Alumna: Consuelo Marín-Vicente

Asignatura: Análisis de datos ómicos (ADO)

PEC 1

1. **Descarga de datos para el análisis:**
   He navegado en la página Metabolomics Workbench y he extraído los datos del dataset ST000567. Este dataset incluye un archivo disponible para ser descargado:

- ST000567_AN000871_Results.txt **(3.7M)**

Resumen del proyecto seleccionado:

La edad es uno de los mayores factor de riesgo en enfermedades cardiovasculares. Dos parámetros que contribuyen a este riesgo son la rigidez de las arterias y el desarrollo de disfunciones endoteliales que se indican mediante EDD (impaired nitric oxide induced endothelium dependent dilation. Disfunción en la dilatación del endotelio causada por óxido nítrico). La reducción de la rigidez en arterias y el incremento de la función endotelial vascular mediante el incremento en la disponibilidad del NO reducen el riesgo de enfermedad. Estudios preliminares han mostrado que el uso de curcumina en la dieta ayuda a reducir dicho riesgo. El objetivo del proyecto seleccionado es el estudio del metaboloma plasmídico tras la administración oral de curcumina.

Al inspeccionar el archivo veo que los metadatos están incluídos en el mismo archivo descargado. En este archivo se incluyen en las columnas los metabolitos identificados con la masa del precursor analizado así como la anotación al metabolito concreto si es el caso.

En las primera y segunda filas se incluyen los nombres de las muestras y la descripción de cada una de ellas, respectivamente.

Me quedo con la información de la primera fila en el archivo a estudiar, y la información de la primera columna. Igualmente, extraigo los metadatos en archivo adicional (dos primeras filas y la primera columna: metadata.xlsx).

2. **Creación de objeto de clase SummarizedExperiment:**
   Ahora que tenemos los datos de expresión en el archivo *data_matrix*, podemos proceder a crear el objeto SummarizedExperiment (ver R script: MARINVICENTE_CONSUELO_ADO_PEC1_3.R y en archivo MARIN_VICENTE_CONSUELO_ADO_PEC1_RCODE_AND_RESULTS.pdf)

Diferencias entre ExpressionSEt y SummarizedExperiment: A resumir

- **ExpressionSet**: La clase ExpressionSet es parte del paquete **Biobase** de Bioconductor. Es una clase de datos más antigua que se utiliza principalmente para almacenar datos de expresión génica, como las matrices de expresión, junto con metadatos de las filas (genes) y las columnas (muestras). Se asocia principalmente con datos de arrays y algunos tipos de datos de expresión. En cuanto a su estructura, los datos de expresión

se almacenan principalmente en una matriz de expresión (gen x muestra). Los metadatos relacionados con las filas (genes) y las columnas (muestras) se almacenan como objetos AnnotatedDataFrame. Es relativamente rígido en su estructura y no soporta de manera eficiente múltiples tipos de datos dentro del mismo objeto. En cuanto a compatibilidad con nuevos paquetes, aunque muchos paquetes antiguos y enfoques de análisis se basan en ExpressionSet, está un poco más limitado en comparación con el estándar actual de clases en Bioconductor.

- **SummarizedExperiment**: La clase SummarizedExperiment, que pertenece al paquete **SummarizedExperiment**, es más moderna y flexible que ExpressionSet. Está diseñada para ser más general y se utiliza para almacenar y manejar datos de cualquier tipo de experimento omico. Es una clase más versátil y extensible que permite manejar datos más complejos. Tiene una estructura más flexible: el objeto puede almacenar **varios "assays"** (por ejemplo, matrices de expresión para diferentes condiciones o tecnologías) dentro de una lista. Además, los metadatos de las filas y las columnas se almacenan en **rowData** y **colData** como objetos DataFrame. Esto permite una mayor flexibilidad en la gestión de diferentes tipos de datos y metadatos. Está diseñado para ser la clase de elección en nuevos desarrollos en Bioconductor.

3. **Análisis exploratorio:**

Un análisis exploratorio de datos (EDA) normalmente incluye varios pasos, como la inspección de las dimensiones del conjunto de datos, la visualización de las distribuciones de las variables, la evaluación de la calidad de los datos y la detección de posibles outliers. Esto lo vamos a hacer con funciones de exploración sobre el objeto creado tal cual se muestra en el R script (ver R script: MARINVICENTE_CONSUELO_ADO_PEC1_3.R y en archivo MARIN_VICENTE_CONSUELO_ADO_PEC1_RCODE_AND_RESULTS.pdf)

Los box plots nos muestran una similitud en la distribución de la señal por muestra. La distribución de las muestras en base a esta señal deja ver en el PCA una acumulación de las muestras en un cluster común mientras que hay algunas muestras que son outliers de este agrupamiento. El análisis de variabilidad no indica cuáles son los metabolitos más abundantes a lo largo del ensayo.

4. Generación de repositorio de github con datos requeridos:

https://github.com/cmarinvic/MARIN_VICENTE_CONSUELO_PEC1_1