Informe_Torrejón_Aquino_Alberto_PEC1

Informe: Exploración de datos metabolómicos en caquexia humana

https://github.com/Albertotorrejon/Torrej-n-Aquino-Alberto-PEC1

Abstract

En esta PEC hemos podido explorar el data set human_cachexia, en el cual nos encontramos datos metabolómicos de 77 pacientes humanos. Este, pertenece a un estudio en el cual alguno de estos pacientes tenían la enfermedad y algunos pacientes pertenecían al grupo control. Estos datos, se han estructurado utilizando un objeto SummarizedExperiment. Este es un paquete de bioconductor que se utiliza principalmente para las ciencias ómicas. Anteriormente, se utilizaba el paquete biobase, pero este paquete no cubre las necesidades de estudio, ya que con SummarizedExperiment, además de tener otros métodos para construir la estructura de los objetos, nos va a ser muy útil para soportar múltiples assays. Esto, nos va a ser muy interesante para la exploración de datos en human_cachexia ya que vamos a integrar una matriz de 63 metabolitos y dos variables clínicas `Patient.ID` y `Muscle.loss`. A través de análisis exploratorios como estadísticas descriptivas, PCA y heatmaps, se han detectado patrones diferenciados entre los grupos. Esto puede permitir la identificación de metabolitos con potencialidad para ser biomarcadores.

Objetivos

- Explorar un conjunto de datos metabolómicos de pacientes con y sin caquexia.
- Identificar patrones generales en los niveles de metabolitos.
- Evaluar si existen diferencias notables en los perfiles metabolómicos entre grupos.
- Aplicar técnicas estadísticas multivariantes para visualizar estructura en los datos.

Métodos

Datos utilizados

Se han utilizado datos metabolímicos de 77 pacientes, cada uno de estos pacientes tenían 63 metabolitos por muestra. Los datos se clasificaron según la presencia o ausencia de caquexia usando la variable `Muscle.loss` ("cachexic" o "control").

Observando la matriz de datos proporcionada, podemos ver que cada fila representa una muestra de un paciente y las columnas poseen un identificador de paciente. Es lo que se denomina como ("Patient ID"). Tanto esta variable como "Muscle loss" son variables categóricas que van a funcionar como nuestros metadatos. Por otro lado, tenemos mediciones cuantitatívas numéricas que van a ser los datos de nuestro dataframe

Estructuración del objeto

Los datos se estructuraron en un objeto de clase `SummarizedExperiment` con:

- **Matriz de expresión**: filas = metabolitos, columnas = pacientes.
- **Metadatos**: variables clínicas `Patient.ID` y `Muscle.loss`.

```
{library(SummarizedExperiment)}
```

```
## Warning: package 'MatrixGenerics' was built under R version 4.4.2
```

```
## Warning: package 'matrixStats' was built under R version 4.4.3
```

```
## Warning: package 'IRanges' was built under R version 4.4.2
```

```
## Warning: package 'GenomeInfoDb' was built under R version 4.4.2
```

```
dataframe <- read.csv("human_cachexia.csv")

metadatos <- dataframe[, 1:2]
metabolitos <- dataframe[, -(1:2)]

matriz <- as.matrix(metabolitos)
row.names(matriz) <- metadatos$Patient.ID

se <- SummarizedExperiment(
  assays = list(metabolomics = t(matriz)),
   colData = DataFrame(metadatos)
)

save(se, file = "human_cachexia_SE.Rda")

save(se, file = "human_cachexia_SE.csv")
```

Resultados

Una vez estructurado el objeto con SummarizedExperiment, ya tenemos un tratamiento de datos que podemos utilizar posteriormente para el análisis de datos. En este caso, hemos utilizado 3 métodos: Estadística descriptiva (Summary, boxplot) ,PCA y heatmap.

Estadísticas descriptivas

```
summary(as.vector(assay(se)))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     0.79    17.46    51.42   347.37   160.77 33860.35
```

```r
sum(is.na(assay(se)))
```

```
## [1] 0
```

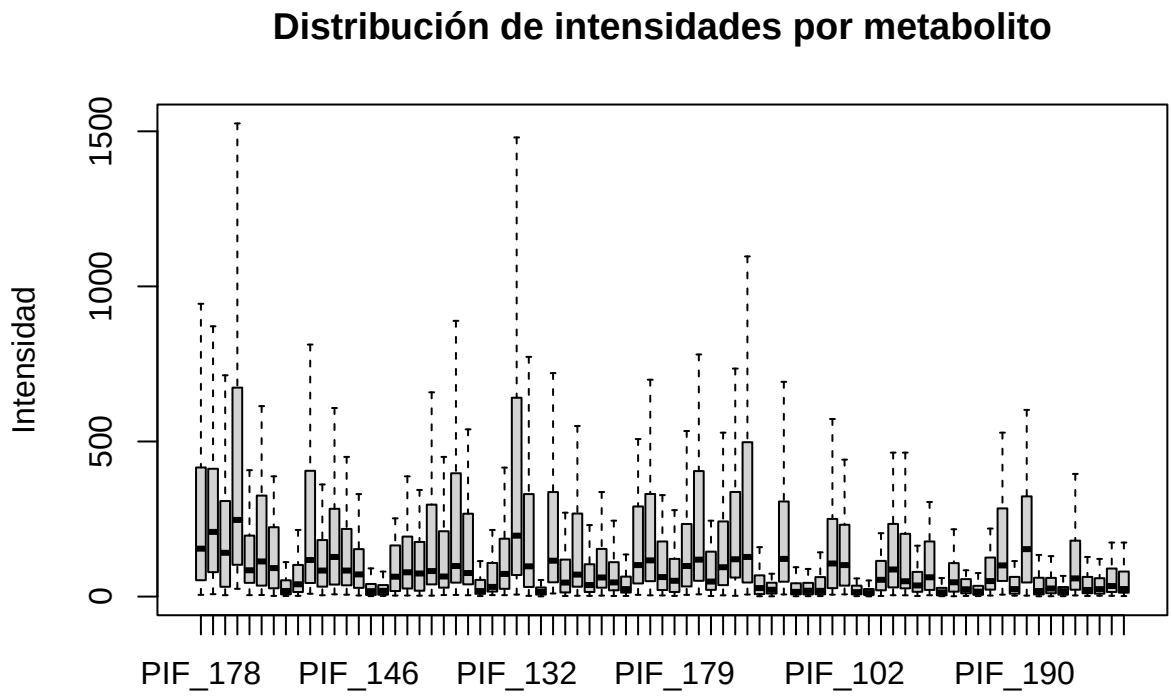
```r
colnames(colData(se))
```

```
## [1] "Patient.ID"  "Muscle.loss"
```

En este caso, con summary, tenemos una visión global del estadístico general sobre el conjunto de intensidades metabolómicas.

Como podemos observar, la media (347,27) es mucho mayor que la mediana (51,42). Esto indica que tenemos algunos valores que van a ser muy elevados que van a descompensar la media. Este dato nos puede resultar bastante interesante a la hora de encontrar algún biomarcador. Este hecho, también lo podemos observar con la gran diferencia que existe entre los cuartiles y la mediana con el valor máximo. Para poder observar esto, vamos a realizar el boxplot.

```r
boxplot(assay(se),
        outline = FALSE,
        main = "Distribución de intensidades por metabolito",
        ylab = "Intensidad")
```



En este boxplot verificamos que hay un sesgo en la distribución, es decir, en la gran mayoría vemos la distrubución en la zonas de intensifades más bajas, mientras que algunos metabolitos superan ámpliamente el rango bajo. También se puede observar con los "bigotes", los cuales se alargan bastante hacia arriba sugiriendo presencia de valores atípicos

PCA

En un gráfico de PCA (Análisis de Componentes Principales), cada punto corresponde a una muestra individual del conjunto de datos. en este caso, partimos de una matriz de metabolitos por muestra y PCA condensa esta información multidimensional en dos ejes. Por lo tanto, cada punto en nuestra grafica representa un metabolito. La posición de un punto nos va a reflejar características combinadas de todos los pacientes.

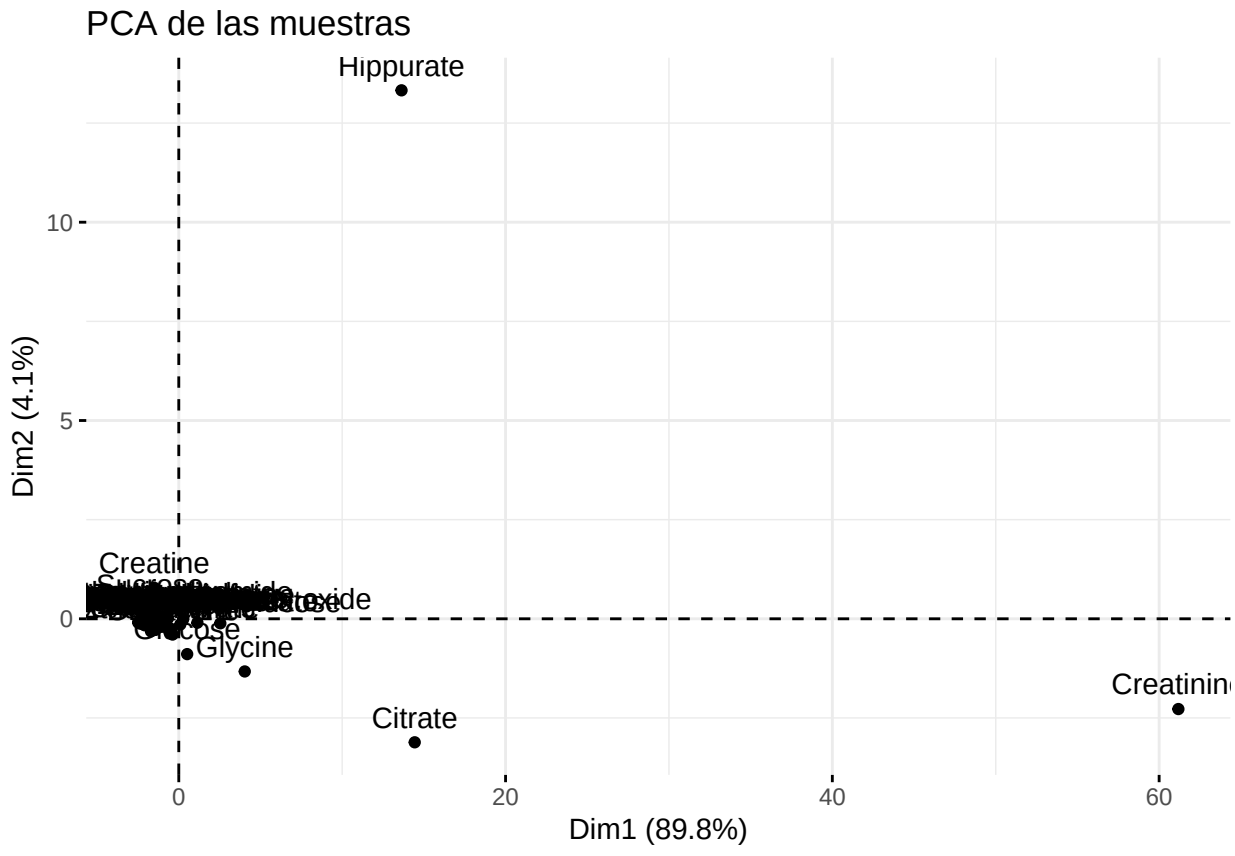
```
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.4.2
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.4.2

## Cargando paquete requerido: ggplot2

## Warning: package 'ggplot2' was built under R version 4.4.2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
pca <- PCA(t(matriz), graph = FALSE)
fviz_pca_ind(pca,
             title = "PCA de las muestras")
```



Podemos observar cómo algunos valores como la creatinina, el citrato y Hippurate nos proporcionan unos valores desmesurados en comparación con el resto de metabolitos. Esto, tiene bastante sentido si hacemos una comparación de este gráfico con el Boxplot. Podíamos observar que los valores que despuntaban eran concretamente 3 metabolítos. El resto, parecía cumplir un patrón en el cual seguían los valores de una distribución similar y no desbalanceada.

Heatmap

Heatmap nos identifican grupos fuertemente correlacionados. En este caso, como hablamos de metabolitos, por lo que la correlación positiva nos indica que los metabolítos tienen una relación fuerte, esto puede ser por ejemplo que tengan una ruta metabólica común. Estas relaciones positivas aparecen representadas en el gráfico en color rojo. En contraposición, tenemos la corelación negativa, que puede reflejar diferencias metabólicas relevantes. Este caso es interesante debido a que estos resultados respaldan los posibles biomarcadores de este estudio.

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
```

```
## Warning: package 'BiocManager' was built under R version 4.4.2
```

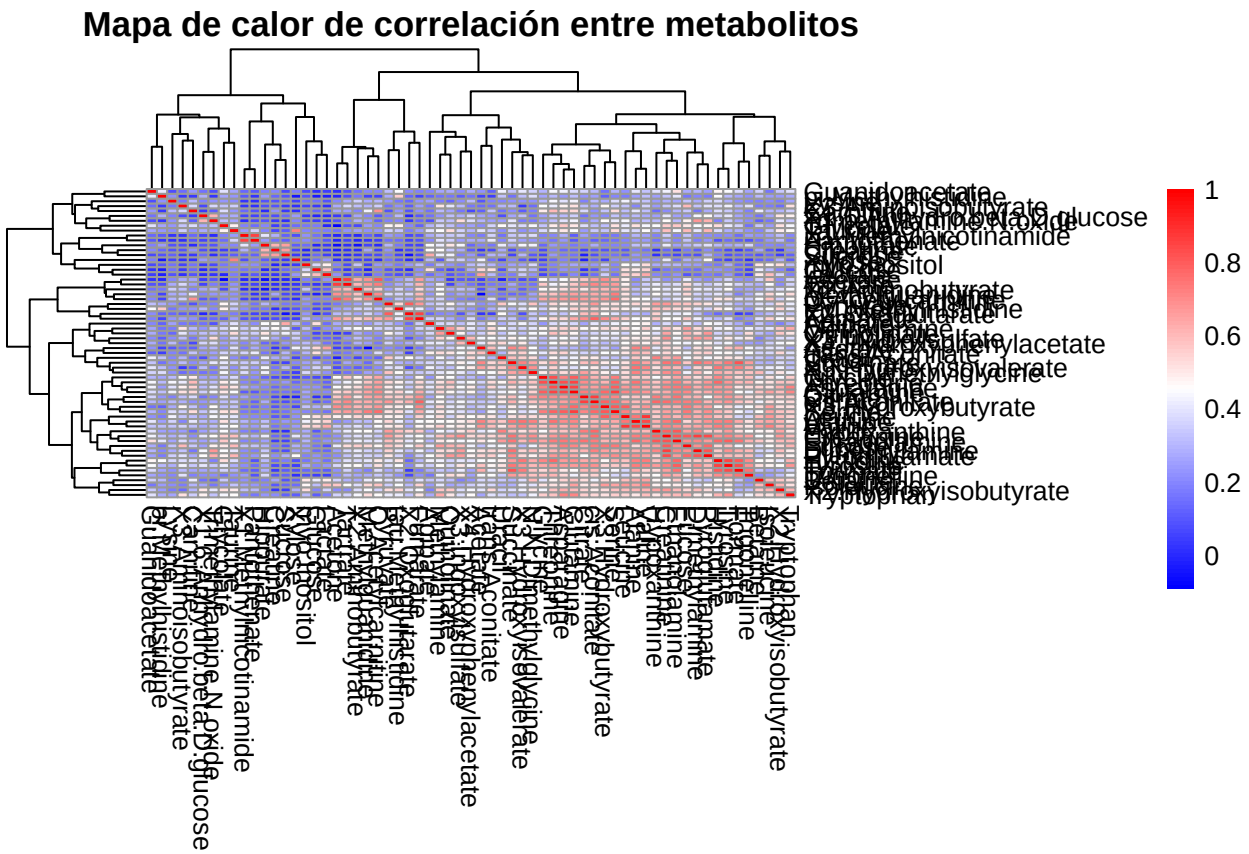
```
BiocManager::install("EnrichedHeatmap")
```

```
## Bioconductor version 3.20 (BiocManager 1.30.25), R 4.4.1 (2024-06-14 ucrt)

## Warning: package(s) not installed when version(s) same as or greater than current; use
##   'force = TRUE' to re-install: 'EnrichedHeatmap'

## Installation paths not writeable, unable to update packages
##   path: C:/Program Files/R/R-4.4.1/library
##   packages:
##     boot, class, cluster, foreign, KernSmooth, lattice, MASS, Matrix, mgcv,
##     nlme, nnet, rpart, spatial, survival

## Old packages: 'cli', 'rlang', 'xfun'
```

```
#Cargamos la librería para poder realizar el heatmap

library(pheatmap)
```

```
## Warning: package 'pheatmap' was built under R version 4.4.3
```

```
# Seleccionamos solo los metabolitos (a partir de la columna 3, que es donde tenemos los datos, evitand
datos_metabolitos <- dataframe[, 3:ncol(dataframe)]

# Calculamos posteriormente la matriz de correlación
matriz_cor <- cor(datos_metabolitos)

# Por último, calculamos el heatmap
pheatmap(matriz_cor,
         color = colorRampPalette(c("blue", "white", "red"))(100),
         main = "Mapa de calor de correlación entre metabolitos")
```



El heatmap refleja agrupamientos consistentes con el patrón observado en el PCA.

Discusión

A partir del análisis realizado, se evidencian diferencias notables y consistentes en los perfiles metabolómicos de los distintos pacientes. Tanto el análisis de componentes principales (PCA) como el mapa de calor reflejan una clara separación entre los grupos estudiados, lo que indica que condiciones clínicas como la caquexia podrían estar asociadas con modificaciones concretas a nivel metabólico. Específicamente, se detectaron alteraciones significativas en metabolitos como la creatinina, el citrato y el hippurato, los cuales parecen jugar un papel importante en la distinción entre los pacientes con caquexia y los controles. Estas variaciones pueden estar vinculadas a procesos fisiológicos como la pérdida de masa muscular, el aumento del estrés oxidativo o disfunciones en el metabolismo energético, fenómenos característicos en estados clínicos de deterioro como este.

Por otro lado, más allá de esta exploración inicial, sería especialmente útil la aplicación de herramientas predictivas que permitan comprobar si estos metabolitos se comportan de forma similar en otros grupos de pacientes o en fases distintas del proceso patológico. La creación de modelos estadísticos o de aprendizaje automático podría no solo confirmar los resultados obtenidos, sino también facilitar la detección de biomarcadores sólidos y con relevancia clínica, capaces de anticipar la aparición de caquexia o de seguir su evolución en el tiempo. Además, profundizar en las condiciones que favorecen la alteración o acumulación de estos compuestos podría ayudar a esclarecer los mecanismos biológicos implicados y a plantear nuevas estrategias para su tratamiento.

Conclusiones

El análisis llevado a cabo ha permitido detectar diferencias claras en los perfiles metabólicos entre los pacientes con caquexia y aquellos pertenecientes al grupo control. Estos hallazgos han sido posibles gracias a la correcta estructuración de los datos mediante el uso del objeto SummarizedExperiment, que ha proporcionado una forma eficiente y ordenada de integrar tanto las matrices de expresión metabolómica como los metadatos clínicos asociados. Esta organización ha facilitado la aplicación de métodos estadísticos multivariantes como el Análisis de Componentes Principales (PCA) y los mapas de calor (Heatmaps), herramientas que han sido clave para visualizar y confirmar la existencia de patrones diferenciados entre ambos grupos.

El uso de este tipo de estructuras no solo mejora la gestión y manipulación de datos complejos, sino que también potencia la reproducibilidad y claridad del análisis. Los resultados obtenidos refuerzan la utilidad de herramientas desarrolladas en entornos como Bioconductor, donde paquetes como SummarizedExperiment permiten llevar a cabo análisis de datos ómicos con mayor robustez. En el contexto específico de enfermedades como la caquexia, estas metodologías no solo optimizan el proceso de análisis, sino que también abren la puerta a futuras investigaciones centradas en la identificación de biomarcadores o en el desarrollo de modelos predictivos con valor clínico.