Informe PEC 1. Asignatura: Análisis de Datos Ómicos

José Alberto Camacho López

2025-04-02

Informe de la PEC 1: Análisis de datos metabolómicos

Tabla de Contenidos

1. Abstract

En este análisis se examinan los perfiles metabólicos de muestras biológicas para entender las diferencias en los metabolitos presentes en diferentes grupos de estudio. Llevamos a cabo un análisis exploratorio de los datos utilizando técnicas estadísticas como el análisis de componentes principales (PCA) y el clustering jerárquico. El análisis de PCA nos ayudó a identificar las principales fuentes de variabilidad en los datos, mientras que el clustering jerárquico sirvió en este caso para agrupar las muestras con características similares.

Los resultados revelaron patrones biológicos y posibles agrupaciones que podrían estar asociadas a las condiciones experimentales. Además, se generaron gráficos que visualizan las relaciones entre las muestras. Este informe describe el proceso, los métodos y los resultados de este análisis.

2. Objetivos

Los objetivos de este análisis son los siguientes:

- **Exploración de datos**: Examinar las características principales de los datos de nuestras muestras y obtener una visión general de la variabilidad entre las muestras.

- **Creación de un objeto de clase SummarizedExperiment**: este objeto debe contener los datos y los metadatos (información acerca del dataset, sus filas y columnas). Además, debemos definir la diferencia de esta clase con respecto a la clase *ExpressionSet*.
- **Análisis de componentes principales (PCA)**: Realizar un PCA para reducir la dimensionalidad de los datos y detectar patrones subyacentes.
- **Clustering jerárquico**: Aplicar clustering jerárquico para agrupar las muestras en función de sus perfiles metabólicos.
- **Visualización de los resultados**: Generar gráficos tales como histogramas, boxplots y heatmaps para representar las distribuciones del counts de muestras y las relaciones entre las mismas.

3. Métodos

3.1. Datos

Los datos analizados en este estudio provienen de un conjunto de muestras biológicas de diferentes grupos de estudio, que se haya en la página de Github del Borenstein Lab y encontramos datos curados de metabolómica (https://github.com/borenstein-lab/microbiome-metabolome-curated-data) y elegimos el dataset de FRANZOSA_IBD_2019 (https://github.com/borenstein-lab/microbiome-metabolome-curated-data/tree/main/data/processed_data/FRANZOSA_IBD_2019).Elegimos este dataset porque nos interesa cómo el microbioma puede estar relacionado con ciertos tipos de cáncer.

Descargamos la matriz de expresión génica (genera.counts.csv), con las muestras en las filas y las bacterias del microbioma en las columnas; el archivo de metadatos de las muestras (metadata.csv), que será utilizados como metadata de las filas (samples); y el archivo que contiene información adicional sobre los metabolitos (mtb.csv). Como veremos, en los datos de metabolitos, cada muestra está representada por un vector que contiene los valores de intensidades de diferentes metabolitos. Estas intensidades corresponden a las mediciones de metabolitos específicos en cada muestra, obtenidas mediante técnicas de análisis como la espectrometría de masas.

Las muestras fueron etiquetadas con identificadores únicos (por ejemplo, PRISM.7122, PRISM.7147, etc.) y se asociaron con metadatos que contienen información adicional, como el grupo experimental y otras características clínicas.

Los datos fueron organizados en un objeto de la clase `SummarizedExperiment`, lo que nos permitió manejar tanto los datos de conteo como los metadatos asociados.

3.2. Procesamiento de Datos

Dado que los datos de conteo tienen una gran dispersión, pues tanto el mínimo como la mediana y primer cuartil tenían valor cero, además de un tercer cuartil realmente bajo, nos vimos obligados a realizar una transformación logarítmica, log1p, es decir (log(1 + x)) -para evitar el log 0- a nuestro assay.

3.3. Análisis de Componentes Principales (PCA)

El PCA fue realizado utilizando la función `prcomp()` de R. Esta técnica reduce la dimensionalidad de los datos al proyectar las muestras en un espacio de menor dimensión (en el caso que nos ocupa, 2 componentes principales) para facilitar la visualización de las relaciones entre las muestras.

3.4 Clustering Jerárquico

El clustering jerárquico fue realizado para agrupar las muestras en función de sus perfiles metabólicos. Se calculó la matriz de distancias utilizando la función dist(), es decir la distancias euclídea,entre las distintas muestras, y luego se construyó un dendrograma con la función hclust(). Este método jerárquico ayuda a identificar las muestras más similares entre sí.

3.5. Visualización

Se generaron varios gráficos para facilitar la interpretación de los resultados.
Entre ellos se incluyen:

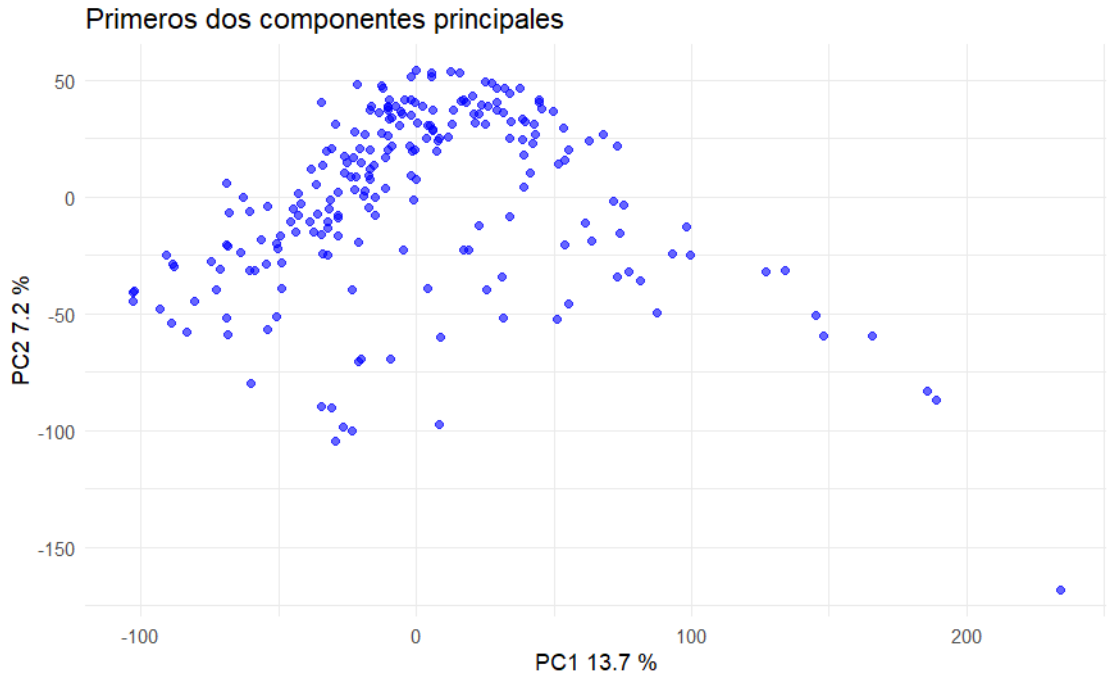
- **PCA**: Un gráfico de dispersión de las primeras dos componentes principales.
- **Clustering Jerárquico**: Un dendrograma que muestra cómo se agrupan las muestras en función de sus perfiles metabólicos.
- **Heatmap**: Un mapa de calor para visualizar las relaciones entre las muestras y los bacterias del microbioma.

4. Resultados

4.1. Análisis de Componentes Principales (PCA)

El análisis de componentes principales (PCA) mostró que las primeras dos componentes explican una buena parte de la variabilidad de los datos, con PC1 explicando un 13,7% y PC2 un 7.2 % de dicha variabilidad.

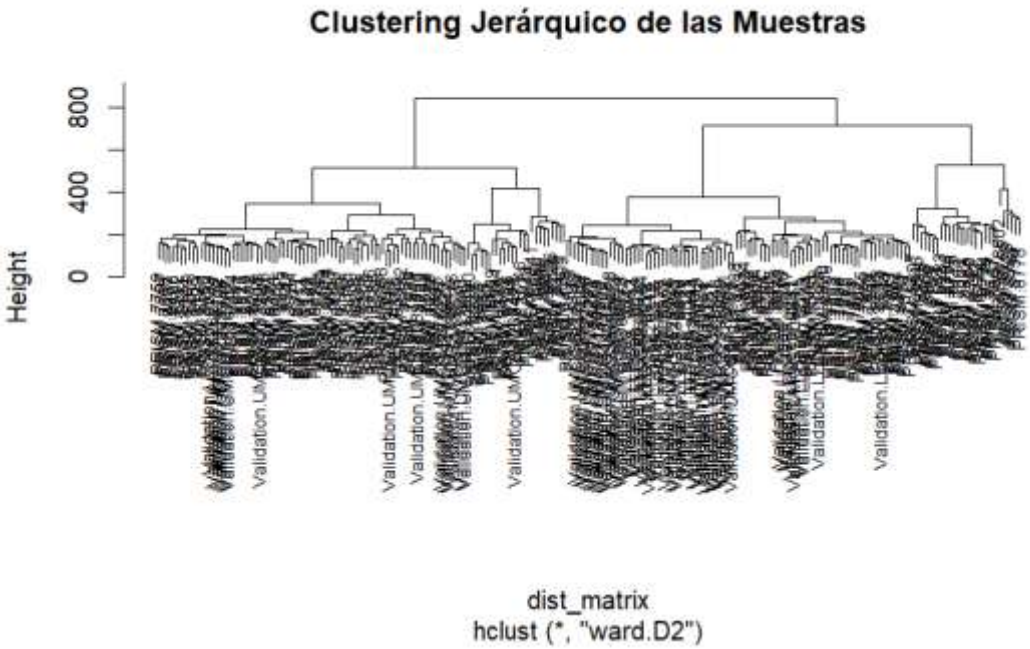
A continuación, mostramos este gráfico:



4.2. Clustering Jerárquico

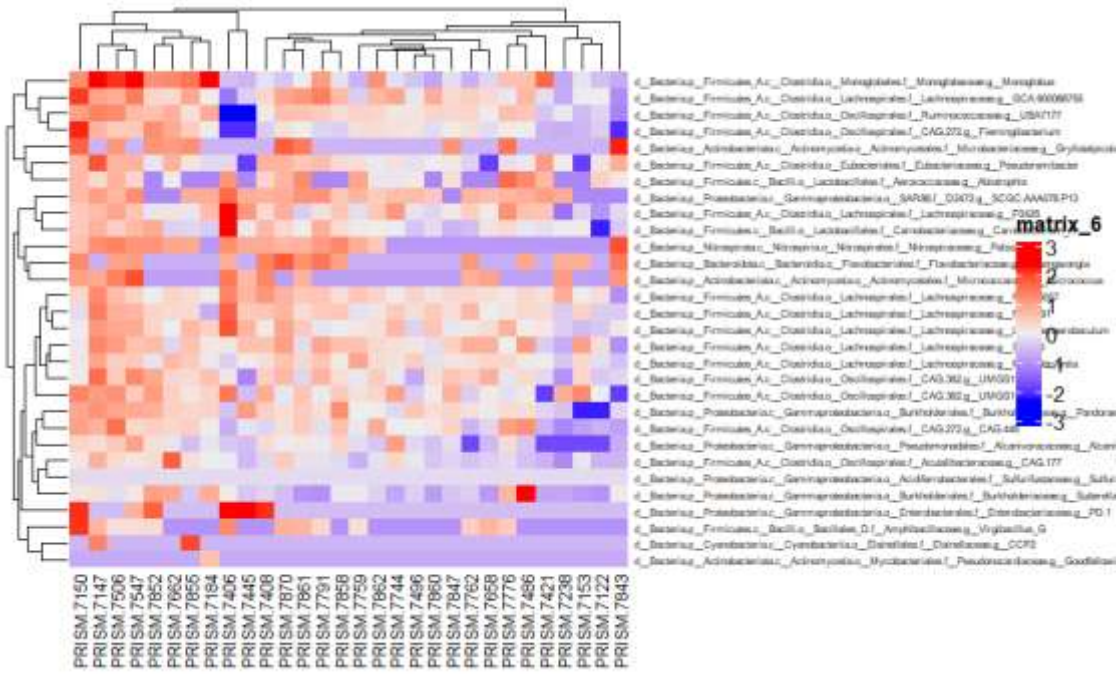
El clustering jerárquico reveló que las muestras se agrupan en tres grandes clusters, lo que sugiere que existen diferencias significativas (biológicas o experimentales) entre cada uno de estos subgrupos.

Mostramos a continuación el dendograma:



4.3. Heatmap

El heatmap parcial de las muestras y los bacterias del microbioma nos marcan unas primeras muestras con bacterias realmente sobreexpresadas y, a su vez, ciertas bacterias (1ª y 27ª) que podrían ser nuestras dianas para futuros estudios.



5. Discusión

En este análisis, se observaron ciertos patrones de agrupamiento entre las muestras, tanto en el análisis de componentes principales (PCA) como en el clustering jerárquico.

Los resultados sugieren que las muestras de diferentes grupos de estudio se segregan de manera significativa, lo que podría reflejar diferencias biológicas o experimentales entre ellas.

Sin embargo, también observamos que algunas muestras dentro de un mismo grupo muestran una variabilidad considerable en sus perfiles de metabolitos, lo que podría indicar la presencia de factores no considerados en este análisis (como la heterogeneidad biológica o variabilidad técnica en la medición).

5.1. Limitaciones

- **Tamaño de la muestra**: El análisis se basa en un número disminuido de muestras, metabolitos y bacterias, pues así tuvimos que realizarlo para mayor legibilidad, aunque asumimos que esto podría no ser representativo de la variabilidad total de las poblaciones.

- **Variables no incluidas**: Consideramos que no se tuvieron en cuenta todas las variables clínicas o experimentales que podrían influir en los perfiles de metabolitos.

5.2. Futuras Investigaciones

- Creemos sería útil explorar el impacto de diferentes condiciones experimentales sobre los perfiles metabólicos mediante un análisis multivariante más detallado.

- Recomendamos incluir más muestras para mejorar la robustez de los resultados, pero apoyándonos esta vez en computación en la nube, pues hemos tenido que recortar asimismo data porque las grandes dimensiones de nuestra matriz de conteo, nos llevaron a demoras excesivas en la generación de, por ejemplo, el mismo informe.

6. Conclusiones

- El análisis PCA y el clustering jerárquico revelaron patrones interesantes en los datos de metabolitos, mostrando agrupamientos significativos entre las muestras de diferentes grupos de estudio.

- Los resultados sugieren que las muestras podrían estar influenciadas por factores biológicos y experimentales, lo que debe ser investigado más a fondo.

- El análisis de los datos y la visualización mediante heatmaps y PCA proporcionan una primera aproximación a las relaciones entre las muestras y los metabolitos.

- Podemos iniciar un estudio de targeting con esas dos bacterias que hemos comprobado qu están sobreexpresadas en nuestro estudio (d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Peptostreptococcales;f__Acidaminobacteraceae;g__Fusibacter_A y d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Eubacteriales;f__Eubacteriaceae;g__Pseudoramibacter).

7. Referencias

Repositorio GitHub con el código y los datos utilizados en este análisis:
Ir a GitHub