Análisis de datos Ómicos PEC1

Ángel I. Pérez Santiago.

A continuación, se presenta el informe final de la PEC1, en el que se describe de forma integrada el proceso de análisis de datos metabolómicos utilizando el dataset "2018-MetabotypingPaper". Se incluye la selección y preparación del dataset, la creación del objeto SummarizedExperiment, el tratamiento de valores faltantes, la transformación de los datos y los análisis exploratorios (boxplots, clustering, mapas de calor y análisis de componentes principales). El informe sigue las directrices establecidas en los documentos de la PEC.

Tabla de Contenidos

1. Abstract

Este estudio presenta el análisis de un dataset de metabolómica obtenido del repositorio "2018-MetabotypingPaper", seleccionado por su calidad y representatividad. Se realizó la creación de un objeto SummarizedExperiment que integra tanto los datos metabolómicos (690 variables) como los metadatos clínicos (39 muestras). Se abordó el tratamiento de valores faltantes, identificados como NA, -9 y -99, reemplazándolos de forma uniforme por 1 para permitir el análisis posterior. Dado el amplio rango de valores y la presencia de valores negativos, se aplicó una transformación asinh, la cual permite estabilizar la varianza sin distorsionar la estructura relativa de los datos. Posteriormente, se realizaron análisis exploratorios mediante boxplots, clustering jerárquico (con dendrograma coloreado) y mapas de calor para visualizar la agrupación de las muestras. Finalmente, se efectuó un análisis de componentes principales (PCA) con screeplot y biplot, que reveló diferencias en los perfiles metabolómicos relacionadas con la variable "SURGERY". Los resultados aportan información clave para la interpretación biológica del estudio.

2. Objetivos

- Seleccionar y descargar un dataset de metabolómica representativo.
- Crear un objeto SummarizedExperiment que contenga los datos y metadatos del estudio.
- Realizar el preprocesamiento: tratamiento de valores faltantes y transformación de los datos.
- Explorar la estructura y agrupación de las muestras mediante visualizaciones (boxplots, dendrograma, mapa de calor y PCA).
- Interpretar los patrones identificados en el contexto biológico.

3. Métodos

3.1 Selección y descarga del dataset

Se optó por utilizar el dataset "2018-MetabotypingPaper" disponible en GitHub, dada su riqueza en información metabolómica y clínica, lo que facilita la realización de análisis multivariante.

3.2 Creación del objeto SummarizedExperiment

Se separaron las primeras 5 columnas (información de muestras) de los datos metabolómicos (690 columnas) y se crearon nombres de fila informativos a partir de las variables SUBJECTS, Group y SURGERY. Posteriormente, se construyó el objeto SummarizedExperiment integrando los datos (ensayo "valores") y los metadatos.

3.3 Tratamiento de valores faltantes

Se identificaron tres tipos de "faltantes" (NA, -9 y -99). Tras convertir los códigos -9 y -99 en NA, se imputaron los valores faltantes asignándoles el valor 1, lo que permitió continuar el análisis sin interrupciones.

3.4 Transformación de los datos

Debido a la presencia de valores negativos y ceros, se aplicó la transformación asinh sobre el ensayo "valores_na_fix", obteniéndose el ensayo "asinh_transform". Esta transformación es robusta ante valores no positivos y conserva las relaciones relativas entre los datos.

3.5 Análisis exploratorio

Visualización (Boxplots):

Se generaron boxplots para visualizar la distribución de los metabolitos (cada caja representa un metabolito) y de las muestras (trasponiendo la matriz para que cada caja represente una muestra).

Clustering y Dendrograma:

Se extrajo la matriz del ensayo "asinh_transform" (filas = muestras) y se escaló cada variable para calcular la distancia euclidiana entre muestras. Posteriormente, se aplicó clustering jerárquico con el método Ward y se generó un dendrograma coloreado en dos grupos.

Mapas de Calor:

Se construyeron mapas de calor de la matriz escalada (sin clustering en filas y columnas y agrupando solo por muestras) para identificar patrones en los perfiles metabolómicos.

Análisis de Componentes Principales (PCA):

Se realizó un PCA sobre la matriz transpuesta (para que las columnas sean muestras, como exige PCAtools) utilizando el ensayo "asinh_transform". Se evaluó la varianza explicada mediante un screeplot y se generaron biplots para visualizar la dispersión de las muestras, coloreadas según la variable SURGERY, además de la visualización de los loadings.

4. Resultados

El flujo de análisis permitió la creación exitosa del objeto *SummarizedExperiment* con dimensiones 39×690, en el que se integraron de forma consistente tanto los datos metabolómicos como los metadatos clínicos. A continuación, se resumen los principales hallazgos:

1. **Exploración de los datos y distribución de variables**
    - **Información de muestras:**

      El análisis descriptivo del *data.frame* derivado del *rowData* evidenció que la mayoría de las muestras presentan edades entre 19 y 59 años, con una mediana de 41 años. La distribución de los grupos en las variables

SURGERY, *GENDER* y *Group* se encontró equilibrada, con 26 muestras clasificadas como “by pass” y 13 como “tubular”; en cuanto al género, 27 son de sexo femenino y 12 de masculino.

- **Distribución de metabolitos:**

  Los boxplots de los metabolitos muestran una amplia dispersión en las concentraciones, lo que indica la presencia de diferencias significativas entre las variables. La visualización reveló que algunas variables presentan rangos de valores extremadamente amplios, lo que justificó la necesidad de realizar una transformación robusta.

2. **Tratamiento de valores faltantes y transformación de datos**

   - Se identificaron tres tipos de “faltantes”: NA, -9 y -99. Estos fueron unificados reemplazando -9 y -99 por NA y posteriormente imputando estos valores con 1. De esta forma se evitó que la presencia de datos inválidos interfiriera en el análisis.

   - Debido a la presencia de valores negativos y ceros en el dataset, se optó por aplicar la transformación *asinh*, la cual mostró ser una alternativa robusta para estabilizar la varianza sin alterar la relación entre los datos. La incorporación del ensayo *asinh_transform* en el objeto *sePEC1* permitió la visualización de los datos en una escala más homogénea, como se aprecia en los boxplots tanto de metabolitos como de muestras.

3. **Visualización y agrupación de muestras**

   - **Clustering y dendrograma:**

     El análisis de clustering jerárquico, realizado sobre la matriz escalada de datos transformados, permitió identificar dos grupos principales de muestras. El dendrograma coloreado resalta la existencia de estos dos clusters, lo que sugiere que existen diferencias sustanciales en los perfiles metabolómicos de las muestras, posiblemente asociadas a las condiciones clínicas, y aparentemente asociadas al tipo de cirugía. La

consistencia en los nombres de las muestras y la validación de la integridad de los datos fortalecen la interpretación de estos grupos.

- **Mapa de calor:**

  El mapa de calor, generado sin aplicar clustering a filas y columnas, facilitó la visualización de la matriz escalada. Este gráfico mostró patrones de intensidad en los metabolitos a lo largo de las muestras, permitiendo identificar visualmente regiones con concentraciones elevadas o reducidas. Además, otro mapa de calor, en el que se aplicó clustering únicamente en las filas (muestras), evidenció la agrupación de las mismas, corroborando los resultados del dendrograma.

4. **Análisis de Componentes Principales (PCA)**

   - La realización del PCA sobre la matriz transpuesta (donde las columnas corresponden a las muestras) reveló que las dos primeras componentes principales capturan aproximadamente el 63% de la varianza total (según el screeplot).
   - Los biplots generados muestran una clara dispersión de las muestras en el espacio definido por estas componentes, con una diferenciación parcial basada en la variable *SURGERY*. Esta separación sugiere que los perfiles metabolómicos están influenciados por el tipo de cirugía, lo que podría tener implicaciones clínicas relevantes.
   - La visualización de los loadings permitió identificar algunos metabolitos que tienen un impacto considerable en la variabilidad de las muestras. Estos metabolitos, al tener altas cargas en determinadas componentes, pueden considerarse candidatos para estudios posteriores de biomarcadores o para investigar rutas metabólicas específicas.

En conjunto, estos resultados demuestran que el proceso de preprocesamiento (tratamiento de valores faltantes y transformación asinh) fue eficaz para estabilizar la variabilidad de los datos, permitiendo así la aplicación de técnicas multivariantes que revelaron agrupaciones y patrones relevantes en el dataset. La combinación de métodos

(boxplots, dendrograma, mapas de calor y PCA) aporta una visión integral de la estructura subyacente en los datos metabolómicos y respalda la hipótesis de que existen diferencias en los perfiles según variables clínicas, especialmente *SURGERY*.

5. Discusión

El análisis realizado sobre el dataset "2018-MetabotypingPaper" evidencia la importancia de un riguroso preprocesamiento en estudios de metabolómica. La identificación y unificación de valores faltantes (NA, -9 y -99) y su posterior imputación mediante la asignación de 1 permitió evitar interrupciones en el flujo analítico, aunque reconocemos que esta estrategia puede introducir ciertos sesgos al reducir la variabilidad en los valores más bajos. La elección de la transformación asinh resultó acertada para manejar la presencia de ceros y valores negativos, logrando estabilizar la varianza sin recurrir a offsets arbitrarios, lo cual se tradujo en una distribución de los datos más homogénea y adecuada para análisis multivariantes.

Los métodos de visualización empleados (boxplots, clustering con dendrograma, mapas de calor y análisis de componentes principales) proporcionaron una visión integral de la estructura del dataset. En particular, el clustering y el mapa de calor resaltaron la existencia de dos grupos principales de muestras, mientras que el PCA –cuya primera y segunda componente explican aproximadamente el 44% y el 19% de la varianza respectivamente, sumando un 63% en conjunto– permitió observar una separación parcial en función de la variable *SURGERY*. Esto sugiere que los perfiles metabolómicos difieren significativamente según la condición quirúrgica, lo cual puede tener implicaciones en la identificación de biomarcadores y en la comprensión de las rutas metabólicas asociadas a cada tipo de cirugía.

No obstante, se deben considerar las limitaciones del estudio. La estrategia de imputación empleada, si bien funcional para el análisis exploratorio, podría beneficiarse en futuros trabajos de métodos más sofisticados que conserven la variabilidad intrínseca de los datos. Además, la presencia de posibles outliers y la amplia gama de concentraciones de metabolitos sugieren la necesidad de validar estos hallazgos en

cohortes independientes y, de ser posible, integrar información de otras plataformas ómicas para obtener una visión más completa del estado biológico.

6. Conclusiones

El presente estudio demostró la viabilidad de construir un pipeline robusto de análisis metabolómico basado en herramientas de Bioconductor. La creación del objeto *SummarizedExperiment* y el tratamiento adecuado de los valores faltantes, seguido de la transformación asinh, permitieron estabilizar la varianza de los datos y facilitar la interpretación multivariante. Los análisis de clustering, mapas de calor y PCA revelaron agrupaciones claras entre las muestras, destacando diferencias en los perfiles metabolómicos asociadas a la variable *SURGERY*. Estos hallazgos indican que la condición clínica influye en la expresión de metabolitos, lo que abre la puerta a la identificación de biomarcadores potenciales.

En conclusión, el flujo de trabajo implementado constituye una base sólida para futuros estudios en metabolómica. Se recomienda la validación externa de los resultados y la exploración de métodos alternativos de imputación y normalización que permitan afinar aún más la detección de patrones biológicos relevantes. La integración de estos hallazgos con otras capas de datos ómicos podría contribuir significativamente a la comprensión de los mecanismos subyacentes en las diferencias clínicas observadas.