muestras, la presencia de valores faltantes puede haber influido en los resultados obtenidos. Si bien se aplicaron técnicas para manejar los valores faltantes, la imputación de valores que hicimos para tratar esos valores pueden introducir sesgos que afectan la validez de nuestros hallazgos.

Además, la representación de las muestras en el conjunto de datos podría no ser completamente representativa de la población general ya que únicamente se disponian de 39 sujetos y además mayoritariamente mujeres. No fueron únicamente esas las limitacdiones, ya que los datos estaban en general desvalanceados (más cantidad de sujetos con un tipo de cirugía en concreto o más sujetos del grupo 1 que del 2 entre otros). Todo esto va a limitar la capacidad de generalizar cualquier resultado obtenido.

En cuanto al análisis realizado, aunque se emplearon métodos robustos de exploración de datos como el análisis de componentes principales (PCA), que permitió reducir la dimensionalidad y observar patrones en los datos, también existe el riesgo de sobreinterpretar los resultados si no se validan con un conjunto de datos independiente.

6 Conclusión.

El análisis de los metabolitos asociados a la respuesta a la cirugía bariátrica ha proporcionado información sobre los posibles patrones metabólicos que podrían influir en el éxito de la intervención. A través de técnicas de análisis exploratorio, como el análisis de componentes principales, la visualización de correlaciones y otras representaciones gráficas, se ha logrado identificar características clave en los datos que podrían ser indicativas de diferencias biológicas en la respuesta a la cirugía.

7 Referencias.

Repositorio del dataset: nutrimetabolomics

Palau-Rodriguez, M., Tulipani, S., Marco-Ramell, A., Miñarro, A., Jáuregui, O., Sanchez-Pla, A., Ramos-Molina, B., Tinahones, F. J., & Andres-Lacueva, C. (2018). Metabotypes of response to bariatric surgery independent of the magnitude of weight loss. *PLoS ONE*, 13(6), e0198214. *https://doi.org/10.1371/journal.pone.0198214*

URL al repositorio creado: https://github.com/MTE23/PEC1_MARIA_TORES_ESPANA.git