Informe PEC1

Asier Iturrate

Noviembre 2024

Contents

Abstract

En este trabajo hemos generado un contenedor SummarizedExperiment a partir de un dataset con datos metabolómicos de un estudio entre pacientes con cachexia e individuos control. Tras importar los datos y crear el objeto SummarizedExperiment, hemos realizado un análisis exploratorio de los datos mediante PCA, clusterización jerárquica y el uso de heatmaps. Este análisis nos ha indicado que podría ser conveniente la eliminación de algunas muestras del dataset original, por lo que hemos creado un nuevo contenedor con este nuevo subset de datos. El análisis exploratorio sobre este nuevo dataset parece indicar que hay un grupo de pacientes que se comportan de manera diferente al resto de individuos del grupo, lo que podría ser indicativo de alguna razón biológica.

Introducción y objetivos

El dataset elegido para este trabajo es el de **human_cachexia**. La caquexia es un síndrome metabólico complejo asociado a una enfermedad subyacente (como el cáncer) y caracterizado por una pérdida de masa muscular con o sin pérdida de grasa (Evans et al., 2008). En este dataset encontramos los resultados de un análisis metabolómico realizado a partir de muestras de orina de 77 individuos, de los cuales 47 son pacientes con caquexia y 30 son individuos control.

Los objetivos de este trabajo son los siguientes:

1. Generar un contenedor del tipo SummarizedExperiment a partir del dataset human_cachexia.
2. Realizar un análisis exploratorio de nuestro dataset.
3. En caso de que sea necesario, realizar un pre-procesamiento de los datos.
4. Identificar mediante el análisis exploratorio posibles sub-grupos dentro de los grupos *caquexia* y *control*.

Materiales y métodos

Puesto que se trara del informe del trabajo, no se ha adjuntado ninguna línea del código utilizado para este estudio. Únicamente se ha adjuntado el código utilizado para la generación del contenedor SummarizedExperiment puesto que se trata de una parte importante del trabajo. El código completo para el desarrollo completo de este trabajo se puede encontrar en el repositorio de Github que se destaca al final de este informe.

Versión de R y RStudio

Este trabajo se desarrolló en R 4.4.2 y Rstudio RStudio 2024.04.1+748 "Chocolate Cosmos" Release

Descarga del dataset, importación a R y pequeñas modificaciones

Se descargó el fichero csv **human_cachexia** del repositorio de Github metaboData https://github.com/nutrimetabolomics/metaboData.git. Una vez situado el fichero en el directorio de trabajo, para su importación a R, se utilizó la función *read.csv()*. Para facilitar la lectura e interpretación de resultados, se cambió el ID de cada uno de los individuos por "paciente#" o "control#" en función del grupo al que pertenecían.

Creación de un objeto SummarizedExperiment

Para la creación del objeto SummarizedExperiment se utilizó la función *SummarizedExperiment()* del paquete SummarizedExperiment. Puesto que se consideró uno de los objetivos del trabajo, la manera en la que se generó el objeto SummarizedExperiment se encuentra explicada de manera más detallada en el apartado **Resultados**.

Principal Component Analysis (PCA)

Para el análisis de componentes principales o PCA se utilizó la función *prcomp()*. La visualización de este análisis se realizó con la función *autoplot()* del paquete de R ggfortify tal y como se muestra en el tutorial https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html

Clusterización jerárquica

La clusterización jerárquica se realizó con la función *hclust()*. Para ello, previamente se normalizaron los conteos de cada uno de los metabolitos con la función *scale()*. Para la visualización del dendograma se utilizó la función *plot()*.

Mapa de calor o *heatmap*

Para la creación y visualización de un *heatmap* con clusters se utilizó la función *pheatmap()* del paquete *pheatmap*.

Resultados

Importación del dataset y análisis inicial de nuestros datos

El dataset **human_cachexia** fue descargado e importado como ha sido descrito en el apartado Materiales y métodos. Tras realizar una breve visualización de las primeras entradas y de la estructura de nuestro dataset podemos observar que se trata de un *dataframe* con 77 observaciones, correspondientes a los análisis metabolómicos realizados sobre las muestras de los 77 individuos (controles y pacientes). Además, encontramos 65 variables. Las dos primeras de tipo *character* corresponden al identificador de cada paciente y al grupo al que pertenecen ("cachexic" o "control"), respectivamente. Las otras 63 variables restantes se corresponden con los valores numéricos obtenidos para cada uno de los metabolitos medidos en este estudio. Como se ha descrito anteriormente, para facilitar el análisis y la interpretación de los resultados, se modificaron los IDs de los individuos por control# y paciente#.

Creación de un objeto SummarizedExperiment

Para la generación de nuestro SummarizedExperiment utilizamos tres componentes y una serie de metadatos con información sobre el SummarizedExperiment:

1. **assays**: Se generó una matriz de conteos a partir del dataframe original con los valores correspondientes de conteos de cada metabolito (dispuestos en filas) para cada uno de los individuos (dispuestos en columnas). Al assay correspondiente a nuestro dataframe lo llamamos "metabolitos".
2. **colData**: Se generó una matriz de dos columnas, donde en la primera de ellas se encontraba el ID de cada uno de los individuos del estudio y a su lado en la segunda columna al grupo al que pertenecían ("control" o "cachexic").
3. **rowData**: Se generó un vector con los identificadores o nombres de cada uno de los metabolitos que se han medido.
4. **metadata**: Se creó una lista con información acerca del SummarizedExperiment (autor, contacto, título del proyecto...).

Puesto que se considera una parte importante del trabajo, se adjunta en el informe el código utilizado para la generación de este SummarizedExperiment

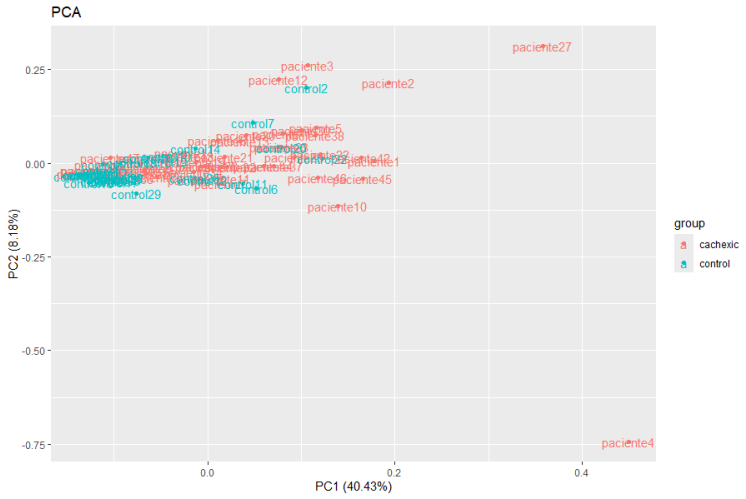
```
library(SummarizedExperiment)

# Cargamos el dataset
human_cachexia <- read.csv("human_cachexia.csv", header = TRUE)

# Visualizamos dataset
head(human_cachexia)

# Hacemos un resumen de la estructura de nuestro dataset
str(human_cachexia)

# Cambiamos los ID de los pacientes por paciente+numeros
human_cachexia$Patient.ID[1:47] <- paste0("paciente", 1:47)

# Cambiamos los ID de los controles por control+numero
human_cachexia$Patient.ID[(nrow(human_cachexia)-29):nrow(human_cachexia)] <- paste0("control", 1:30)

## Generamos matriz de conteos

# Generamos la matriz transpuesta de nuestro dataset con los datos de los metabolitos
myCounts <- data.frame(t(human_cachexia[3:65]))
```

```r
# Añadimos el nombre de los individuos a cada columna
colnames(myCounts) <- human_cachexia[, 1]

# Visualizamos la matriz de conteos
head(myCounts)


## Generamos matriz de covariables

# Extraemos las dos primeras columnas de nuestro dataset original
myGroups <- data.frame(sampleName = human_cachexia[, 1],
                       group = human_cachexia[, 2],
                       row.names = 1)

# Visualizamos matriz de covariables
head(myGroups)


## Generamos vector con identificadores

# Guardamos los nombres de los metabolitos
myMetabolites <- c(colnames(human_cachexia[3:65]))


## Generamos información del SummaryzedExperiment
## Generamos información del SummaryzedExperiment
myInfo <- list(myName = "Asier Iturrate",
               myInstitution = "UOC",
               myContact = "aiturrates@uoc.edu",
               myTitle = "SummarizedExperiment generado para la PEC1",
               Description = "Objeto original creado a partir del dataset human_cachexia" )


## Generamos SummaryzedExperiment

# Creamos SummarizedExperiment
mySE <- SummarizedExperiment(assays = list(metabolomica = myCounts),
                             colData = myGroups,
                             rowData = myMetabolites)

# Añadimos metadatos
metadata(mySE) <- myInfo

# Visualizamos información de SummarizedExperiment
show(mySE)

# Guardamos el SummaryzedExperiment con formato .Rda
save(mySE, file = "mySE.Rda")
```

Análisis exploratorio de nuestro dataset

Análisis por PCA

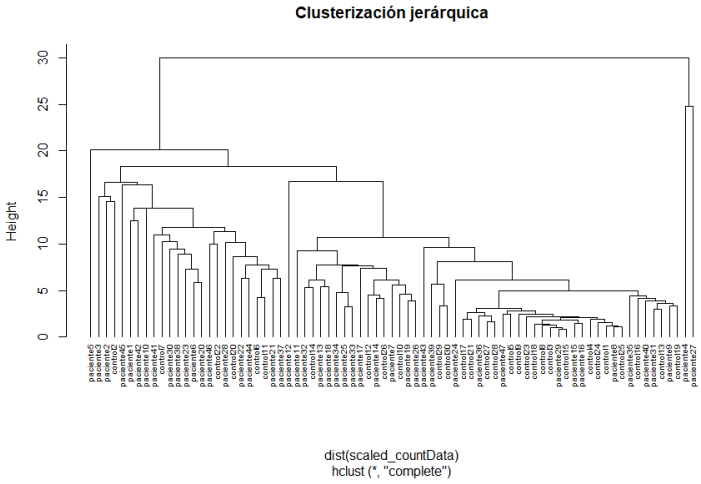
En primer lugar, se extrajo la información correspondientes a los datos de expresión a partir del objeto SummarizedExperiment, así como los datos correspondientes a los grupos. Para realizar el análisis de PCA se normalizó la expresión de cada uno de los metabolitos. El análisis de componentes principales o PCA arrojó el siguiente resultado:



Podemos observar en primer lugar que la suma de ambos componentes principales PC1 y PC2 no llega al 50%. Esto puede ser debido a una alta variablidad interna o que no haya una tendencia clara entre las muestras. Sin embargo, si analizamos la distribución de las muestras, parece que hay una mayor tendencia de los controles a agruparse a la izquierda del gráfico (azules) y una mayor tendendia de pacientes a agruparse a la derecha (rojos). No es una separación estricta, pero podemos observar cierta tendencia. Por otro lado, podemos observar que las muestras **paciente4** y **paciente7** se separan más hacia la derecha en comparación con los mismos individuos de su grupo y del conjunto general de individuos (pacientes y controles).

Análisis por clusterización jerárquica

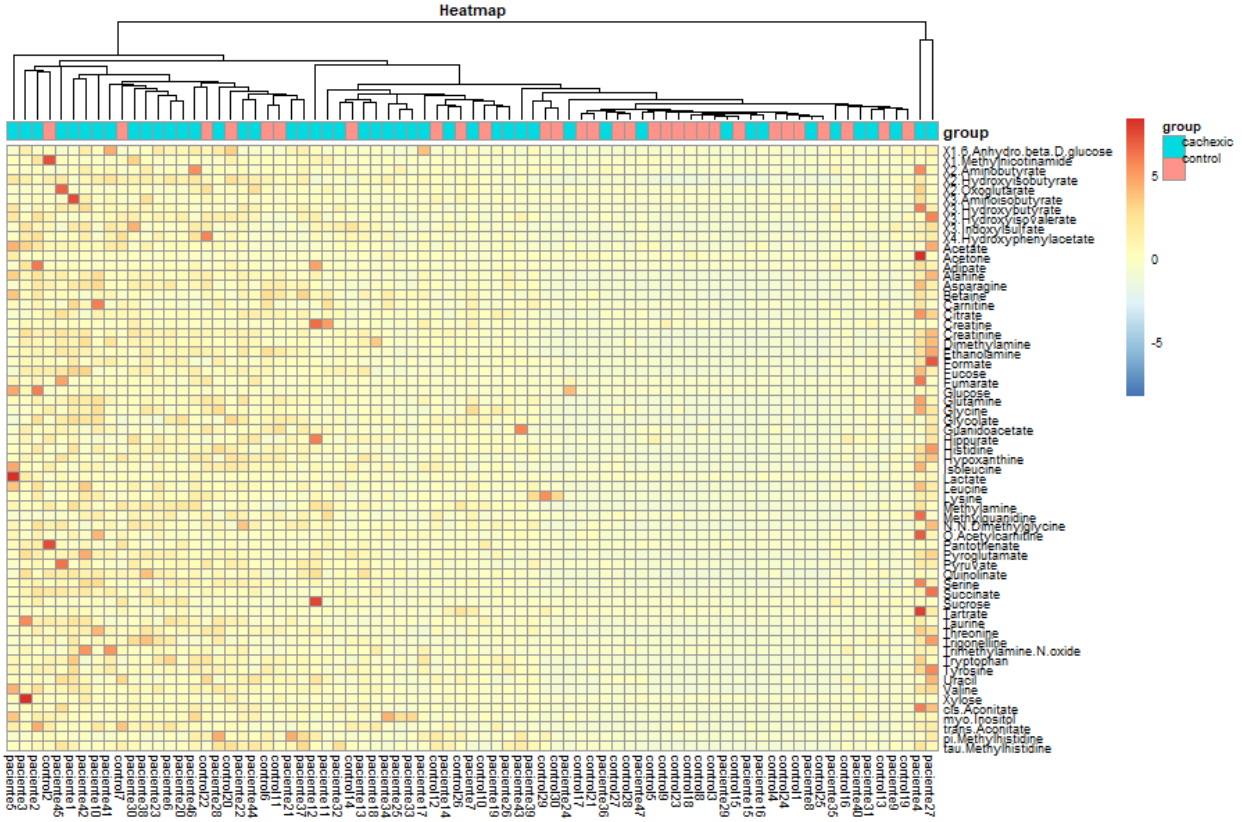
Para ver si existe alguna agrupación entre controles y pacientes (o incluso descubrir alguna nueva dentro de cada grupo) realizamos una agrupación jerárquica usando los conteos de los metabolitos normalizados y representamos los resultados en un dendograma. El análisis reveló los siguientes resultados:



Una vez más, podemos observar que los **pacientes4** y **paciente27** se agrupan muy alejados del resto de muestras. Adicionalmente, vemos que **paciente5** tampoco parece agrupar de la manera adecuada con el resto de muestras.

Análisis por heatmap con clusterización jerárquica

Para poder visualizar un poco mejor las diferencias observadas en estos 3 pacientes y debido a que el número de metabolitos analizados es un número razonable, decidimos realizar un análisis visual rápido de los conteos normalizados de los diferentes metabolitos de nuestro dataset. Para ello, representamos un heatmap con una clusterización jerárquica y vimos lo siguiente:

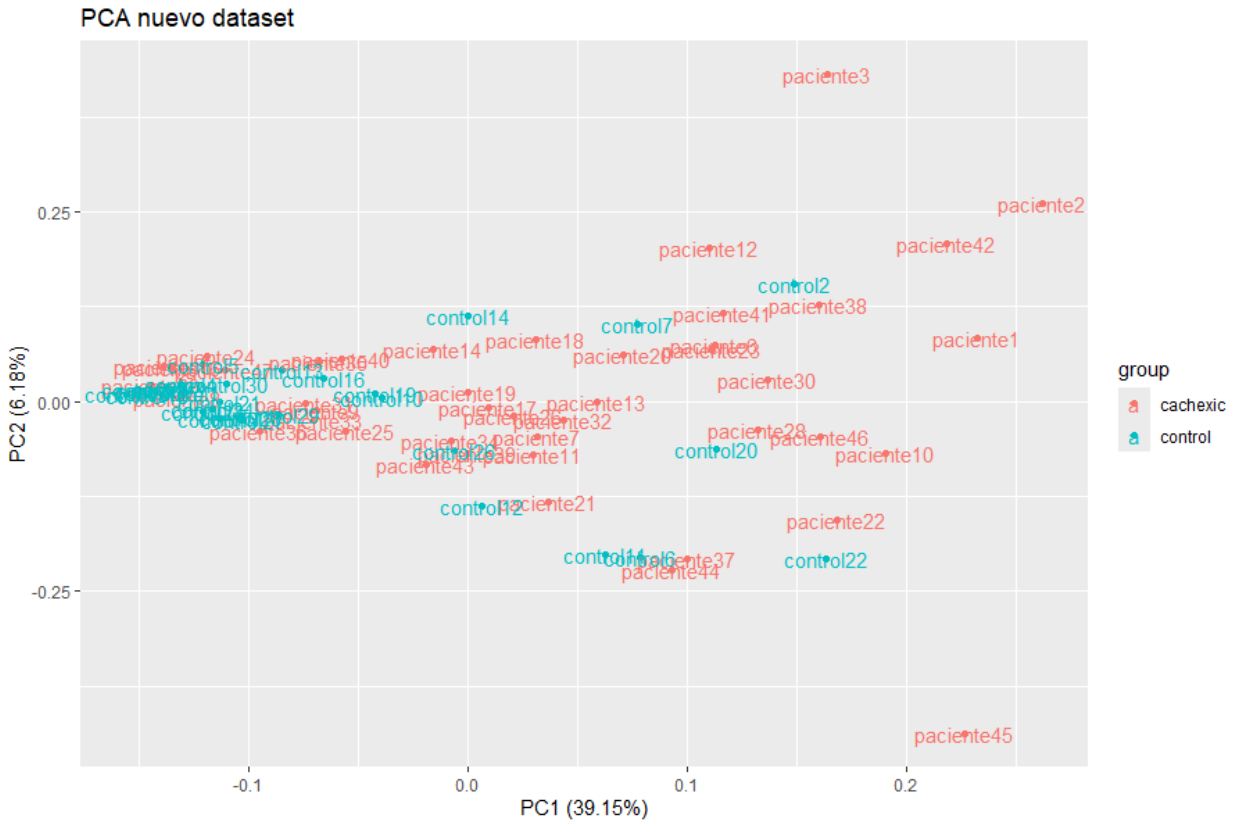


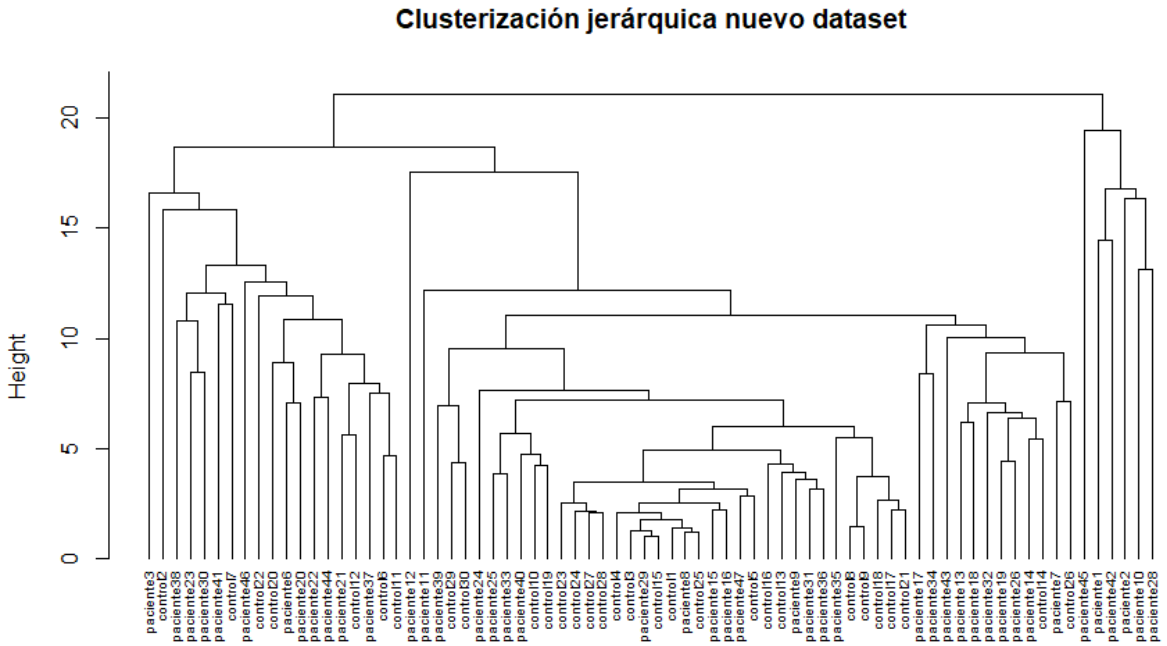
Este análisis nos reveló que los **paciente4** y **paciente27** presentan unos mayores niveles de una gran cantidad de metabolitos en comparación con el resto de muestras del dataset, tanto de controles como de pacientes. Por ello, se agrupan de manera diferente al resto de muestras. Parece pasar algo parecido con el *paciente5* que tampoco agrupa de manera firme con otros individuos del dataset. De todas maneras, podemos observar una vez más como los controles (esta vez representados con un cuadrado rojo) tienden a agrupar hacia la derecha del heatmap, mientras que los pacientes de cachexia tienden a agrupar hacia la izquierda (cuadrados azules).

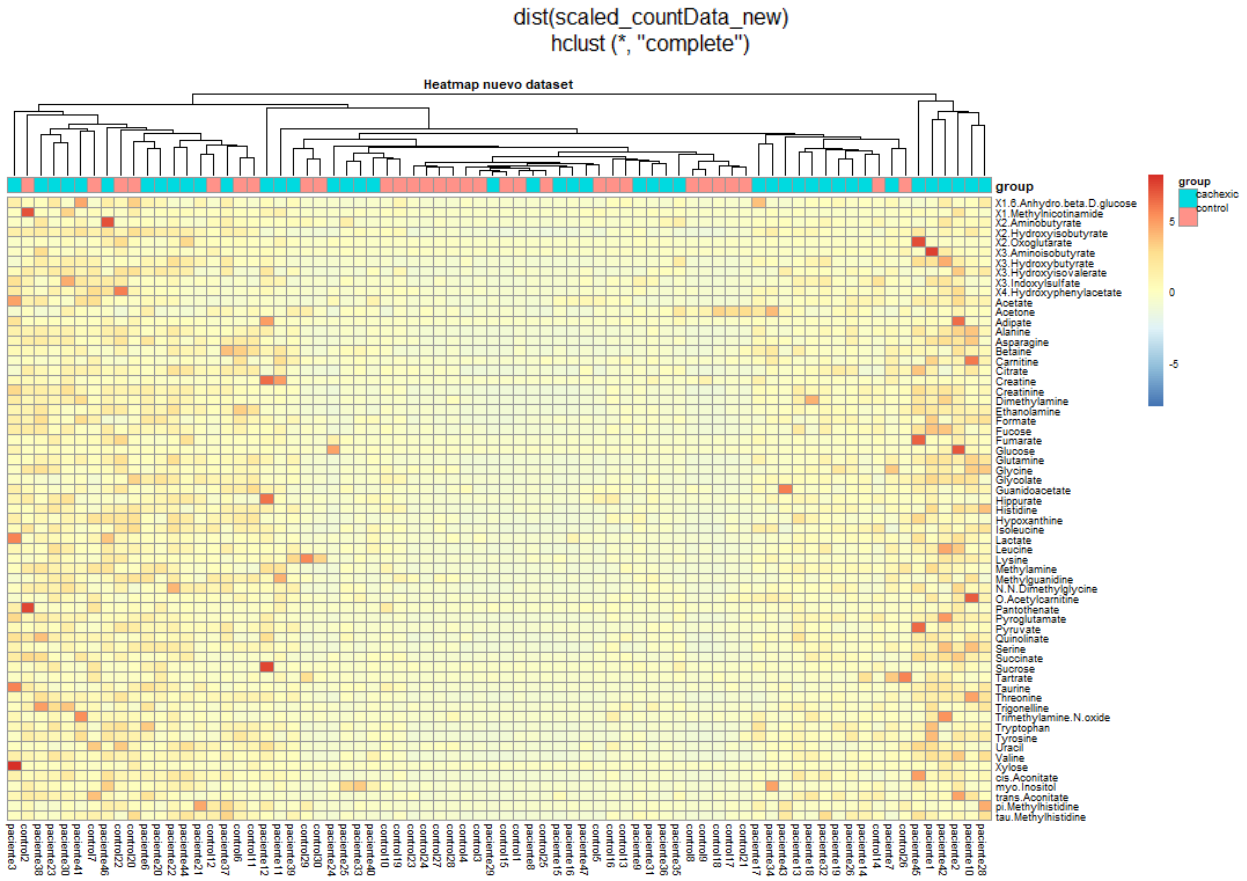
Generación y análisis de un nuevo dataset filtrado

Puede que las diferencias observadas en nuestro dataset con respecto a los niveles de metabolitos de los pacientes 4, 5 y 27 sean reales. Sin embargo, puesto que no tenemos información adicional de estos datos, puede que las diferencias que se observan sean debido a un artefacto o a un error en el procesado de las muestras, por ejemplo. Además, ya que otro de los objetivos del trabajo sería encontrar posibles "sub-grupos" dentro del grupo de controles o de pacientes, vamos a eliminar los datos correspondientes a estas muestras para así poder ver otras diferencias más sutiles dentro de nuestro dataset.

Para ello, retiramos de nuestro SummarizedExperiment los datos correspondientes a estos individuos y repetimos el análisis exploratorio:







En este nuevo análisis observamos que en el gráfico PCA todas las muestras se encuentran más cercanas (lo

cual indicaría que no hay diferencias muy grandes entre los grupos cachexia y control), pero seguimos viendo esa tendencia de los controles a agruparse a la izquierda y los pacientes a la derecha del gráfico PCA.

En cuanto a la clusterización de las muestras, vemos que esta vez, todas ellas parecen estar más agrupadas entre sí en comparación con el dendograma de todo el dataset sin filtrar. Además, si vemos en conjunto los datos del dendograma y del heatmap, vemos que a la derecha de ambos gráficos aparece un grupo de pacientes (pacientes 1, 2, 10, 28, 42 y 45) que tienden a agrupar entre ellos. Además, vemos que en este nuevo dataset son los que mayor niveles de determinados metabolitos presentan. Se trata más sobre este razonamiento en el apartado de conclusiones y discusión. Por último, seguimos viendo cierta tendencia a agrupar de los controles (cajass rojas), en este caso en el centro del gráfico, mientras que la mayoría de los pacientes (cajas azules) se agrupan en las esquinas de los gráficos. No sería una agrupación absoluta, pero vemos cierta tendencia.

Conclusiones y discusión

En este trabajo hemos realizado un análisis exploratorio de un dataset que contenía información de datos de un análisis metabolómico de un grupo de pacientes de cachexia en comparación con un grupo de controles. Sin embargo, una limitación a la hora de analizar y discutir los posibles resultados es que no disponemos de demasiada información técnica y/o biológica sobre estas muestrs.

El análisis que hemos realizado ha mostrado que parece existir cierta agrupación entre las muestras de los controles y de las muestras de pacientes. No es una agrupación absoluta, puesto que vemos que hay controles que se encuentran más cercanos a los subgrupos de pacientes que a su propio grupo, y viceversa. Esto puede ser debido a que únicamente estamos teniendo en cuenta la variable "enfermo" o "no enfermo". No tenemos información sobre la edad de los sujetos, sexo, etiología de la cachexia... u otro tipo de información biológica que nos pueda estar informando sobre estas posibles diferencias que haga que no agrupen al 100% las muestras control y las paciente. Por otro lado, esto también es esperable, ya que se trata de un dataset con un número considerable de individuos (47 pacientes y 30 control) y que por lo visto en los heatmaps, no parece que haya una tendencia o una especie de biomarcador/biomarcadores fuerte que haga que se agrupen las muestras en función de si son pacientes o no.

Por otro lado, hemos decidido repetir el análisis con un nuevo subset de datos eliminando los valores correspondientes a tres pacientes. Como ya se ha mencionado anteriormente, puesto que no disponemos de información adicional de estas muestras, ni de otro tipo de datos, hemos decidido eliminarlos en el nuevo subset puesto que hemos asumido que podrían trartarse de arterfactos a la hora de recoger las muestras, procesarlas, que haya ocurrido algún evento sobre ellas... Sin embargo, la decisión de eliminarlas debería ser confirmada con por ejemplo, si estos pacientes tienen un fenotipo más grave y por eso presentan esa tendencia, o si hay otras muestras que hayan sido procesadas ese mismo día y pueda descartar que se deba a un artefacto de la muestra.

Para finalizar, hemos observado en el nuevo dataset que parecen aparecer una serie de sub-grupos dentro del grupo cachexic. Esta formación de clusteres podría tener algún significado biológico como que por ejemplo esos pacientes tengan un fenotipo más grave en comparación con el resto de pacientes, o su cachexia venga originada por una patología subyacente diferente.

Para completar este estudio, habría que realizar un análisis de expresión diferencial de esos metabolitos y ver si hay alguno que se encuentre diferencialmente expresado en estos pacientes. Esto podría ser de gran utilidad a la hora de encontrar biomarcadores de la enfermedad o incluso servir como marcador para estratificar a los pacientes en más o menos grave o que puedan ser susceptibles a algún tratamiento.

A partir de este estudio podemos concluir:

1. Hemos creado un objeto SummarizedExperiment a partir del dataframe human_cachexia del cual hemos extraido luego la información para realizar un análisis exploratorio.
2. Del análisis exploratorio hemos asumido que había ciertas muestras que no se comportaban como el resto del dataset, por lo que hemos generado un nuevo subset de datos para su análisis.
3. El análisis de este nuevo subset ha arrojado cierta tendencia de un conjunto de pacientes a agrupar entre ellos, lo que podría ser indicativo de algún motivo biológico.

Disponibilidad del código

Los archivos derivados de este trabajo para completar la PEC1 se encuentran disponibles en el siguiente repositorio de Github:

https://github.com/aiturratesuoc/Iturrate-Soleto-Asier-PEC1.git