PEC 1

David Rodríguez Temporal

Tabla de contenido

Abstract

En el presente trabajo se ha seleccionado un dataset de metabolómica relacionado con la microbiología, que consiste en mediciones de varias muestras de bacterias y varios controles. Se ha realizado un análisis exploratorio con R con tal de obtener una visión general de los datos, aplicando representaciones con histogramas y boxplot. Seguidamente se ha realizado un sencillo análisis mutivariado por Análisis de Componentes Principales y agrupación jerárquica. En las muestras analizadas, se ha observado que una de ellas difiere del resto, por lo que podría expresar características diferentes de los metabolitos analizados.

Objetivos

Objetivo general:

Realizar un análisis exploratorio de un dataset de metabolómica mediante R.

Objetivos específicos:

- Descargar un dataset de metabolómica y crear un objeto *SummarizedExperiment* en R.
- Realizar un pequeño análisis exploratorio de los datos, creando varias representaciones gráficas.
- Generar un repositorio en GitHub que contenga los datos utilizados y el código creado.

Métodos

Se ha buscado en metabolomicsWorkbench un dataset relacionado con microbiología, y se ha seleccionado el ST003777. Se trata de un dataset de metabolómica titulado "Separation and characterization of a clinically-derived *Staphylococcus epidermidis* strain HE23: Revealing its antibiotic resistome and metabolic potential". El objetivo de dicho studio era revelar determinantes relacionados con la Resistencia antibiótica y metabolitos tóxicos de esta cepa. Se han descargado los datos y metadatos y creado un proyecto de R.

Este estudio contiene dos tipos de datos, unos adquiridos en modo positivo por el instrumento y otro en modo negativo, pero para el ejercicio se han utilizado únicamente los resultados del modo positivo.

Se ha utilizado el programa R para el análisis de los datos, en el que se ha seguido el siguiente proceso: 1) Carga del paquete *BiocManager* y las librerías *metabolomicsWorkbenchR* y *SummarizedExperiment*; 2) Carga en el entorno de R el archivo de resultados (*datos*) y el de metadatos (*metadatos*); 3) Creación del objeto de clase *SummarizedExperiment* (*sum_exp*); 4) Análisis exploratorio de los datos. Para este análisis, se ha realizado principalmente un histograma, boxplot, Análisis de Componentes Principales (PCA) y un agrupamiento jerárquico mediante un dendrograma.

Resultados

El dataset seleccionado se trata de una lectura de 9 muestras: 3 de bacterias (C1, C2 y C3), 3 del blanco (H1, H2 y H3) y 3 de control de calidad (QC01, QC02, QC03). En el archivo de metadatos había mucha información y no se podía cargar correctamente en R, por lo que lo he modificado dejando sólo los datos de las muestras para una correcta interpretación.

Primero, hemos visto la cantidad de datos que obtenemos, que nos indica 743 variables (metabolitos medidos) en 9 muestras.

Para comenzar el análisis se realiza un histograma de cada muestra (Figura 1).

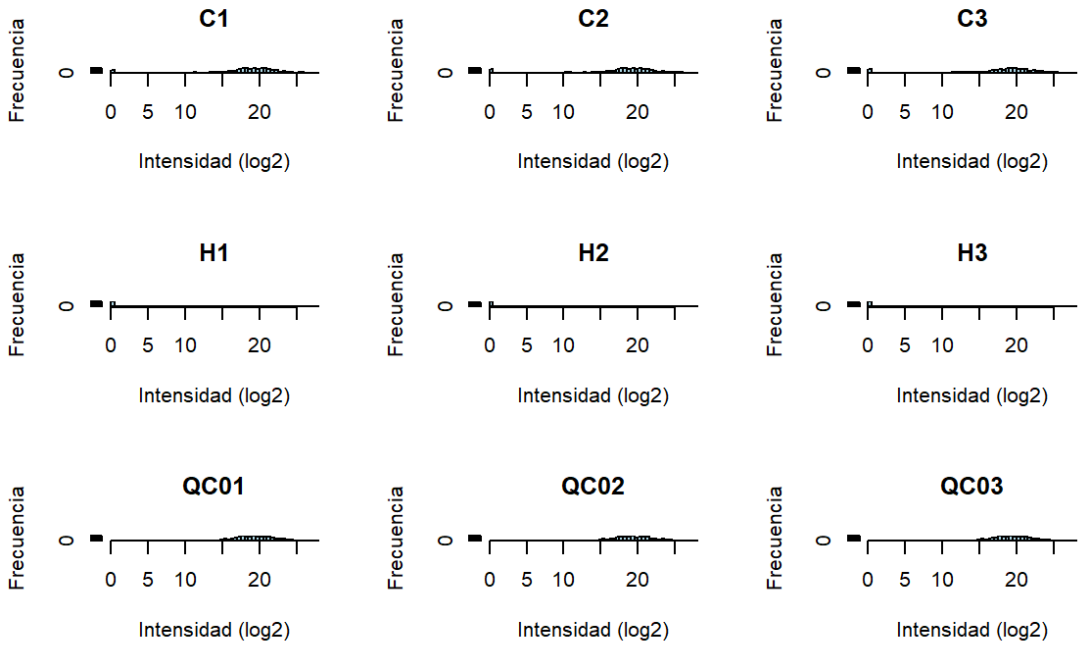

Figura 1. Histograma de intensidades para cada muestra del estudio.

Aunque no consigo ajustar correctamente el eje vertical para mejorar la visualización, podemos apreciar que las mediciones siguen una distribución similar en las muestras de bacterias (C) y en los controles de calidad (QC). Es esperable que en las muestras blancas prácticamente no veamos grandes intensidades y sólo una pequeña barra cercana al 0.

Podemos ver un enfoque parecido mediante boxplots, que realizo también de cada muestra (Figura 2).

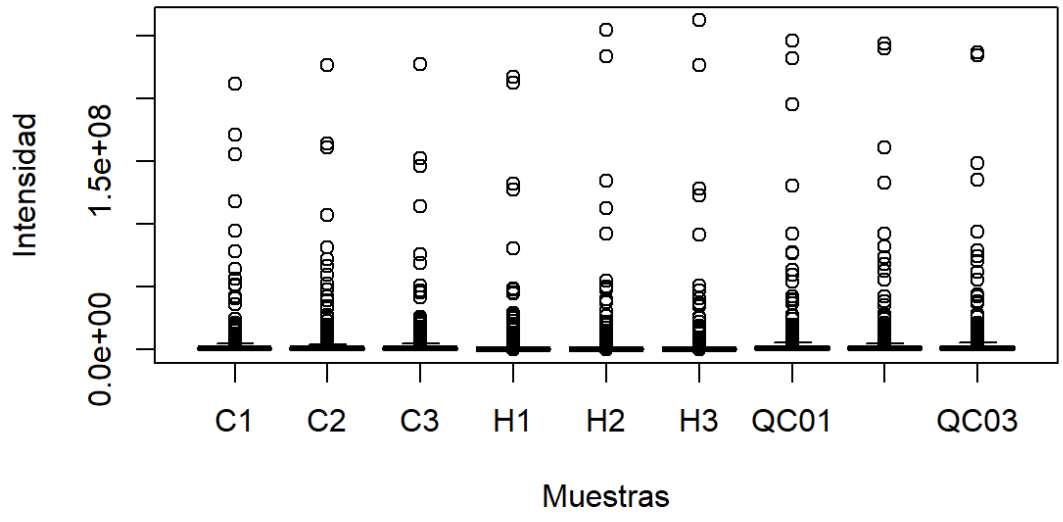

Figura 2. Boxplot de intensidades para cada una de las muestras.

En el caso del boxplot, vemos gran cantidad de valores cercanos al 0, lo que tiene sentido biológico, ya que, de todos los metabolitos medidos, no todos van a expresarse o detectarse a la misma vez, por lo que la mayoría tendrán intensidad baja. Seguramente, los más interesantes sean aquellos con alta intensidad y que, posiblemente, sean los característicos de cada muestra.

Posteriormente paso a realizar un análisis multivariado mediante agrupamiento, primero por un Análisis de Componentes Principal (PCA; Figura 3) y luego mediante un agrupamiento jerárquico (Figura 4).

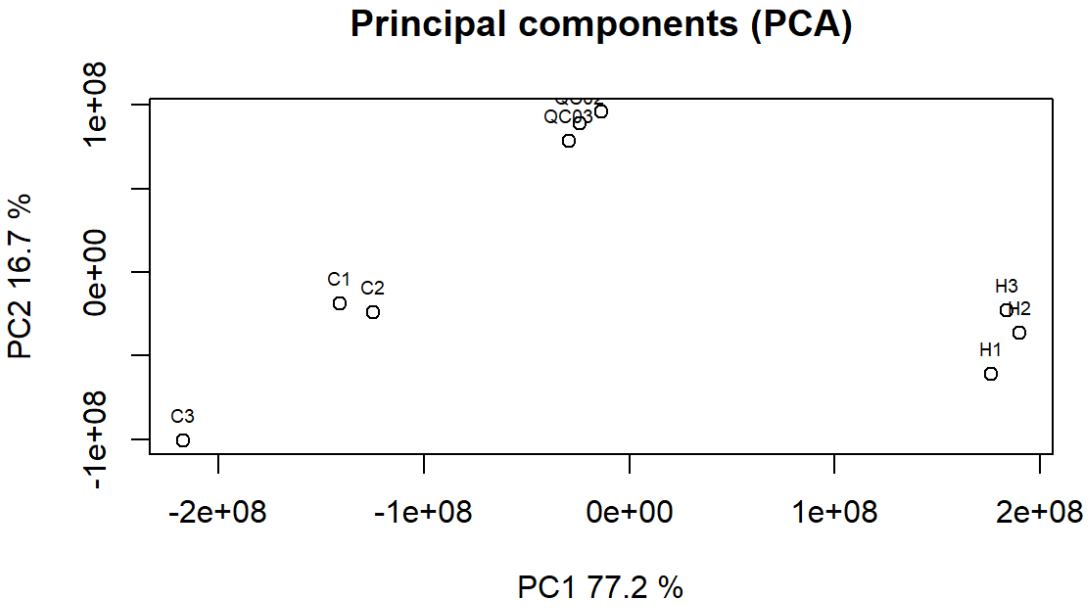

Figura 3. Análisis de Componentes Principales (PCA) de las 9 muestras del estudio. Se han representado solamente los 2 primeros componentes.

En el caso del PCA, en el que represento únicamente los dos primeros componentes, vemos un agrupamiento que era de esperar. Por un lado, los controles de calidad (QC) y los blancos (H) agrupan muy cercanamente entre ellos. Por otro lado, tenemos las 3 muestras bacterianas en otra zona. En ellas, podemos apreciar que parece que la muestra C3 es más diferente de las otras dos, ya que aparece más alejada de éstas. Vemos, además, que el PCA es una buena herramienta para estos datos concretos, ya que con sólo dos componentes logramos explicar el 93,9% de la variabilidad total.

Para terminar de visualizar y confirmar esta agrupación, realizo un dendrograma de agrupamiento jerárquico (Figura 4).

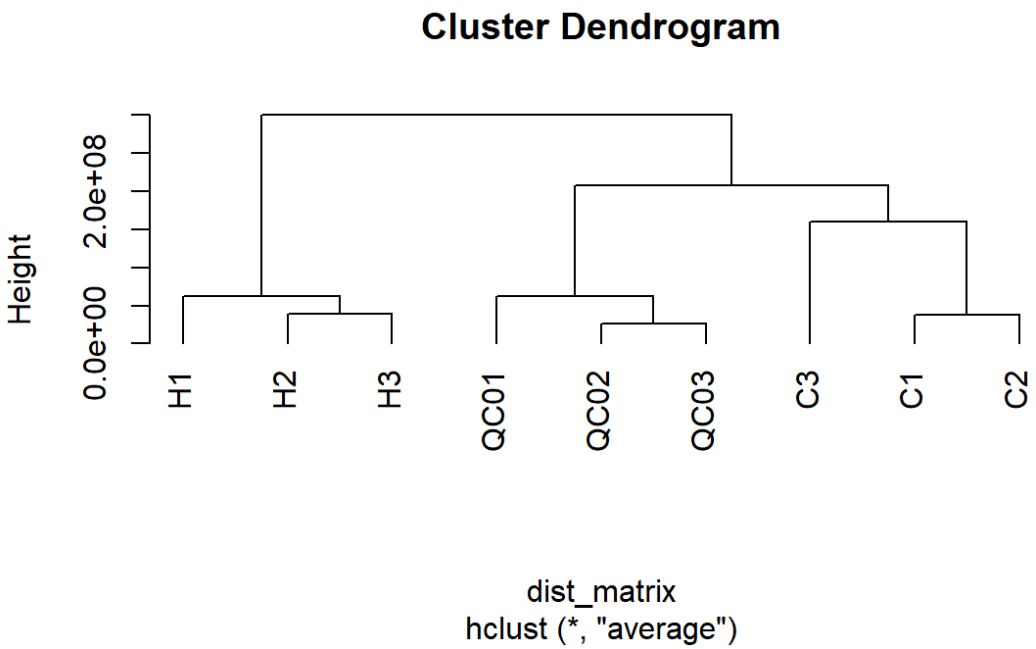

Figura 4. Dendrograma de las 9 muestras analizadas para visualizar su relación y agrupamiento jerárquico.

En este último caso volvemos a confirmar las observaciones previas con el PCA: entre las muestras de bacterias, la C3 es la más diferente, ya que se encuentra alejada del grupo que forman C1 y C2.

Discusión

Mediante este primer análisis, se ha podido poner en práctica como realizar un primer estudio de un conjunto de datos de metabolómica, aplicando distintas técnicas exploratorias en el entorno de R.

Respecto al estudio, se ha comparado un conjunto de 9 muestras. Tres de ellas consistían en mediciones en bacterias, mientras que el resto englobaba controles de calidad y muestras que representaban el blanco. A través del primer análisis por el histograma y el boxplot, se podía confirmar que las muestras de blanco prácticamente no mostraban mediciones de gran intensidad, lo que era esperable.

Mediante un segundo análisis algo más profundo, utilizando el PCA y el dendrograma, se ha podido confirmar la gran diferencia entre los distintos tipos de muestras incluidas en el estudio. Interesantemente, entre las 3 muestras bacterianas, se ha podido detectar una de ellas (C3) con mayor diferencia respecto a las otras dos.

Aunque no se ha profundizado más de forma práctica en el estudio, la continuación teórica se haría de la siguiente forma. Viendo que la muestra C3 es algo diferente del resto, lo interesante sería buscar qué metabolitos o qué mediciones son lo que la hacen diferente. De esta forma, podríamos comparar las mediciones de C3 con las de C1 y C2, realizando un test estadístico sobre cada metabolito para seleccionar aquellos con

una intensidad significativamente diferente, que bien se expresen en C3 y no en las otras muestras, y viceversa. De esta forma podríamos identificar aquellos que están marcando esta diferenciación entre las muestras y poder buscarle una explicación biológica.

Conclusiones

A través de este trabajo hemos podido realizar un pequeño análisis exploratorio de los datos de metabolómica mediante R. Las técnicas utilizadas nos han permitido confirmar la naturaleza de las muestras y detectar diferencias entre ellas mediante la comparación de los datos obtenidos en cada una.

Referencias

Con todos estos datos, creamos un repositorio en Github.

https://github.com/DavidRT23/Rodriguez-Temporal-David-PEC1