Análisis de datos Ómicos PEC_1

Javier Barrios

Abstract

Un aspecto muy importante a la hora de manejar datos ómicos es saber a qué datos nos estamos enfrentando y tener una visión general acerca de los mismos. Por eso, en este trabajo se va a realizar un análisis generalista de un conjunto de datos. Para ello se va a cargar un fichero con datos de metabolómica de un repositorio de github proporcionado por los profesores y se va realizar un breve análisis y preprocesado de los datos. Para ello se va a usar el lenguaje R y se van a proporcionar los resultados en un fichero Rmdarkdown(Rmd) subiéndolos a un repositorio en github.

Objetivos

Los objetivos de este trabajo son varios. * En primer lugar, familiarizarse con un dataset de metabolómica. Observar cómo pueden ser, qué tipo de datos se pueden observar, cómo se puede dar la descripción del estudio,etc. * El segundo objetivo es familiarizarse con el entorno de programación del lenguaje R, y es que, para analizar el conjunto de datos, se va a usar este lenguaje de programación, que es ampliamente usado en bioinformática. * En tercer lugar, dentro del lenguaje de programación R, está el objetivo de familiarizarse más concretamente con el paquete Bioconductor. Este paquete es muy potente si hablamos del campo de la bioinformática y aprender a manejarse con él es un paso importante. * El cuarto objetivo es ver el proceso de análisis de un conjunto de datos. Cómo se comienza con unos datos en bruto y prepararlos para poder analizarlos, así como el propio proceso de análisis de los mismos datos.

Métodos

En este estudio se ha utilizado el conjunto de datos human_cachexia. La caquexia es un trastorno metabólico que deja a la persona en un estado de desnutrición, fatiga y debilidad generalizada. Además, implica pérdida de masa muscular y de peso.

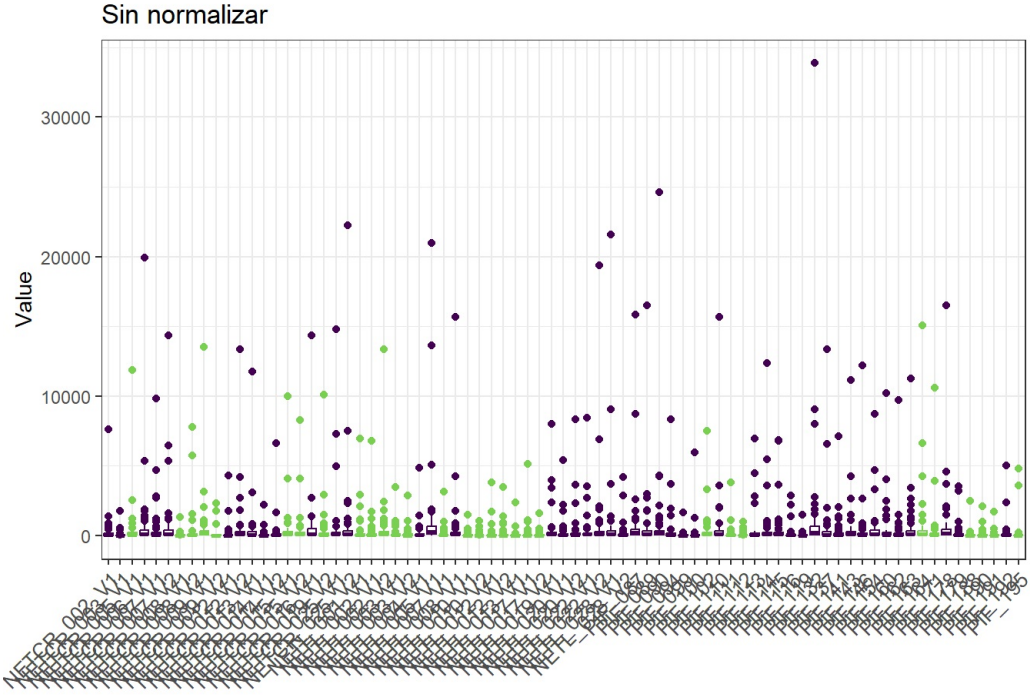
En primer lugar se han cargado los datos. Tras cargarlos se ha visto que este dataset contiene 77 filas y 65 columnas. Las filas son las muestras, es decir, los distintos pacientes a los que se les ha realizado el estudio metabólico. De las 65 columnas que se tienen, la primera es el ID o identificador del paciente, la segunda es si es un paciente de control o es un paciente afectado por el trastorno metabólico y las otras 63 son distintos datos metabólicos que se han obtenido en el estudio.

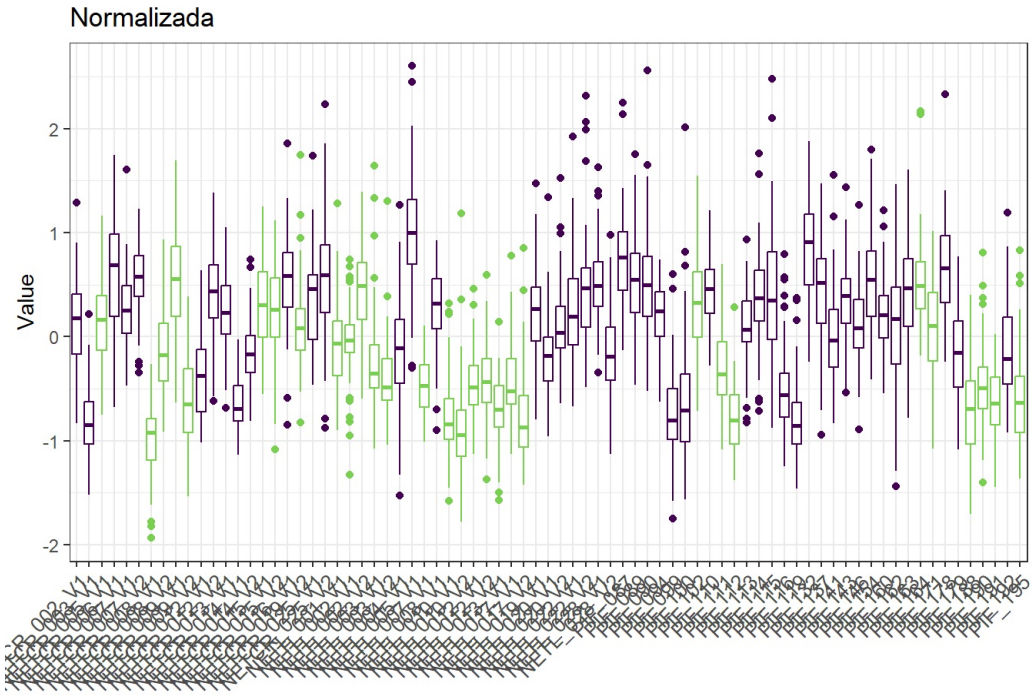
Lo primero que se realiza es una observación de los datos para ver si todo es correcto. Lo primero es ver que no hay ningún elemento que sea NA, ya que es una cosa que puede ocurrir en estos casos; sin embargo, no se vio ninguno, ya que al parecer este dataset ya se había limpiado de este tipo de valores previamente.

A partir de aquí queremos hacer una exploración de los datos. Para ello se va a hacer uso del paquete POMA dentro del paquete de Bioconductor. Este paquete nos proporciona herramientas para el preprocesado de nuestro conjunto de datos. Una particularidad de las funciones de este paquete es que funcionan con objetos de la clase *SummarizedExperiment*, con lo que una de las primeras cosas que se deben hacer es crear un objeto de esa misma clase.

Para ello se carga el fichero, se cogen solamente las columnas numéricas y se trasponen para conseguir el formato requerido por los assays de dicha clase, donde se tienen en filas los metabolitos y en las columnas los pacientes. Las columnas PatientID y MuscleLoss, se pasan a la clase como datos de las columnas(colData). Con eso ya se pueden usar las funciones del paquete POMA.

Una vez visto que se tiene todo ordenado de la forma apropiada, se normalizan las variables para eliminar posibles bias. En las siguientes figuras se pueden ver los datos antes y después de la normalización.



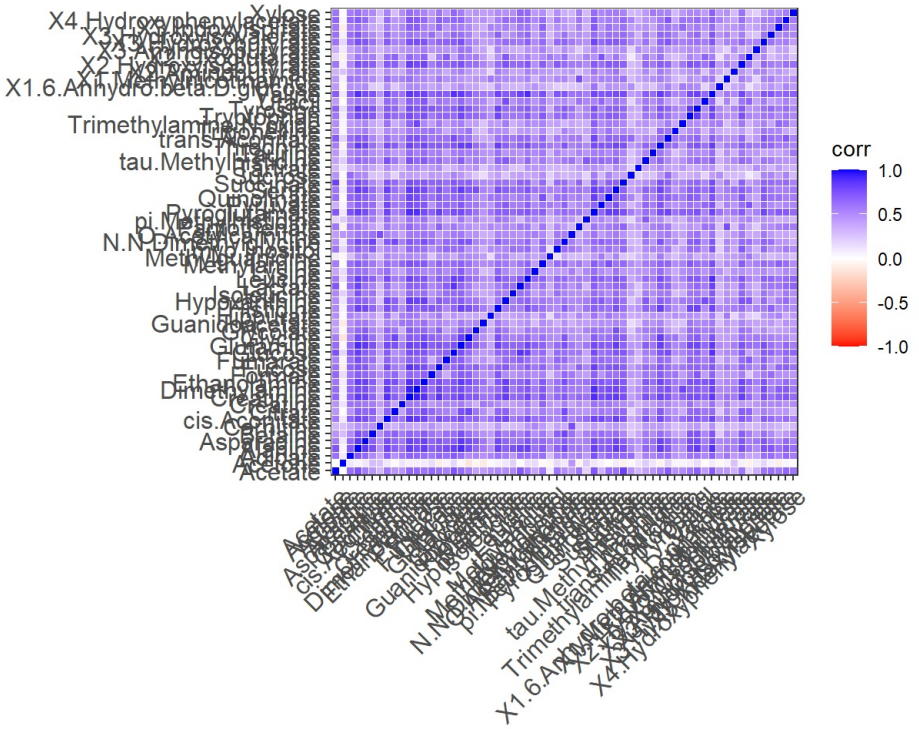


Otro ajuste que podemos hacer es la eliminación de outliers. Los outliers son muestras que se alejan mucho del resto de elementos del mismo tipo y que, por tanto, sus valores pueden deberse a errores en la lectura o en el procisimiento de adquisición de los mismos. En nuestro caso se ha considerado su eliminación.

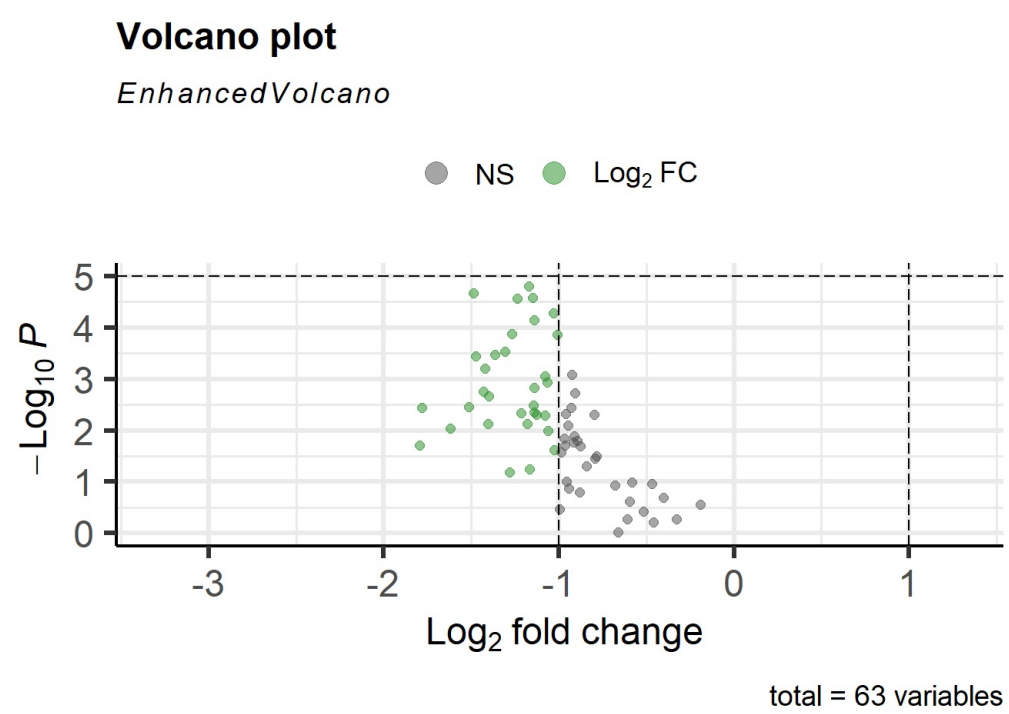
Una vez pre-procesados los datos vamos a ver los análisis estadísticos. En primer lugar veremos los p-values de cada variable, y veremos su nivel de significación. También realizaremos un PCA como técnica de análisis multivariante para ver si se puede ver algún agrupamiento y se pueden clasificar más fácilmente los dos grupos.

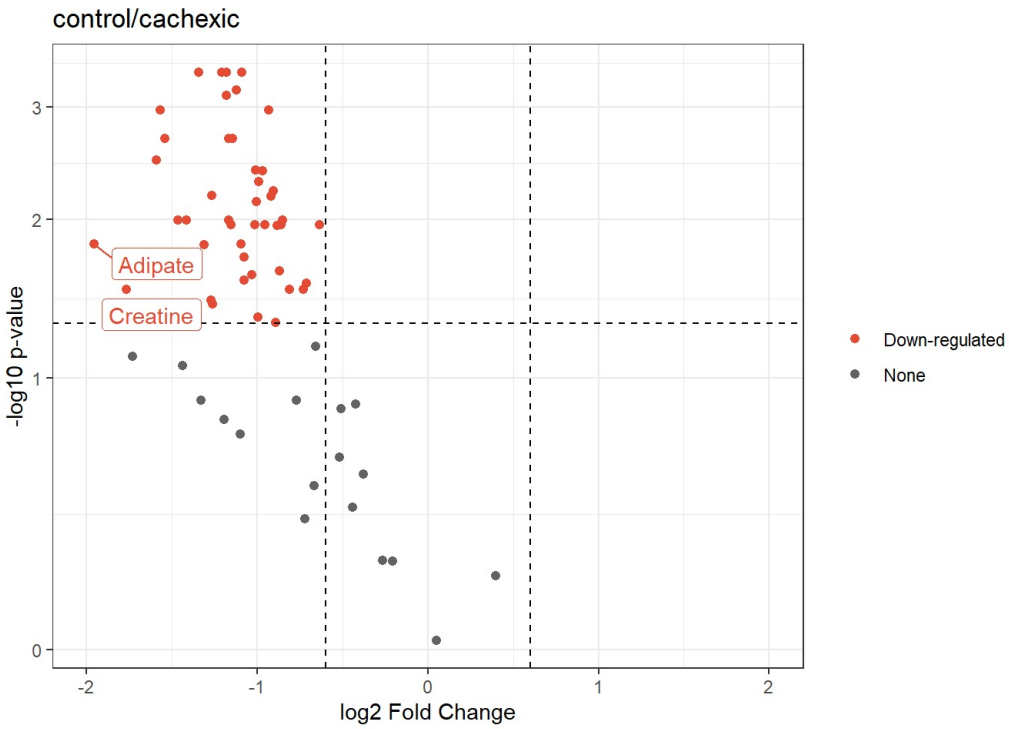
Resultados

Lo primero que queremos ver es si las variables están correlacionadas entre sí. Para ello vamos a ver las los valores de correlación. Si no hubiésemos normalizado los datos, usaríamos la matriz de covarianzas. Al ver la matriz de correlacione, podemos ver que muchas variables están correlacionadas entre sí a unos valores altos, por tanto, lo que se va a hacer es reducir la dimensionalidad para ver si podemos explicar mejor estos datos.

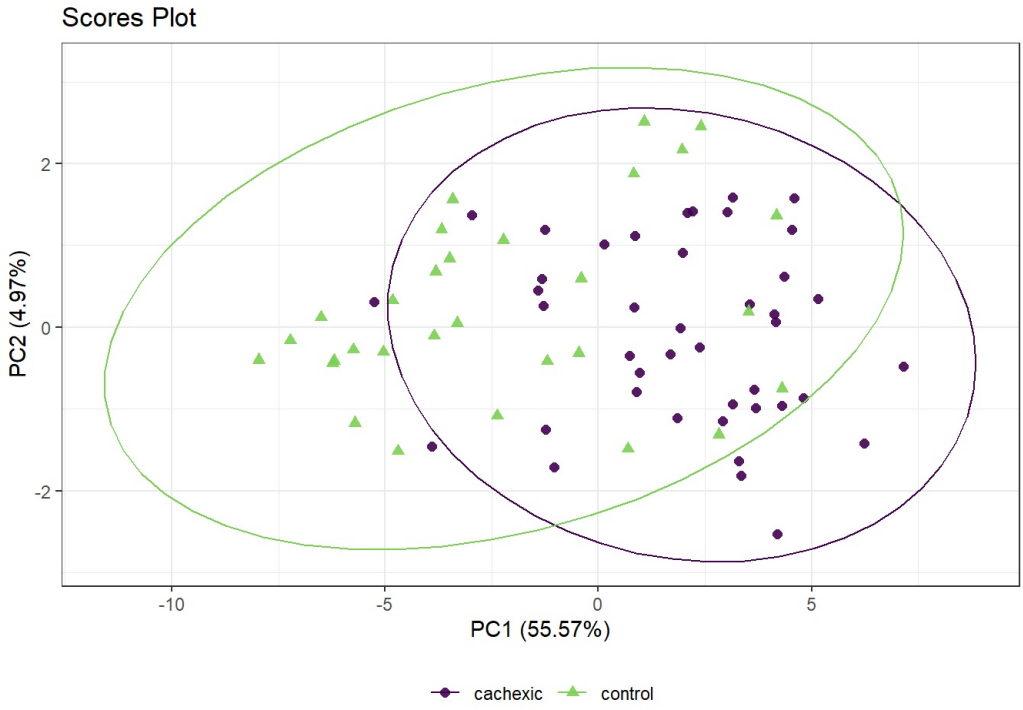


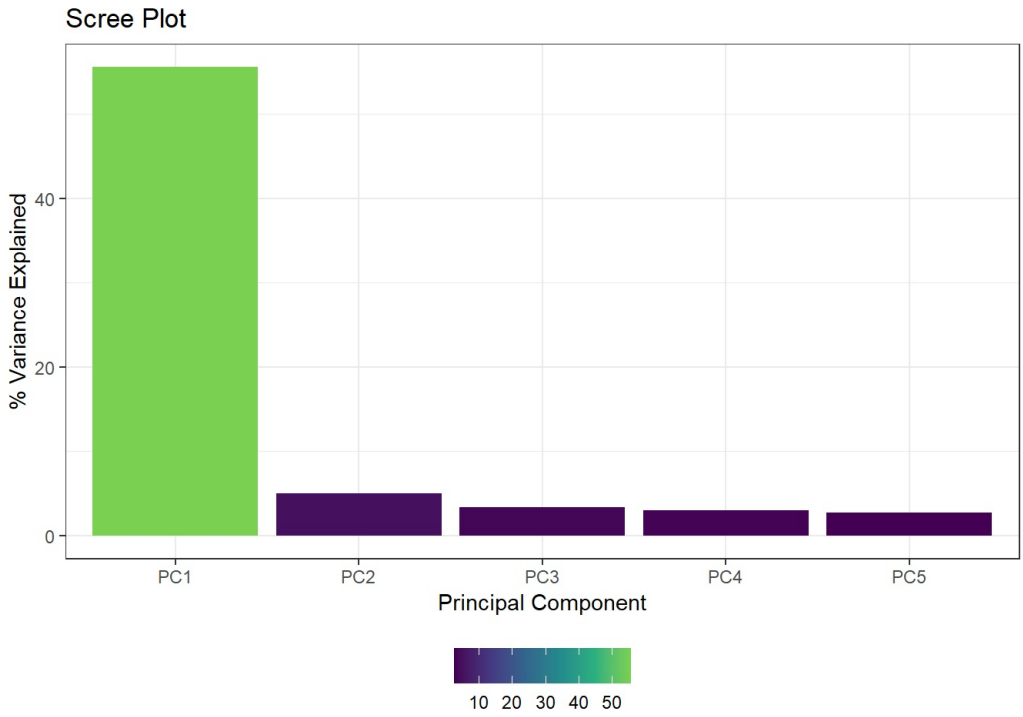
Con el volcano plot podemos ver si hay diferencias significativas entre los grupos de control y test, así como ver si hay diferencias significativas. Hice 2 pruebas, tanto con la función PomaVolcano, como con EnhancedVolcano, para ver si se podían ver cuáles son los diferentes, pero los labels parecen ser demasiado largos.

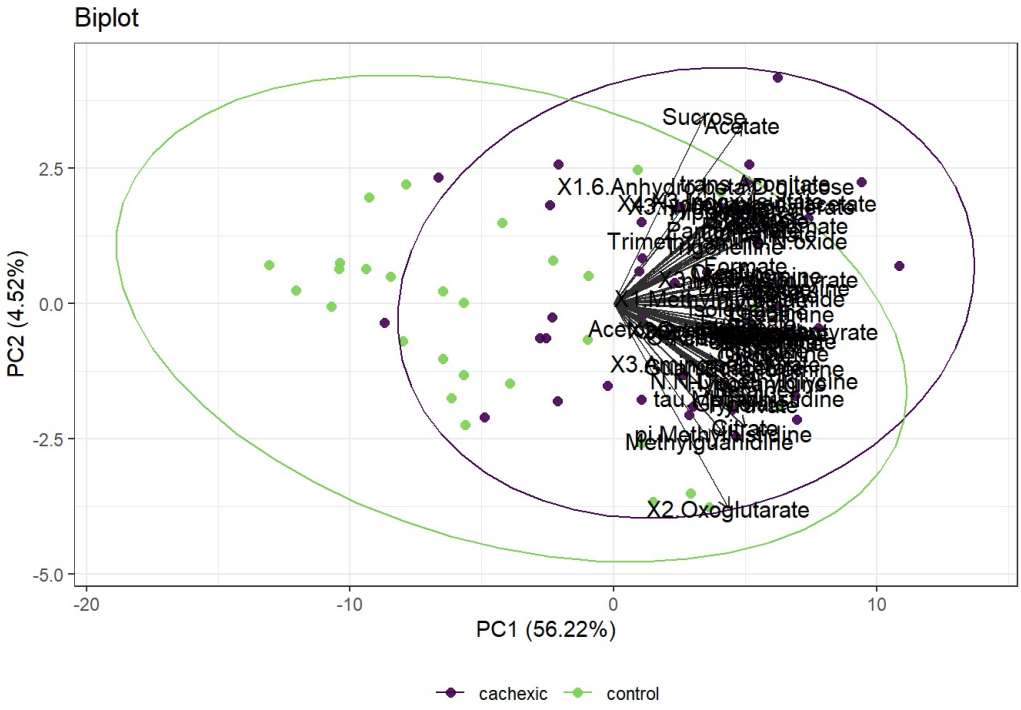




En vista de que esto no nos aclara gran cosa, salvo que sí que hay variación y significación entre los grupos. Así pues, con el PCA se va a intentar ver si se pueden agrupar de una manera sencilla. En el resultado del PCA se puede ver por las elipses dibujadas que se puede diferenciar entre los grupos, sin embargo, están bastante entremezclados. Cabe destacar que la primera componente principal nos presenta una variación del 56%, tal y como se muestra con el scree plot, pero que la segunda ya se ve reducida al 4%, con lo que se podría decir que con la primera componente principal se describe bien la variabilidad de nuestro conjunto de datos.


Scores Plot
PC2 (4.97%)
2
0
-2
-10
-5
0
5
PC1 (55.57%)
cachexic
control


Scree Plot
% Variance Explained
40
20
0
PC1
PC2
PC3
PC4
PC5
Principal Component
10 20 30 40 50



Discusión y Conclusiones

Este dataset de 77 muestras de 63 variables tiene dos grupos diferenciados, que son test(afectados por caquexia) y control. A pesar de tener más muestras de uno que de otro se puede considerar pareado, por lo que, a falta de más datos debido a la anonimización de los pacientes, se puede considerar un estudio bien planteado si la variabilidad entre pacientes es la correcta. Del fichero se puede ver que hay metabolitos que tienen diferencias significativas si están en test o en control, y que se pueden llegar a agrupar las muestras de forma apropiada al aplicar PCA.

Referencias

Enlace al repositorio de github https://github.com/barriosjavier/Omicas_PEC.git (https://github.com/barriosjavier/Omicas_PEC.git)