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

Angie Serrano García

2025-03-28

Tabla de contenidos

1 Abstract

En este trabajo se exploraron perfiles metabolómicos asociados a diferentes estados de la enfermedad hepática grasa no alcohólica (NAFLD), utilizando datos del estudio ST000915 del repositorio Metabolomics Workbench. Se seleccionaron los datos correspondientes al análisis de fosfolípidos (Core H), que fueron incorporados en un objeto llamado SummarizedExperiment, integrando tanto los datos de expresión como los metadatos clínicos. Posteriormente, se realizó un análisis exploratorio que incluyó estadísticas descriptivas, visualización de la distribución de intensidades y un análisis de componentes principales (PCA). Se observó una variabilidad entre metabolitos, y el PCA mostró cierta separación entre diagnósticos, especialmente entre los grupos de pacientes normales y con cirrosis, aunque con solapamientos entre clases. Estos resultados sugieren patrones metabólicos diferenciables entre estados de NAFLD, aunque no de forma completamente discriminatoria. El trabajo demuestra el valor de la metabolómica y las herramientas bioinformáticas para el estudio de enfermedades complejas como la NAFLD.

2 Introducción

La metabolómica, estudio sistemático de los perfiles metabólicos en muestras biológicas, es clave en los análisis ómicos, ya que permite identificar biomarcadores que sirven como diagnóstico y pronóstico en enfermedades tales como la enfermedad hepática grasa no alcohólica (NAFLD) (1,2). La detección de marcadores metabólicos en esta población es fundamental para poder identificar pruebas diagnósticas no invasivas.

Por tanto, la presente prueba de evaluación continúa (PEC) se centra en el análisis de un dataset de metabolómica humana proveniente del repositorio de Metabolomics Workbench, que incluye metabolitos de pacientes con NAFLD (3). Para esto, se utilizó el entorno R, el ecosistema Bioconductor y Git hub.

2.1 Objetivos

Objetivo General: - Planificar y ejecutar una versión simplificada del proceso de análisis de datos ómicos a partir de un dataset de metabolómica, aplicando técnicas exploratorias para obtener una visión general de los datos y su relevancia biológica.

Objetivos específicos: - Estructurar los datos y metadatos dentro de un objeto SummarizedExperiment, integrando adecuadamente la matriz de expresión y los metabolitos estudiados. - Realizar un análisis exploratorio del dataset que incluya estadísticas descriptivas, visualizaciones de distribución y un análisis de componentes principales (PCA). - Interpretar los resultados desde una perspectiva biológica, considerando su aplicabilidad en la caracterización de la progresión de NAFLD.

3 Materiales y métodos

3.1 Selección de los datos para el estudio

Para el desarrollo de esta PEC se utilizó un dataset de metabolómica de Metabolomics Workbench, correspondiente al estudio ST000915 titulado ''Biomarkers of NAFLD progression: a lipidomics approach to an epidemic. Part 1:Liver''. Este estudio consta de perfiles lipídicos de muestras humanas, para identificar biomarcadores asociados a NAFLD. Se seleccionó especificamente la data MS: Core H (Phospholipids) debido a que tenía un buen número de metabolitos, y la calidad era también optima. Integrar a esta PEC los otros subestudios que tenia esta dataset implicaba una mayor complejidad, especialmente por el manejo de datos heterogeneos, lo cual no se contemplaba para el alcance de esta PEC.

El dataset está en formato .txt tipo mwTab, e incluye los valores cuantitativos de los metabolitos, así como el diagnóstico clínico (Normal, esteatosis, esteatohepatitis no alcoholica, cirrosis).

3.2 Procedimiento de análisis y herramientas empleadas

El análisis de los datos se realizó utilizando R y herramientas del ecosistema Bioconductor. Inicialmente, se descagaron los datos en formato texto y se inspeccionó la estructura del archivo para identificar especificamente desde dónde comenzaban los datos cuantitativos de los metabolitos, que

fue a partir de la línea 263, por lo que esto debio especificarse en el código para el análisis de los datos. Se obtuvo una tabla con 226 metabolitos (filas) y 88 muestras (columnas).

A partir de lo anterior, se generó un objeto llamado SummarizedExperiment, que permite integrar la matriz de expresión (que es la intensidad de los metabolitos) y los metadatos de las muestras y las variables (identificador del metabolito). Por eso, SummarizedExperiment fue seleccionado ya que permite utilizar diferentes paquetes para el análisis ómico de Bioconductor.

Por otra parte, para el análisis de los datos, se realizó un análisis exploratorio evaluar la distribución de los datos, así como posibles valores atipicos. Se calcularon estadísticas descriptivas básicas (media y desviación estandar por metabolito) mediante el paquete matrixStats. Así mismo, se generaron boxsplots para las intensidades log-transformadas para cada muestra, para observar variabilidad. Finalmente, se realizó un análisis de componentes principales (PCA), utilizando paquetes como FactoMineR y factoextra, para visualizar la estructura multivariada de los datos.

3.3 Estructura de datos: SummarizedExperiment vs ExpressionSet

SummarizedExperiment y ExpressionSet son objetos que permiten representar datos experimentales, sin embargo, SummarizedExperiment es más moderno y está basado en objetos como Assays, rowData, ColData, entre otros, lo que permite una diferenciación entre datos y anotaciones. Además, SummarizedExperiment es más fácil de usar al momento de manipular los datos, facilitando de esta manera un análisis reproducible (4).

Aunque la tabla original contenía 227 filas, una de ellas correspondía a los factores experimentales (diagnóstico), y no a un metabolito. Por eso, el análisis se realizó finalmente sobre 226 metabolitos.

```
## Tablón cargado:  227  filas y  89  columnas
```

```
## objeto SummarizedExperiment creado con éxito.
```

4 Resultados

4.1 Estructura de los datos

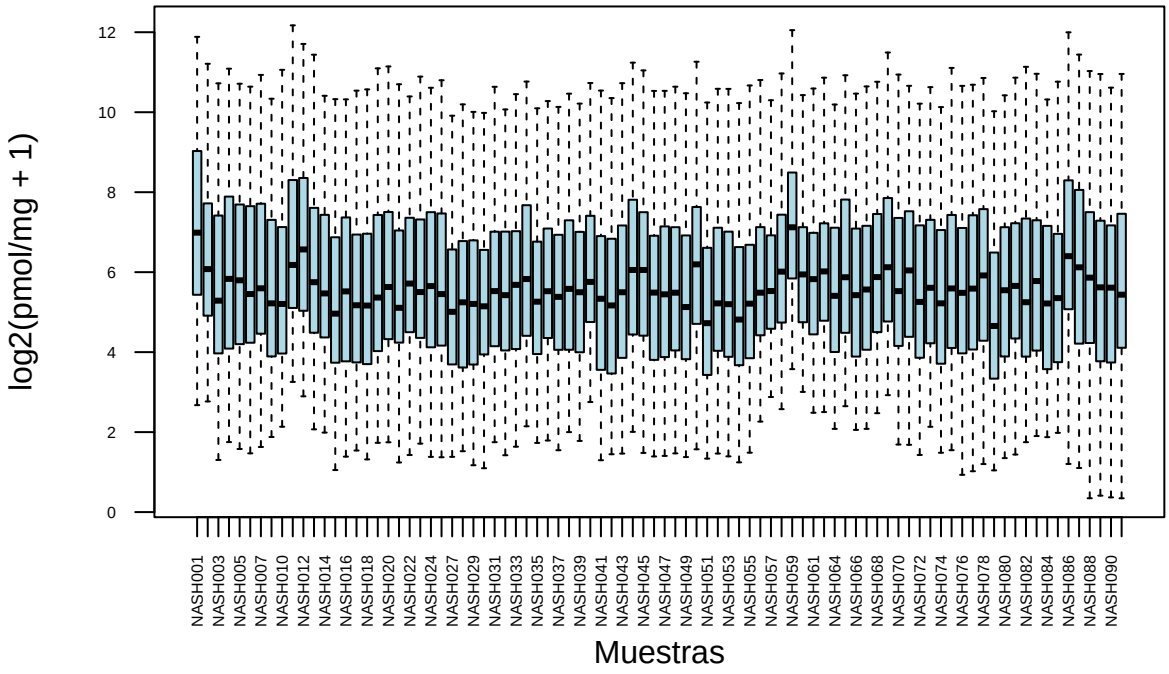
El dataset que se seleccionó contiene datos metabolómicos obtenidos por espectrometría de masas en muestras de tejido hepático humano, procedentes del estudio ST000915 del repositorio Metabolomics Workbench. En este estudio se analizaron un total de 88 muestras, clasificadas según el diagnóstico clínico en cuatro grupos: cirrosis (20), NASH (20), esteatosis simple (17) y normales (31). Esta distribución, aunque no está equilibrada, permite una comparación inicial entre las diferentes etapas de la enfermedad hepática por NAFLD. Llama la atención que el grupo con mayor número de muestras es el de individuos con hígado normal, lo cual es útil para tener una referencia de los perfiles metabólicos considerados ‘’saludables’’.

El grupo más pequeño fue el de esteatosis simple, lo que podría limitar un poco la capacidad de observar diferencias claras en esta fase intermedia de la enfermedad. Aun así, la cantidad de muestras disponibles en los otros grupos permite tener una visión general de cómo varía el metabolismo a lo largo del espectro de NAFLD.

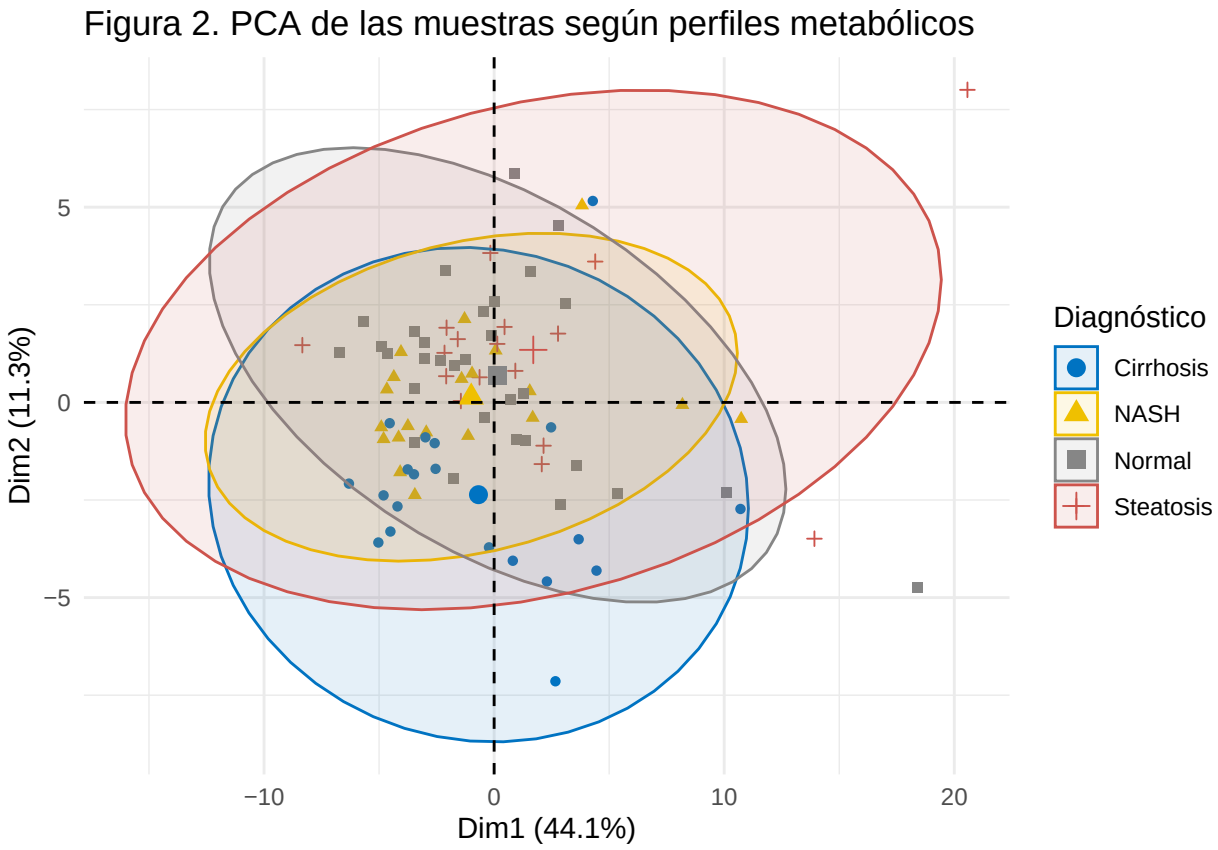
4.2 Análisis exploratorio de los datos

```
## Número de muestras: 88

## Número de metabolitos: 226

##
## Tabla 1. Distribución de muestras por grupo diagnóstico y estadística descriptiva.

##
## Cirrhosis      NASH    Normal Steatosis
##        20        20        31        17

##
## Resumen de medias por metabolito:

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##    4.726   19.219   48.641  171.053  150.588 1581.575      116

##
## Resumen de desviaciones estándar por metabolito:

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   2.243  12.896  27.535  78.504  72.790 656.333     116
```

Figura 1. Distribución log2 de intensidades por muestra



```
## Expr_t dimensions after NA filter: 88 57
```



4.3 Datos generales

El análisis descriptivo de los niveles de expresión de los metabolitos mostró una gran variabilidad entre ellos. Algunas moléculas tienen intensidades medias bastante bajas (por debajo de 5 pmol/mg), mientras que otras alcanzan valores mucho más altos (más de 1500 pmol/mg) (tabla 1). Esta diferencia probablemente refleja tanto la diversidad en la abundancia natural de los distintos fosfolípidos como la variabilidad entre los pacientes.

En cuanto a la desviación estándar, también se observó una dispersión importante. Algunos metabolitos presentaron muy poca variabilidad entre muestras, mientras que otros tuvieron valores mucho más altos, lo cual podría indicar diferencias biológicas relevantes o incluso la presencia de posibles biomarcadores. Estos datos, aunque preliminares, ya sugieren que el perfil metabolómico varía de forma significativa entre individuos y posiblemente entre grupos diagnósticos.

No se detectaron valores atípicos extremos en las muestras, aunque sí metabolitos con alta dispersión. La transformación logarítmica permitió estabilizar la varianza y visualizar mejor la distribución general de los datos (Figura 1).

Algo importante a considerar es la presencia de valores faltantes; se identificaron 116 datos que faltaban en la matriz de expresión. Esto no es raro en estudios ómicos, pero sí plantea retos en el análisis posterior, especialmente en técnicas como PCA, que requieren una matriz sin valores perdidos. Por eso, antes de realizar ese análisis, se eliminaron las variables con valores faltantes.

4.4 Distribución de intensidades metabolómicas

La figura 1 mediante cada caja, representa la variabilidad en la abundancia de cada metabolito. Se observa una distribución homogenea entre muestras, lo que sugiere que la mayoria tienen perfiles de intensidad similares. Sin embargo, algunas muestras tienen una dispersión más amplia o más estrecha, lo cual podría ser debido a diferencias biologicas o en la variabilidad de la técnica empleada. Además, también se aprecia que algunos muestras presentan bigotes mas largos, probablemente por alteraciones metabolicas de ciertos diagnósticos, que será explorado mediante el análisis multivariado del PCA.

4.5 Análisis de componentes principales (PCA)

La Figura 2 muestra el resultado del PCA, la cual es una técnica multivariante que permite disminuir la dimensionalidad de los datos al identificar las direcciones de mayor variabilidad. En este caso, se analizaron las intensidades de 57 metabolitos medidos en 88 muestras, clasificadas en los cuatro grupos diagnósticos ya mencionados previamente.Los dos primeros componentes explican conjuntamente el 55.4% de la variabilidad total (Dim1: 44.1% y Dim2: 11.3%). Cada punto en el gráfico representa una muestra individualmente, codificada por color y forma según su diagnóstico. Las elipses indican la dispersión y tendencia central de cada grupo. Se observa una cierta separación parcial entre los grupos, especialmente entre las muestras con diagnóstico de cirrosis (azul) y aquellas con diagnóstico normal (gris), lo cual sugiere diferencias en los perfiles metabólicos entre estos estados fisiopatológicos. Las muestras clasificadas como NASH (amarillo) y steatosis (rojo) se sitúan en regiones intermedias, con cierto solapamiento entre ellas.nAunque el solapamiento entre grupos impide una separación completamente clara, la estructura general del PCA sugiere que existen patrones de expresión metabolómica que dependen de la progresión de la enfermedad hepática. Estos resultados son consistentes con el objetivo del estudio original, que busca identificar biomarcadores capaces de discriminar entre estados como esteatosis, NASH y cirrosis.

5 Discusión

El análisis realizado permitió explorar los perfiles metabolómicos de pacientes con diferentes estados de NAFLD, utilizando datos obtenidos de Metabolomics Workbench del estudio ST000915. Para estructurar los datos, se utilizó un objeto de clase llamado SummarizedExperiment, que facilitó la organización conjunta de los datos de expresión y la información clínica de las muestras.

Una de las limitaciones principales del estudio fue la presencia de valores faltantes en varios metabolitos, lo que llevó a excluir algunas variables del análisis exploratorio, especialmente en el PCA. Esto podría haber reducido la capacidad de análisis para identificar la variabilidad biológica. No obstante, el análisis de los datos permitió observar diferencias relevantes en los niveles de expresión entre las diferentes muestras, evidenciado a través de la distribución log-transformada de intensidades y el PCA.

En el PCA se evidenció cierta separación entre los pacientes con cirrosis y los clasificados como normales. Por el contrario, los grupos con NASH y esteatosis mostraron un mayor solapamiento, lo que podría deberse a similitudes metabólicas entre estas fases o a una mayor heterogeneidad dentro de esos subgrupos. Este resultado concuerda con lo reportado en la literatura (5), donde se describe la progresión de NAFLD como un espectro continuo en el que algunos perfiles metabolómicos pueden

superponerse entre fases clínicas, dificultando su diferenciación sin el apoyo de herramientas más complejas o un mayor tamaño muestral.

Desde el punto de vista metodológico, se emplearon herramientas ampliamente validadas en bioinformática, como SummarizedExperiment, matrixStats y FactoMiner, que son robustas y adecuadas para el análisis ómico. Sin embargo, dado que el enfoque de esta PEC fue descriptivo y no se aplicaron pruebas estadísticas inferenciales (como ANOVA o modelos lineales), no se puede afirmar si las diferencias observadas son estadísticamente significativas, ni si ciertos metabolitos podrían tener un rol diagnóstico o pronóstico.

Así mismo, se debe tener presente que para esta PEC se trabajó solamente con los datos del core H (fosfolípidos), dejando por fuera otras clases lipídicas que también podrían ser relevantes en la fisiopatología de la NAFLD. No obstante, esta decisión fue tomada para delimitar el alcance del trabajo y mantener un solo enfoque. Sin embargo, es importante señalar que en un estudio más amplio, la integración de distintas clases de metabolitos podría ser esencial para construir un modelo diagnóstico más robusto.

Por último, aunque este trabajo se ha centrado en un análisis exploratorio, el enfoque seguido es el análisis inicial para futuros trabajos e investigaciones. Estos datos podrían ser utilizados en modelos de clasificación, análisis multivariados supervisados (como PLS-DA) o integrarse con otros tipos de datos ómicos (transcriptómica, proteómica) para obtener mayor información de la enfermedad hepática.

6 Conclusiones

- El enfoque utilizando SummarizedExperiment permitió organizar de una manera más eficiente los datos metabolómicos y sus metadatos, facilitando el análisis exploratorio.
- Se identificaron patrones en los perfiles metabolómicos que podrían sugerir una diferencia entre los diferentes estados de la NAFLD, especialmente entre aquellos pacientes normales y aquellos con cirrosis.
- El PCA permitió identificar agrupamientos de acuerdo al tipo de diagnóstico, lo que sugiere que el metabolismo lipídico se va alterando progresivamente a medida que avanza la enfermedad.

7 Referencias

Enlace al repositorio de GitHub: https://github.com/Angieserranouoc/Serrano-Garcia-Angie-PEC1

1. Klassen, A. et al. (2017) “Metabolomics: Definitions and significance in systems biology”, Advances in experimental medicine and biology, 965, pp. 3–17. https://doi.org/10.1007/978-3-319-47656-8_1.
2. Masoodi, M. et al. (2021) “Metabolomics and lipidomics in NAFLD: biomarkers and non-invasive diagnostic tests”, Nature reviews. Gastroenterology & hepatology, 18(12), pp. 835–856. https://doi.org/10.1038/s41575-021-00502-9.

3. LIPID MAPS y Fahy, E. (2018) Metabolomics Workbench : NIH data repository, Metabolomicsworkbench.org. Disponible en: https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Study&StudyID=ST000915&StudyType=MS&ResultType=1

4. ExpressionSet and SummarizedExperiment (sin fecha) Sthda.com. Disponible en: https://www.sthda.com/english/wiki/expressionset-and-summarizedexperiment

5. Friedman, S.L. et al. (2018) "Mechanisms of NAFLD development and therapeutic strategies", Nature medicine, 24(7), pp. 908–922. https://doi.org/10.1038/s41591-018-0104-9.