Informe PEC1

Análisis de Datos Ómicos

María Luisa Reyes Conde

2025-04-01

Contents

1 Resumen

Este informe trata el estudio realizado sobre el dataset ST003776 que ha sido descargado de la plataforma *Metabolomics Workbench*. Este dataset está ligado al estudio titulado *Investigation of Hepatic Lipid Alterations Following Micro- and Nanoplastic-Ingestion*. Contiene información espectrométrica de 28 muestras que se encuentran en en formato `.mzML`. Estas muestras eran de tejido hepático de ratones y se analizaron sus perfiles lipídicos mediante técnicas lipidómicas.

El análisis realizado se ha centrado en crear un objeto del tipo *SummarizedExperiment*, que nos permite estructurar los datos de manera que podamos trabajar de forma eficiente y a la vez con las intensidades de los metabolitos, con los metadatos de las muestras dadas y con sus características. Una vez obtenido este objeto, se realiza un análisis exploratorio con el que poder obtener más información acerca de la distribución de los metadatos de nuestro dataset.

Cuando observemos los resultados obtenidos, veremos que existen diferencias en la distribución de las intensidades entre las muestras, demás de un patrón de agrupamiento en el PCA realizado. Esto nos lleva a pensar que hay cierta variabilidad entre los perfiles metabolómicos de los datos estudiados. Además,

obrevaremos que existe cierta correlación entre algunas muestras, de modo que nos permite evaluar si los datos experimentalos son coherentes y/o si tienen anomalías.

2 Objetivos

Los objetivos de este estudio son:

1. Tener los datos *crudos* organizados: Con el objeto *SummarizedExperiment* estrucutramos los datos de manera que sea más sencillo manejar toda la información que contiene el dataset.
2. Evaluar los datos: Analizando las intensidades de los metabolitos en las muestras de nuestro dataset podemos detectar posibles valores atípicos y las tendencias que puedan tener dichos datos.
3. Siguiendo con la línea del objetivo anterior, explorar posibles patrones en las muestras: Aplicando un Análisis de Componentes Principales podemos visualizar los posibles grupos y tendencias que hay en el conjunto de datos.
4. Determinar si existe correlación entre las muestras: Construyendo una matriz de correlación y representarla para observar si hay similitudes, evaluarlas en el caso de que hayan y estudiar la intensidad de estas.

La idea de este análisis que realizamos es que nos permita identificar la calidad de los datos y proporcionar información clave para los futuros estudios metabolómicos que puedan surgir de esta línea de investigación.

3 Métodos

3.1 Origen y naturaleza de los datos

Los datos utilizados en este estudio provienen del estudio obtenido de la plataforma *Metabolomics Workbench* con ID ST003776. En este dataset encontramos 28 archivos en formato `.mzML`, que representan muestras provenientes de tejido hepático de ratones analizadas mediante espectrometría de masas. En este estudio, llamado *Investigation of Hepatic Lipid Alterations Following Micro- and Nanoplastic-Ingestion* y realizado por la Universidad de Bonn (Alemania), se investigan las consecuencias metabólicas de la ingesta de microplásticos y nanoplásticos.

El dataset dado en este estudio incluye:

- Las intensidades de los metabolitos en cada muestra,
- Los metadatos de los metabolitos: nos encontramos con 35 variables que miden cada metabolito, como son la m/z (relación masa/carga) o el tiempo de retención (tiempo que tarda en atravesar una columna cromatográfica, desde que se inyecta la muestra hasta que se alcanza la respuesta máxima),
- Información genérica de las muestras: por ejemplo, los nombres y archivos de origen.

3.2 Metodología empleada

Aunque el apartado de los Resultados lo dividiremos en 2, a esta sección le debemos añadir un "apartado" previo:

3.2.1 Análisis preparatorio

Antes de crear el objeto *SummarizedExperiment* hemos hecho un pequeño análisis y reestructuración de los datos de nuestro conjunto, para que la creación de dicho objeto se lleve a cabo correctamente. Brevemente, los pasos seguidos han sido:

1. Carga de los archivos `mzML`: Importamos los 28 archivos contenidos en el dataset ST003776 utilizando el paquete MSnbase.

2. Obtención de los metadatos de los metabolitos y de las muestras: Extraemos la información clave de los metabolitos y lo almacenamos en un dataframe De igual manera se recopila la información de las muestras, guardándose también en un dataframe.
3. Extracción de intensidades: Obtenemos las matrices de intensidades de los metabolitos de cada muestra mediante la función chromatogram().
4. Alineación de datos: Como el número de filas de la matriz de intensidades no coincidía inicialmente con la cantidad de metabolitos de la rawdata, se filtró para que dichas filas coincidieran y las dimensiones fueran las correctas para la creación del objeto *SummarizedExperiment*. Como el número de columnas correspondía a las muestras en colData, no tuvo que hacerse ningún filtrado sobre la metadata.

3.2.2 Creación del objeto SummarizedExperiment

Para poder gestionar de manera eficiente el dataset, hemos creado un objeto del tipo *SummarizedExperiment* que tiene la siguiente estructura:

- Assay (*counts*): Matriz en la que nos encontramos las intensidades de los metabolitos. Se compone de 99 filas (los metabolitos) y 28 columnas (las muestras).
- rowData: La información detallada de cada metabolito. Son las 35 variables que explican cada metabolito.
- colData: Los ,etadatos de las muestras. En este caso es solo una variable, que corresponde con el nombre de cada archivo `.mzML`.

3.2.3 Análisis de datos exploratorio

Se han realizado variostipos de análisis para explorar y evaluar la calidad y estructura de los datos de nuestro conjunto:

1. Distribución de las intensidades de los metabolitos: Mediante boxplots (o diagramas de cajas y bigotes) podemos observar cómo se distribuyen las intensidades para cada muestra del dataset. Mediante este tipo de representaciones, podemos detectar la presencia de outliers (o valores atípicos) y así poder evaluar si existe homogeneidad entre los datos.
2. Análisis de Componentes Principales (PCA): Este tipo de análisis se usa para reducir la dimensionalidad del conjutno de datos. Así, podemos visualizar si existen agrupaciones entre las muestras y si es así, nos indicaría si estas muestras que se dividen en grupos tienen características comunes. Los dos primeros componentes principales los hemos representado en un gráfico de dispersión (realizado con `ggplot2`).
3. Análisis de correlación entre las muestras: Se construye una matriz de correlación entre las muestras del dataset que representamos mediante un heatmap para poder observar la intensidad de estas correlaciones.

3.3 Herramientas bioinformáticas y estadísticas

Es imprescindible tener en cuenta que se ha usado el Lenguaje R y el entorno de RStudio para todo el desarrollo y análisis. Además, mediante la herramienta de RStudio hemos realizado el informe en un archivo `Rmd` (RMarkdown) usando `TeX` para la composición de texto.

Los paquetes cargados para nuestro desarrollo en R han sido:

- SummarizedExperiment: Para la estructuración de los datos y para la creación del objeto con este mismo nombre.
- ggplot2: Para visualizar gráficas.
- pheatmap: Para el heatmap de las correlaciones.
- MSnbase: Para la gestión de archivos mzML y la extracción de las intensidades de los metabolitos.
- BiocParallel: Para la paralelización de procesos en análisis de espectrometría de masas.

4 Resultados

Este es de loa apartados más importantes de nuestro informe, ya que vamos a mostrar e interpretar los resultados que hemos obtenido mediante nuestro análisis. Lo dividimos en dos partes:

4.1 El objeto SummarizedExperiment y la estrucutra de los datos

Como el dataset que hemos descargado posee bastante cantidad de datos, decidimos usar el objeto *SummarizedExperiment*. Gracias al uso de este comando, podemos visualizar nuestros datos de una forma más estructurada. Veamos el resultado:

```
## class: SummarizedExperiment
## dim: 99 28
## metadata(0):
## assays(1): counts
## rownames(99): F01.S0001 F01.S0002 ... F01.S1259 F01.S1260
## rowData names(35): fileIdx spIdx ... scanWindowUpperLimit spectrum
## colnames(28): Sample68.mzML Sample69.mzML ... Sample94.mzML
##   Sample95.mzML
## colData names(1): sampleNames
```

Con esta salida podemos concluir los siguientes datos:

- El objeto tiene dimensión 99x28, es decir, posee 99 filas que corresponden a los metabolitos que hemos filtrado y los que hemos considerados más imprescindibles de incluir en el estudio y 28 columnas, que son las muestras que se han estudiado.
- La línea "metadata(0): " nos está indicando que no hay información adicional en la metadata, que es algo que se puede almacenar con el objeto "SummarizedExperiment". Un ejemplo sería información sobre la toma de muestras, si se hubieran tomado mediante mecanismos distintos o el tiempo en el que se toman dichas muestras si suponemos que hay una línea temporal en la que se recogen (por ejemplo, estudios en los que se recojan muestras en el día 1, día 20 y día 50).
- La línea "assays(1): counts" nos indica que hay una única matriz de datos, en este caso, la que se corresponde con los *counts* o, como lo hemos denominado, intensidades, que son los valores de expresión que poseen los metabolitos.
- Sobre las filas (metabolitos): el nombre de cada fila corresponde al nombre que se le asigna a cada metabolito. La rowData contiene 35 variables, que corresponden a las variables que miden los metabolitos y que nos proporcionan información sobre ellos, como son la masa/carga (m/z), el tiempo de retención o la intensidad total, entre otras.
- Sobre las columnas (muestras): el nombre de estas es el nombre de cada muestra (coincide con el nombre de cada fila de la metadata). De mismo modo, la colData contiene 28 variables, que se corresponden a las muestras de la metadata.

Como vemos, gracias a este objeto tenemos una imagen general del dataset, de cómo se organizan los datos y de la posible información que podemos obtener.

Otro objeto que se podríaa haber aplicado habría sido el *ExpressionSet*, que también sirve para manejar datos de experimentos de alto rendimiento en R. Y, ¿por qué escogemos utilizar *SummarizedExperiment*? Podemos enumerar algunas diferencias entre estos dos objetos:

1. Para empezar, *ExpressionSet* se suele usar para experimentos que están basados en microarrays donde las filas suelen llevar información sobre las características de las muestras a estudiar. Sin embargo, como hemos observado con *SummarizedExperiment*, las filas representan rasgos biológicos, como son los metabolitos en este caso.

2. En nuestro caso no aplica, ya que tenemos un solo assay, pero *SummarizedExperiment* permite trabajar con múltiples assays, manejarlas y asociarlas a los metadatos de manera más eficiente.

3. El objeto *SummarizedExperiment* está diseñado para integrarse de mejor manera con datos genómicos, lo cual hace que el manejo de estos datos sea cómodo y más sencillo que si se usara *ExpressionSet*.

Como podemos ver, el uso de *SummarizedExperiment* es más adecuado según qué tipos de análisis y datos vayamos a manejar. Si usamos datasets sobre secuenciación y genómica, es la mejor opción que podemos elegir.

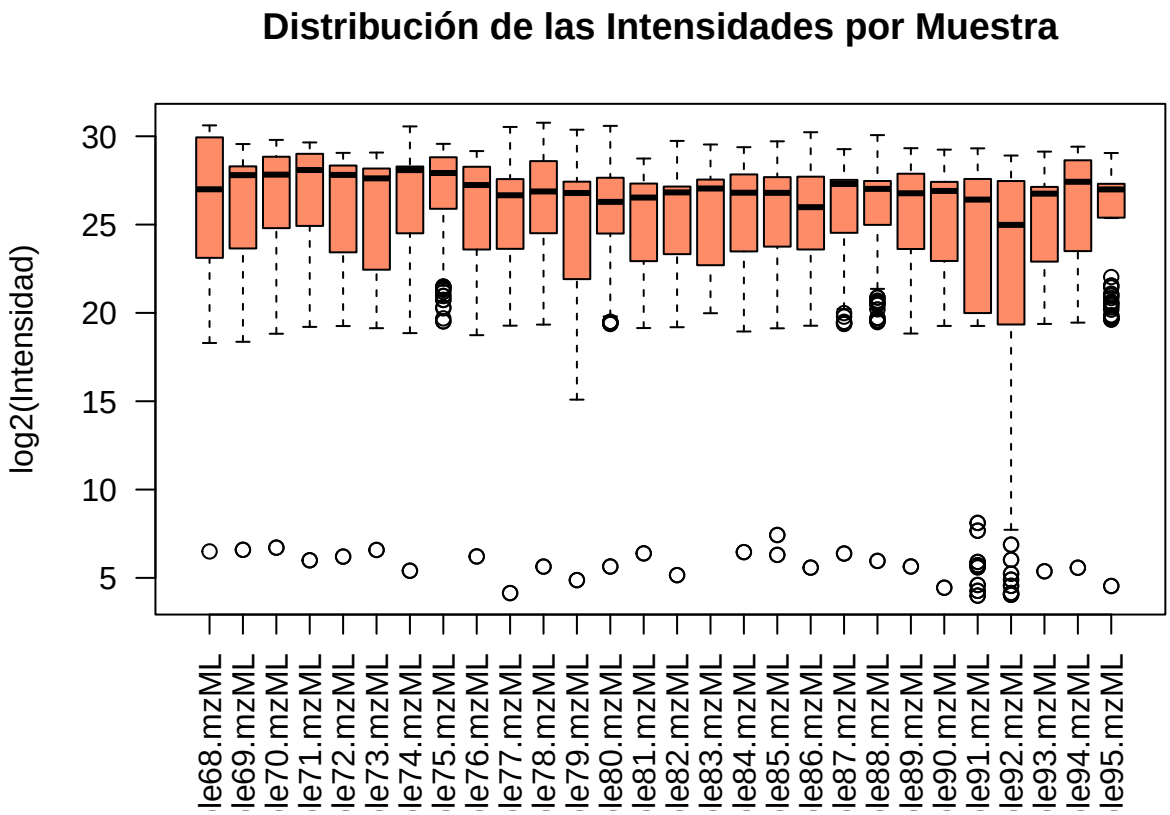
Una vez hecho el pre-análisis y el estrucuturamiento de los datos mediante el comando en R de *SummarizedExperiment*, podemos pasar al análisis exploratorio de los datos para entender los datos y ver cómo se relacionan, qué información nos dan y qué podemos hacer con ellos.

4.2 Análisis exploratorio

Se realizan varias gráficas para poder visualizar la información obtenida gracias a nuestra estructuración del dataset.

4.2.1 Distribución de intensidades

El primer paso en nuestro análisis exploratorio es ver si realmente el dataset es de calidad. Para esto, generamos en un mismo gráfico, un boxplot de las intensidades de los metabolitos de cada muestra:



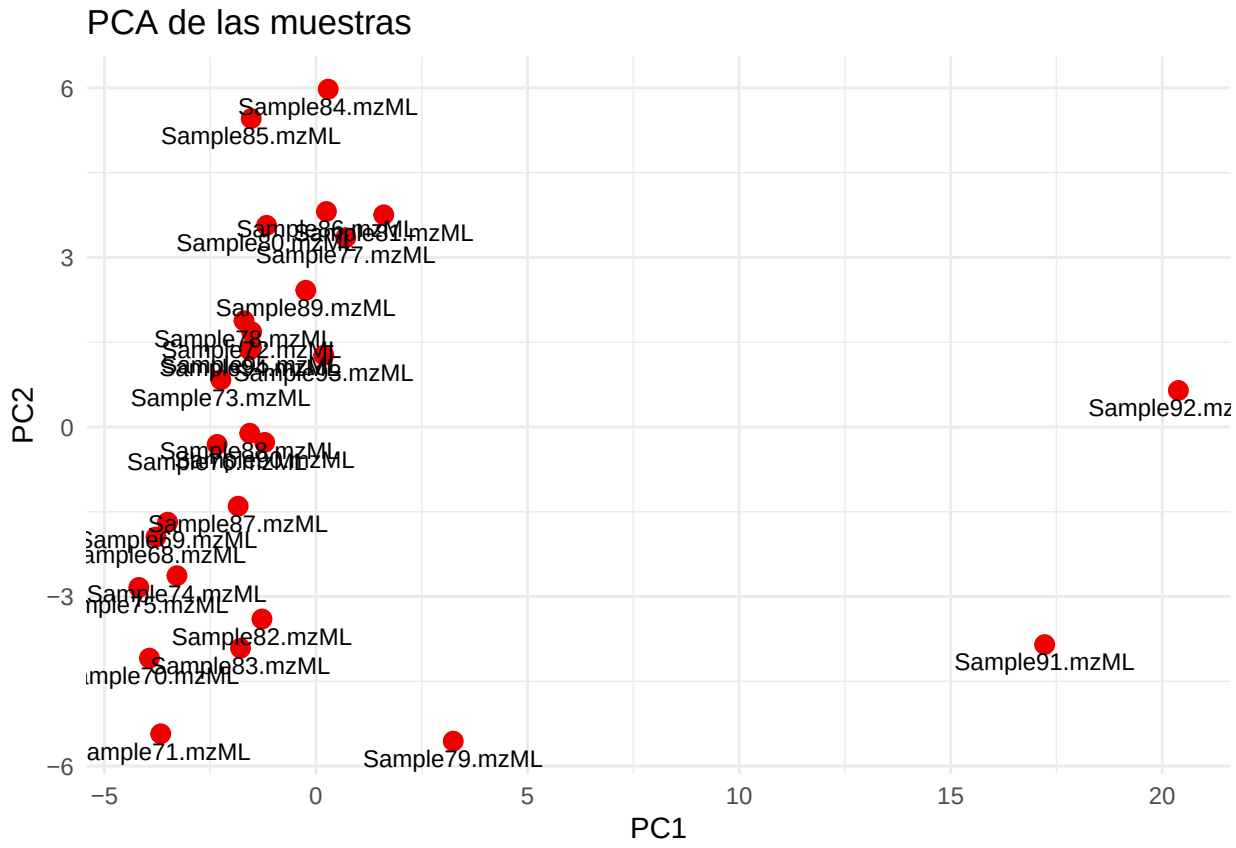
Es importante destacat que la conversiónde logarítmica de las intensidades se ha realizado a modo de normalización para que sea más fácil la interpretación del gráfico y reduzca la variabilidad y posibles efectos técnicos que puedan generar diferencias que no sean biológicas en nuestros datos.

Como vemos la mayoría de las muestras tienen una mediana que se encuentra dentro del rango 25-27. La dispersión de los datos parece ser bastante homogénea entre las muestras y los boxplots tienen alturas

similares, lo que indica que la variabilidad de cada muestra es comparable. SIn embargo, se observa que ciertas muestras poseen valores atípicos (outliers), lo que puede indicarnos que haya algún tipo de probelma en los datos o que haya diferencias biológicas entre las muestras. Podemos destacar la muestra 24, que es la que más outliers presenta y con valores muy bajos: esto nos puede alertar de que la muestra sea de menor calidad que el resto.

4.2.2 Análisis de Componentes Principales (PCA)

Aplicamos un PCA para reducir la dimensionalidad y para visualizar si existen agrupaciones entre las muestras. Obtenemos el siguiente gráfico:



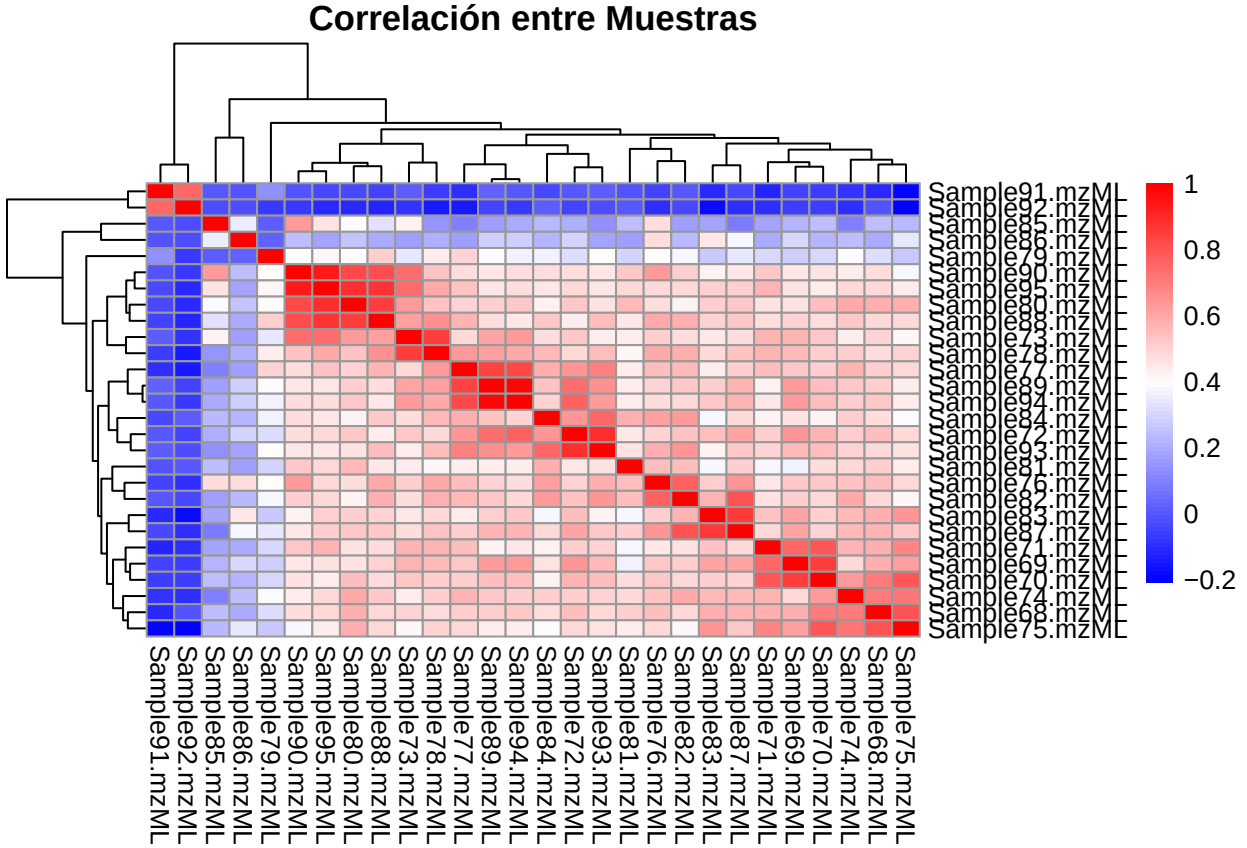
Observamos que a mayoría de las muestras están agrupadas cerca del origen, lo que nos puede sugerir que las muestras tienen perfiles similares en los datos originales. Sin embargo, vemso que hay 3 muestras que se alejan notoriamente del grupo principal: Sample92.mzML, Sample91.mzML y Sample79.mzML. Estas muestras podrían representar valores atípicos o informarnos de que poseen cierta diferencia biológica o experimental que es significativa con respecto del resto de muestras. De hecho, comprobando el conjunto metadata y observando los boxplots nos podemos percatar de que corresponden, respectivamente, con las muestras 25, 24 y 12 que son representadas en el boxplot y vemos que son muestras que presentan outliers con intensidades muy bajas. Aún así, que estas muestras estén tan alejadas puede deberse a que haya contaminación o fallos en el procesamiento de los datos, así como a problemas que hayan surgido en la normalización de estos. Los componentes representan la variabilidad en los datos (teniendo PC1 mayor peso). Por tantom cuanto más extremos sean las muestras, más diferentes serán estas del resto.

La continuación de este análisis sería revisar de forma más exhaustiva las muestras que están alejadas del grupo principal. Además, sería conveniente analizar los valores de los dos componentes para ver qué variables son las que más contribuyen en la agrupación de las muestras y rehacer el PCA sin muestras atípicas para ver si el patrón que se ha obtenido se sigue manteniendo.

En resumen, el PCA obtenido nos sugiere que la mayoría de las muestras son similares, pero que hay algunas que podrían ser atípicas y/o podrían ser fuente de una variabilidad importante en el experimento, contribuyendo a que el análisis no sea lo satisfactorio que nos gustaría que fuera.

4.2.3 Correlación entre muestras

Por último, se construyó una matriz de corelación entre las muestras y se representó mediante un mapa de calor:



Observamos que hay algunas muestras que poseen colores intensos lo cual indica que la correlación es más alta, ya sea de manera negativa como positiva. Vemos que los tonos rojos indican correlación positiva: esto nos hace pensar que los perfiles metabolómicos de estas muestras tienen similitudes y que están asociadas de alguna manera. En cambio, los tonos azules más oscuros son aquellas muestras que están correlacionadas negativamente, de manera que sus perfiles metabolómicos presentan diferencias bilógicas entre sí. Un aspecto positivo de este mapa de calor: las correlaciones positiva son altas. Es decir, aquellas muestras con colores rojos intensos indican que están muy correlacionadas y que los perfiles son muy parecidos. En cambio, en las correlaciones negativas aunque los azules sean intensos, los valores no son altos, lo cual indica que la correlación no tiene suficiente fuerza.

Nos podemos dar satisfechos con los resultados, aunque debería de continuarse con el análisis, intentar obtener más información de las muestras para poder realizar de nuevo el análisis exploratorio sin los datos atípicos. Así, podríamos ver realmente qué relación hay entre las muestras y si su correlación posee verdaderamente peso sobre el estudio biológico que se está realizando.

5 Discusión

Los resultados del análisis exploratorio que hemos realizado y que acabamos de observar nos revelan aspectos clave sobre la calidad y estructura del dataset que hemos estudiado. Aún así, encontramos ciertas limitaciones

que nos pueden empujar a continuar estudiando y analizando este dataset para encontrar, posteriormente, información más relevante sobre el estudio.

5.1 Sobre la calidad de los datos

Observamos que hay variabilidad en las intensidades de los metabolitos lo que nos sufiere que se debería realizar una normalización sobre las muestras. Con la normalización de los datos, eliminaríamos elementos redundantes y posibles datos dependientes de otros. Además, corregiríamos los datos duplicados o "anormales" y mejoraríamos la coherencia de los datos. Todo ello nos llevaría a optimizar el tiempo y el almacenamiento que se necesita para el procesamiento de los datos y nos haría más llevadero el proceso del análisis. De hecho, observando los boxplots de las intensidades, la presencia de valores extremos o atípicos en algunos de los metabolitos nos podría indicar que diferencias biológicas reales o posibles errores experimentales entre las muestras analizadas.

5.2 Patrones en los datos

El Análisis de Componentes Principales sugiere que hay cierta agrupación entre muestras, aunque no son grupos demasiado diferenciados y están poco compensados entre sí. Aún así, esto podemos interpretarlo suponiendo que existe cierta influencia por los factores experimentales o por los factores biológicos sobre algunos de los perfiles metabolómicos. Además, la matriz de correlación confirma que ciertas muestras están correlacionadas (tanto positiva como negativamente), pero se observan algunas con baja correlación, lo que podría indicar ruido en los datos. Es por ello que un análisis más exhaustivo y que "limpie" un poco los datos sería una buena continuación para este estudio.

5.3 Las limitaciones del estudio

De nuevo, nos reiteramos al hecho de que una normalización para "corregir" los datos podría mejorar mucho la calidad del análisis. Así, se podrían corregir posibles efectos *batch* y llegar a una mejor interpretación de los resultados. El análisis exploratorio que hemos realizado no nos permite identificar cuáles son los metabolitos que se "diferencian", por lo que sería conveniente realizar análisis con modelos estadísticos más concretos y/o apliar aprendizaje automático. Además, noo se han considerado los posibles factores experimentales adicionales, como la posible degradación de muestras o diferencias en la preparación del experimento cuando se realizó el estudio del que proviene nuestro dataset.

5.4 Implicaciones biológicas

Como hemos nombrado anteriormente, podemos definir este estudio como un primer paso para la explroación de este conjunto de datos. Con un análisis más exhaustivo podríamos detectar cuáles son los perfiles metabolómicos que se diferencian o que destacan más en la respuesta del tejido hepático de los ratones a la ingesta de microplásticos y nanoplásticos.

6 Conclusiones

Con este análisis y ssu resultados podemos sacar las siguientes conclusiones:

1. Buena estructuración de los datos: Con el objeto *SummarizedExperiment* se logró organizar el dataset, de manera que se facilitó la compresión de la estrctura y forma de los datos metabolómicos.
2. Variabilidad en las muestras: Mediante el bozplot se detectó una distribución bastante heterogénea de las intensidades de las muestras, lo que sugiere la necesidad de normalización para futuros análisis.
3. PCA y correlaciones: Con el Análisis de Componentes Principales se identificaron patrones de agrupamiento. Además, gracias a la visualización del heatmap de correlacones, observamos que existen similitudes entre algunas muestras. Estos dos resultados nos indica que los datos contienen información estructurada y que pueden tener bastante relación entre sí.

4. Futuras direcciones y posibles estudios: Para mejorar la interpretación biológica del dataset, recomendaría realizar la normalización de los datos y eliminar valores atípicos. Una vez realizado esto, sería conveniente aplicar algunos análisis estadísticos más avanzados para identificar metabolitos diferenciales que nos puedan dar información de peso sobre el objeto de estudio, que era la observación y análisis de las consecuencias metabólicas de la ingesta de microplásticos y nanoplásticos.

7 Rreferencias

Repositorio de GitHub: PEC1-Análisis de Datos Ómicos - María Luisa Reyes Conde [1]