Análisis de Datos Ómicos – PEC 1

Master de Bioinformática y Bioestadística de la UOC
Diana Campos López

Índice

1. Objetivos 2
2. Métodos 2
3. Resultados 4
4. Discusión 6
5. Conclusiones 7
6. Aclaraciones 8
7. Referencias 8

Resumen

Este trabajo tiene como objetivo explorar y analizar los datos de un experimento metabolómico utilizando el objeto SummarizedExperiment (SE) en R. El estudio seleccionado investiga la relación entre los metabolitos de la microbiota intestinal y los cambios neuroendocrinos en la transición a la menopausia. Se procesaron los datos de 12 muestras de ratonas, con metabolitos obtenidos por cromatografía líquida acoplada a espectrometría de masas (LC-MS). Se realizaron análisis de normalización, componentes principales (PCA) y mapas de calor. Los resultados mostraron una separación entre muestras según la edad de las ratonas, aunque la falta de datos completos impidió una interpretación más detallada.

1. Objetivos

El objetivo principal de este trabajo es explorar y analizar los datos de un experimento seleccionado en una base de datos de metabolómica utilizando para ello un objeto SummarizedExperiment (SE). Este objetivo puede desglosarse en:

a) Importación de los datos y creación del objeto SE
b) Análisis de los datos y procesamiento de los datos
c) Interpretación biológica de los resultados obtenidos

2. Métodos

Para realizar este trabajo he buscado en el repositorio Metabolomics Workbench la palabra clave “microbiota” debido a que este tema me resulta altamente interesante y útil debido a la compleja relación que ocurre entre estas bacterias, hongos y virus con el resto del organismo en el que se alojan. A continuación, para seleccionar uno de los experimentos me fijé en el número de muestras que se habían analizado con el fin de obtener un número suficientemente alto como para que el estudio resulte interesante pero no tan alto como para que me suponga problemas por la capacidad de mi ordenador.

Por estas razones he seleccionado el estudio ST003003, llamado “Gut microbiota and metabolites in estrus cycle and their changes in a menopausal transition rat model with typical neuroendocrine aging” (Dai et al., 2023). El objetivo de este proyecto es estudiar la relación entre los metabolitos de la microbiota para asociarla con los cambios en el hipotálamo relacionados con la transición a la menopausia. Este tema me parece de vital importancia dado que la microbiota intestinal desempeña un papel clave en la modulación del eje intestino-cerebro y su impacto en la menopausia no ha sido muy estudiado.

Este experimento se ha realizado sobre 27 ratonas divididas en 3 grupos:

- QC: Controles
- Y: ratonas jóvenes, entre 2 y 3 meses de vida
- MA: ratones de edad media, entre 9 y 10 meses de vida

A su vez, estos podían dividirse en proestrus (fase previa al celo) y diestrus (fase posterior al celo) según en la fase del ciclo estral en la que se encuentran. Estas fases son relevantes porque la fluctuación hormonal podría influir en los metabolitos presentes en la microbiota intestinal. Se realizó un análisis de metabolitos con una cromatografía líquida acoplada a un espectrómetro de masas (LC-MS).

Para trabajar con estos datos el primer paso ha consistido en importarlos al entorno de R y unir los archivos “ST003003_AN004933_results.txt” y “ST003003_AN004934_results.txt” correspondientes con los resultados de la LC-MS en modo positivo y negativo respectivamente. Estos archivos están

compuestos por 13 columnas, 12 correspondientes a las distintas muestras de ratonas y una correspondiente al nombre del metabolito que se ha analizado. Este formato corresponde con el que se necesita para crear el SE como podemos ver en la Figura 1. SE puede considerarse una versión más versátil de ExpressionSet que permite analizar estudios con características más específicas y diversas. Según la documentación de Bioconductor, ambos objetos son similares en estructura y uso, pero la principal diferencia es que SummarizedExperiment ofrece mayor flexibilidad en la información de las filas, permitiendo integrar objetos *GRanges* que contienen la localización genómica. Esto lo hace especialmente útil para estudios basados en secuenciación como RNA-seq y ChIP-seq, mientras que ExpressionSet no posee esta capacidad.

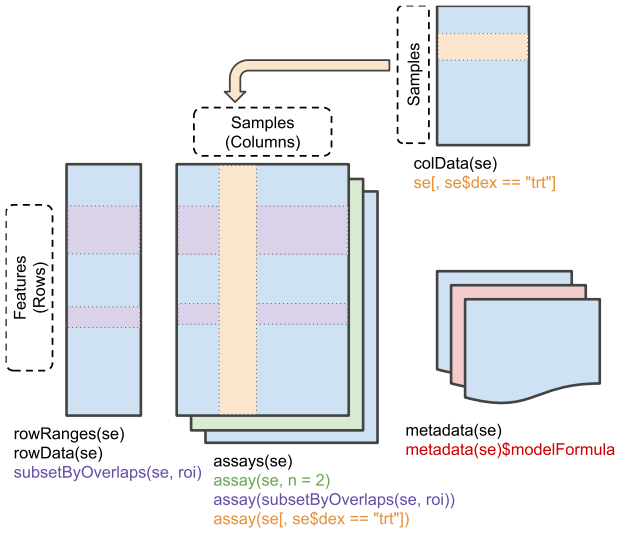

Figura 1: Estructura de un objeto SummarizedExperiment y el código necesario para poder acceder a las distintas partes cuando este objeto se ha guardado con el nombre se. Imagen obtenida de (*SummarizedExperiment for Coordinating Experimental Assays, Samples, and Regions of Interest*, n.d.)

A continuación, se creó el SummarizedExperiment usando la suma de estos documentos como assay, y usando los documentos “ST003003_AN004933.txt” y “ST003003_AN004934.txt” por separados como metadatos. Estos últimos son archivos descargables directamente desde el repositorio que contienen información general acerca del experimento, tanto para el modo positivo como para el modo negativo del análisis. El SummarizedExperiment obtenido puede ser descargado en *Didi-491/PEC1-ADO*, n.d.

A continuación se han realizado diversos estudios sobre este experimento, cuyo código puede ser visualizado en el archivo .rmd de *Didi-491/PEC1-ADO*, n.d. Primero se han utilizado diversos comandos para analizar el tamaño de nuestros datos. En este momento he podido ver que no había 27 muestras en los archivos descargables, sino que solo estaban disponibles las ratonas con en la fase proestrus faltando los datos de los controles y de las ratonas en fase diestrus. No he podido averiguar porque el resto no están subidos al repositorio. Aun así, he seguido trabajando con estos datos.

Para analizar la expresión en los metabolitos en las distintas ratonas se han realizado diversas gráficas que permiten analizar los resultados visualmente. En el proceso he observado que los datos requerían ser normalizados para poder trabajar con ellos, por lo que se ha aplicado una función logarítmica que ha reducido la amplitud de los valores atípicos. Posteriormente se ha realizado un PCA tanto de las expresiones crudas como normalizadas y se realizó un mapa de calor utilizando el método de clustering Ward.D y el método de correlación de Pearson para el cálculo de distancias. He probado otras combinaciones de métodos, pero está ha sido la más eficaz, aunque ninguna presentó una agrupación especialmente diferente debido a la clara separación de los datos en todas ellas, como se comentará más adelante.

Las librerías requeridas para este trabajo han sido: SummarizedExperiment, readr, dplyr, tidyr, ggplot2, DESeq2, pheatmap, FactoMineR y factoextra.

3. Resultados

El objeto SE permite almacenar los datos en un único objeto con tres componentes principales:

- **Assay**: Matriz de expresión de metabolitos.
- **ColData**: Información sobre las muestras
- **RowData**: Información sobre los metabolitos analizados.

Las dimensiones de nuestro experimento se pueden observar con la función dim(se), cuyo resultado ha sido 729 filas (metabolitos) y 12 colunas (muestras). Con el comando colnames(se) podemos ver el nombre de las muestras, que son: MA1, MA4, MA6, MA7, MA13, MA26, MA34, Y5, Y7, Y8, Y9 y Y13. Como se ha explicado anteriormente faltan los datos de los controles y de las ratonas en fase diestrus, lo que va a impedir inferir conclusiones sobre la relación de los metabolitos con las distintas fases del ciclo.

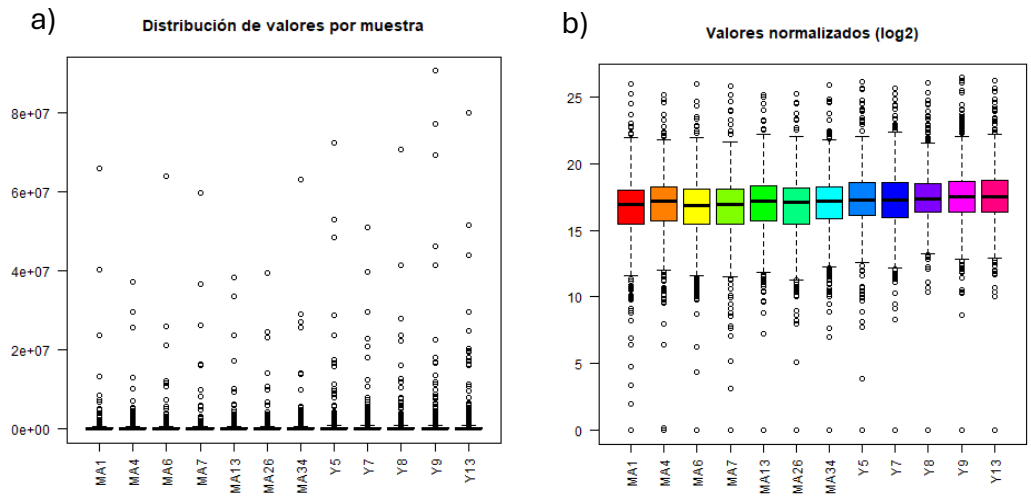

Figura 2: a) Boxplot de la distribución de los datos crudos por muestra. b) Boxplot de la distribución de los datos tras normalizar por muestra. Imagen de elaboración propia.

La figura 2 representa un gráfico de cajas sobre la expresión de los metabolitos en las distintas muestras. Como podemos ver en la figura 2.a, los datos sin normalizar presentan una gran dispersión y un gran número de valores atípicos en la parte superior. Para poder realizar una comparación más acertada de los datos se ha realizado una normalización de los datos que se puede ver en la figura 2.b. Podemos comprobar que la normalización ha sido efectiva ya que la escala del eje Y ha disminuido notablemente, indicando que la transformación ha reducido la amplitud de los valores extremos. Podemos observar que la media de metabolitos no varía entre las ratonas independientemente de su edad.

Para complementar este análisis se ha realizado un análisis de componentes principales (PCA) danto de los datos crudos como de los datos normalizados (Figura 3). Podemos ver que los componentes principales (PC) solo recogen aproximadamente un 70% de los metabolitos. Además, podemos observar que en ambos PCA presentan una separación de las muestras según la edad. Esta dispersión podría estar relacionada con un efecto batch, es decir, diferencias técnicas introducidas por el procesamiento de las muestras en distintos momentos, aunque en el artículo no se cuenta con información detallada sobre la prevención del efecto batch de manera explícita. Esta diferencia también puede deberse simplemente por una diferencia en composición de metabolitos relacionada con la edad. Además, dentro de cada grupo existe una dispersión en el eje vertical que indica una diferencia dentro del grupo entre los componentes principales del grupo 2.

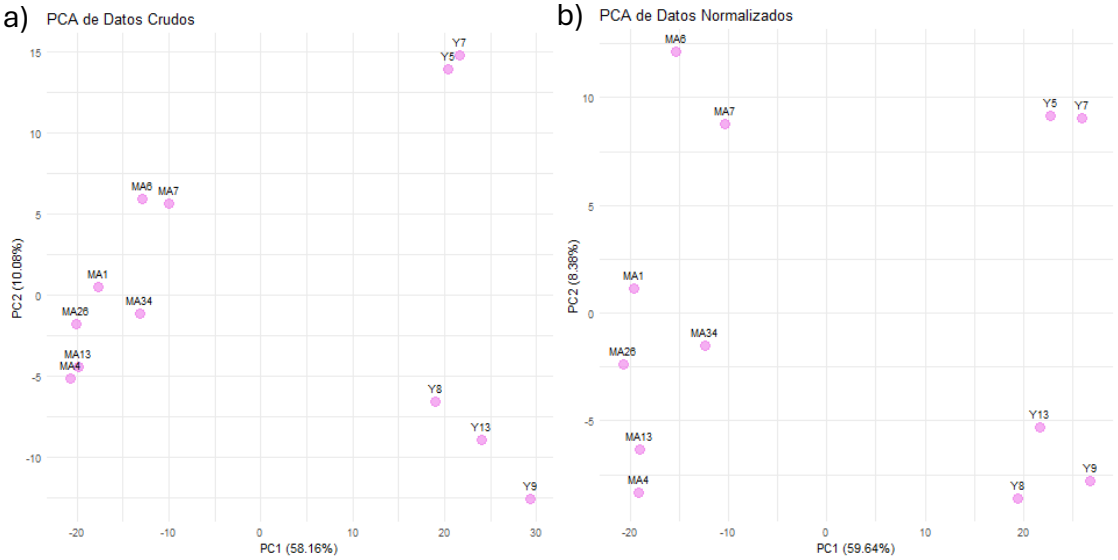

Figura 3: a) PCA de los datos crudos. b) PCA de los datos normalizados. Imagen de elaboración propia.

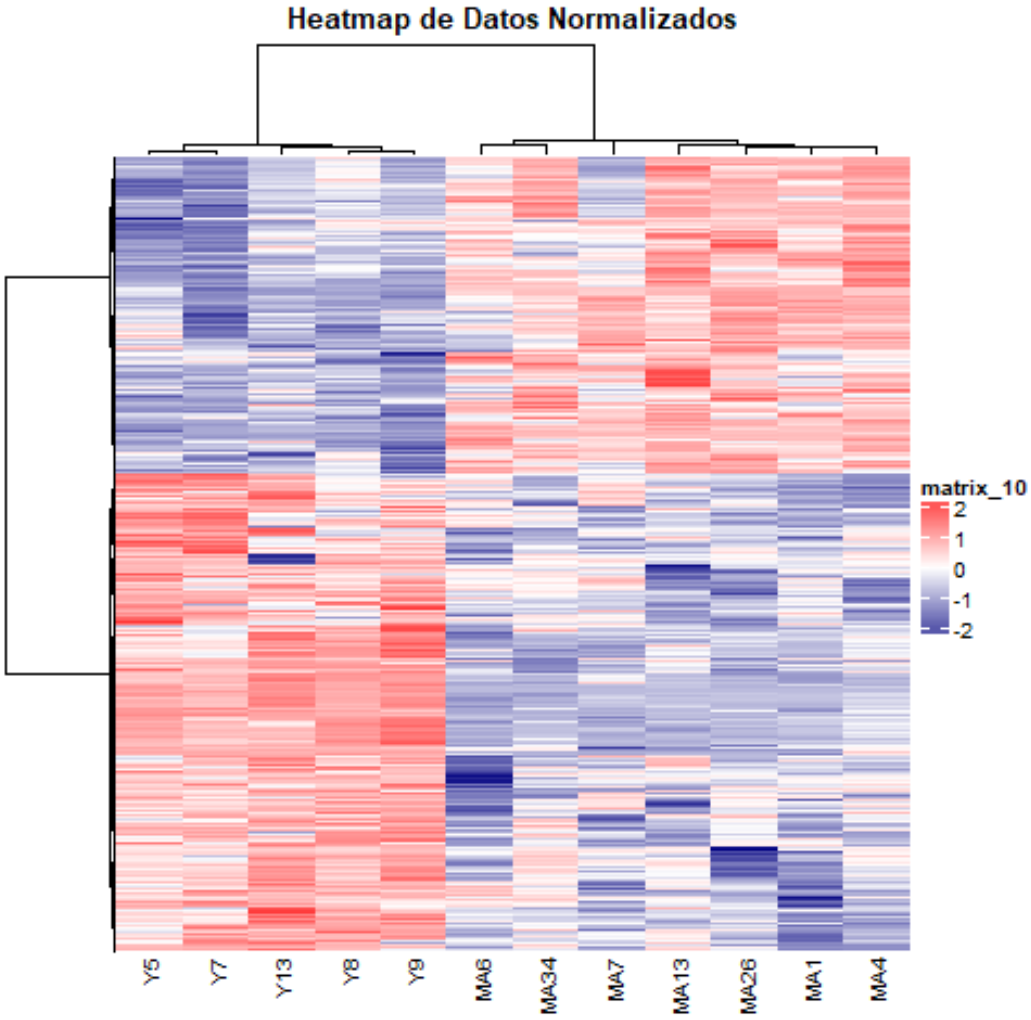

Figura 4: Mapa de calor sobre la expresión de los datos normalizados por paciente. Imagen de elaboración propia.

Por último, se realizó un mapa de calor sobre los datos normalizados, en el que se puede observar un patrón claro sobre la expresión de los metabolitos según la edad de los ratones. Aún así encontramos ciertas diferencias dentro de cada grupo. Por ejemplo, la muestra Y13 presenta un metabolito infra expresado en comparación a todas las demás muestras de su grupo. Esto también ocurre con la muestra MA7.

4. Discusión

El análisis de los datos de metabolómica en ratonas jóvenes y de mediana edad ha permitido observar patrones diferenciados en la expresión de metabolitos. Sin embargo, la falta de datos para los grupos de control y las ratonas en fase diestrus limita la posibilidad de extraer conclusiones sobre la influencia del ciclo estral en la composición metabólica de la microbiota intestinal.

El PCA mostró una clara separación entre los grupos de edad en el eje principal (PC1), lo que sugiere diferencias metabólicas asociadas al envejecimiento. No obstante, dentro de cada grupo se observa una dispersión en el eje PC2, lo que

indica que puede haber otros factores biológicos o técnicos influyendo en la variabilidad de los datos. Además, no se cuenta con información sobre el control del efecto batch, lo que podría estar introduciendo sesgos en la agrupación de las muestras.

Otro punto relevante es que, tras la normalización, la dispersión de los datos se redujo, lo que sugiere que la variabilidad observada en los datos crudos estaba en parte influenciada por diferencias en la magnitud de los valores y no necesariamente por diferencias biológicas reales. Esto subraya la importancia de aplicar métodos de normalización adecuados en estudios de metabolómica.

Para mejorar la interpretación biológica de los resultados, habría sido ideal contar con anotaciones metabólicas detalladas, ya que, el dataset provee la masa a la que el MS encontró el metabolito, pero no su nombre. Sin esto no se pueden hacer estudios sobre la agrupación de los metabolitos según sus propiedades físico-químicas ni realizar análisis de enriquecimiento que permitan relacionar estos metabolitos con rutas bioquímicas.

Aún así, se ha demostrado que la edad esta directamente relacionada con la presencia de ciertos metabolitos, lo que podría sugerir una alteración en el microbiota intestinal influenciado por el cambio hormonal relacionado con la edad.

5. Conclusiones

El objeto SummarizedExperiment ha demostrado ser una herramienta valiosa para el análisis de datos bioinformáticos de estudios propios o ajenos con gran versatilidad y capacidad para trabajar en múltiples dimensiones de datos. Considero que en el trabajo no se ha podido ver reflejado el todo el potencial de este tipo de objetos, que reside en su capacidad de almacenar distintos tipos de análisis siempre que estos tengan las mismas dimensiones, lo que permitiría acceder a todos estos análisis a través del mismo objeto y manejarlos simultáneamente. Por ejemplo, la eliminación de una muestra de los datos resultaría en su eliminación automática de todos los análisis relacionados, lo que simplifica la gestión de los datos a gran escala.

Aun así, las dos primeras partes del objetivo han sido cumplidas con el objeto SE: la importación de los datos, la creación del objeto SE y el análisis preliminar de los datos. Además, he podido realizar una interpretación biológica acorde a los datos, aunque incompleta debido a la falta de muestras y de una librería que permita acceder al nombre de los metabolitos en lugar de a su relación masa carga (m/z) obtenida por el espectrómetro de masas.

A pesar de estos obstáculos, el trabajo ha logrado proporcionar una comprensión más clara del análisis multivariante tanto a nivel biológico como computacional. He adquirido nuevas habilidades que considero muy valiosas para mi carrera profesional y que sin duda serán de gran utilidad en mi TFM.

6. Aclaraciones

El trabajo ha sido realizado en formato Word para facilitar la inclusión de las referencias y debido a que el código utilizado se visualiza de manera más cómoda en GitHub. Además, en el repositorio de GitHub se encuentra un archivo .html que combina el texto con el código utilizado. También se incluye el archivo .rmd, el cual puede ser utilizado para replicar los análisis. El enlace para acceder al repositorio de GitHub es el siguiente:

https://github.com/didi-491/Campos_L-pez_Diana_PEC1

7. Referencias

Dai, R., Huang, J., Cui, L., Sun, R., Qiu, X., Wang, Y., & Sun, Y. (2023). Gut microbiota and metabolites in estrus cycle and their changes in a menopausal transition rat model with typical neuroendocrine aging. *Frontiers in Endocrinology*, *14*, 1282694. https://doi.org/10.3389/fendo.2023.1282694

Didi-491/PEC1-ADO. (n.d.). GitHub. Retrieved 2 April 2025, from https://github.com/didi-491/PEC1-ADO

SummarizedExperiment for Coordinating Experimental Assays, Samples, and Regions of Interest. (n.d.). Retrieved 2 April 2025, from https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html