Análisis de datos ómicos en Cachexia Humana

PEC 1 - Análisis de Datos Ómicos

Giuliana Ricco

Índice

Abstract

El análisis de datos ómicos es fundamental para entender los procesos biológicos que dan lugar a distintas condiciones. En este estudio, se analizan un conjunto de datos de metabolitos en pacientes con cachexia utilizando herramientas bioinformáticas y estadísticas en R. Se organiza la información en un objeto SummarizedExperiment y se realizaron tres análisis principales: un boxplot para examinar la distribución de los metabolitos, una matriz de correlación para evaluar relaciones entre ellos y un análisis de componentes principales (PCA) para explorar la variabilidad en los datos. Los resultados muestran que algunos metabolitos, como la creatina, presentan valores significativamente elevados en comparación con otros. La matriz de correlación evidenció relaciones entre ciertos metabolitos, lo que sugiere posibles interacciones metabólicas. El PCA por su parte no da una estructura clara para diferenciar a los pacientes según la información dada.

Objetivos

El principal objetivo de este informe es realizar un análisis estadístico del conjunto de datos proporcionados acerca de la cachexia humana. Este análisis se realizará utilizando diversas técnicas estadísticas para así poder extraer la información más relevante para poder discutirla. Los objetivos específicos son:

- Crear un objeto SummarizedExperimente que permita organizar y gestionar los datos y metadatos, facilitando su análisis.
- Analizar la distribución de los metabolitos mediante un boxplot, para visualizar estas distintas distribuciones y poder identificar ciertas características generales.
- Estudiar la distribución de los metabolitos mediante una matriz de correlación, y así identificar asociaciones entre los mismos.
- Realizar un análisis de componentes principales (PCA) con el cual se pretende visualizar las principales fuentes de variabilidad del conjunto.

Métodos

El conjunto de datos utilizado corresponde a un archivo CVS que contiene datos sobre la cachexia humana, denominado human_cachexia.cvs, obtenido a partir de los materiales a disposición para la realización de esta PEC. En este archivo cada columna corresponde a un metabolito distinto mientras que cada fila se refiere a un paciente del cual han obtenido las mediciones. El conjunto de datos también incluye la identificación de cada paciente así como la pérdida muscular de cada uno.

Los datos fueron tratados con el programa R, en un primer lugar siendo tratados y transformados en un data frame. A continuación se organizó la información en un objeto SummarizedExperient para facilitar su análisis.

Posteriormente se llevó a cabo una visualización de los datos con el objetivo de obtener una primera impresión acerca de sus distribuciones. Para esto se realizó un diagrama de cajas tipo bloxplot (anexo 1) con el cual es posible visualizar las distribuciones de los distintos metabolitos en el conjunto de datos y las posibles tendencias que estos tengan. Este tipo de gráficos facilita la visualización de posibles valores atípicos así como proporcionar una visión más general de la variabilidad de los mismos.

Para poder identificar las distintas relaciones entre los metabolitos, se calculó una matriz de correlación. Esta matriz de correlación permitió evaluar qué metabolitos presentaban asociaciones entre sí. Esto resulta muy util a la hora de identificar patrones o interacciones entre los distintos metabolitos. Los resultados fueron representados mediante un mapa de calor (anexo 2).

Por último se realizó un Análisis de Componentes Principales (PCA). Este análisis tiene como fin minimizar la dimensionalidad de los datos, y así poder estudiar las fuentes de variabilidad dentro del conjunto de datos. Este tipo de análisis permiten identificar patrones y tendencias a la hora de la manifestación de la condición en los pacientes. Los resultados de este análisis están representados en un gráfico de dispersión (anexo 3) que muestra los dos primeros componentes principales.

Cada uno de estos análisis se llevó a cabo con herramientas y paquetes de R, incluyendo ggplot2, pheatmap, tidyverse y SummarizedExperiment.

Resultados

SummarizedExperiment

Lo primero ha realizar del análisis es la transformación de los datos del dataset en un objeto SummarizedExperiement, lo que facilita la organización y trabajo de la información de una manera más eficiente. Este objeto permite organizar los datos de tal forma que se obtiene una matriz con los valores de los metabolitos asociados a cada paciente, una tabla con la información de las muestras y otra con los nombres de los metabolitos.

Una vez hecho esto reviso que se ha creado correctamente así como genero un resumen de los contenidos del mismo. Este proceso se puede ver tanto en el archivo generado en html por R como en el propio archivo del programa facilitado.

```r
ruta <- "F:/UOC/2 cuatri/Análisis datos ómicos/PEC1/human_cachexia.csv" #cargar datos
cachexia <- read_csv(ruta) #leer datos
head(cachexia)
```

```r
cachexia <- as.data.frame(cachexia) #convertir en data frame
assay_data <- as.matrix(cachexia[, -c(1, 2)])  #quitar patient ID y mucleloss

col_data <- DataFrame(
  SampleID = colnames(assay_data)  # definir columnas
)

row_data <- DataFrame(
  Metabolite = cachexia[[1]]  # definir filas
)

se <- SummarizedExperiment( #crear objeto SummarizedExperiment
  assays = list(counts = assay_data),
  colData = col_data,
  rowData = row_data
)

se
```

```r
assay(se)  # ver la matriz de datos
rowData(se)  # Ver metadatos de metabolitos
colData(se)  # Ver metadatos de muestras
```

```r
#guardar el objeto SummarizedExperiment
objeto_se <-"F:/UOC/2 cuatri/Análisis datos ómicos/PEC1/human_cachexia_SE.Rda"
save(se, file = objeto_se)
```

```r
#cargamos el objeto y visualizamos para comprobra que se ha creado bien
load("F:/UOC/2 cuatri/Análisis datos ómicos/PEC1/human_cachexia_SE.Rda")
se
```

```r
#ver resumen de SE para hacernos una idea de los datos que manejamos
summary(assay(se))
```

```r
#generar data frame de se y verificar columnas
df <- as.data.frame(assay(se))
colnames(df)
```

Boxplot

La realización de una boxplot (diagrama de cajas) es útil a la hora de explorar la distribución del conjunto de datos del archivo. Se pueden observar como ciertos metabolitos presentan una variabilidad considerable entre los pacientes asu como otros muestran distribuciones mucho más homogéneas entre los distintos pacientes. Este tipo de análisis es útil para poder identificar valores atípicos , lo que puede indicar diferencias relevantes biológicas.

Este tipo de análisis es además muy útil para poder detectar posibles inclinaciones en la distribución de los datos a los que tal vez haya que dar más importancia.

```r
```{r Boxplot}
#convertir a formato largo para usar con base r
df_long <- df %>%
  pivot_longer(cols = everything(), names_to = "Metabolite", values_to = "Value")

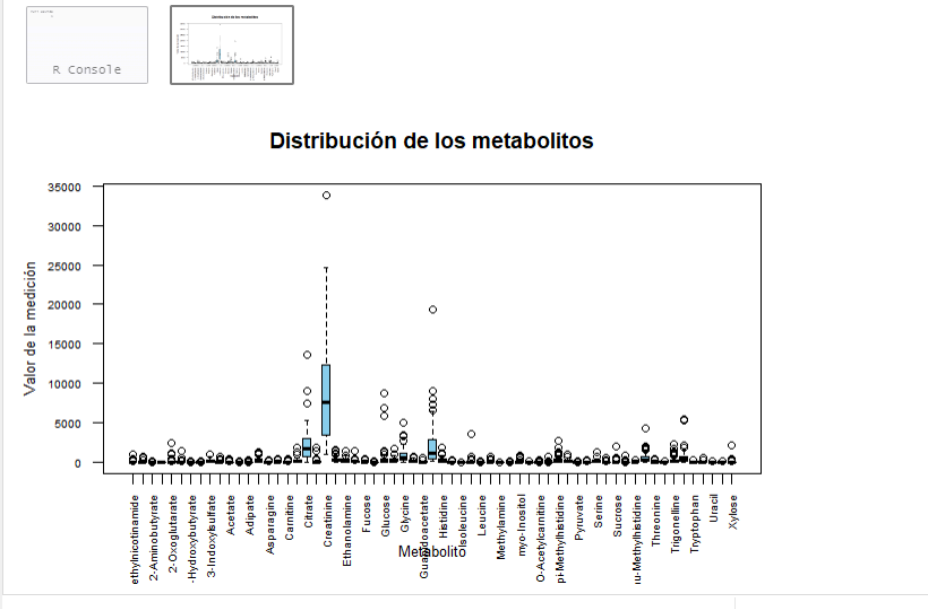
#SVG
svg("boxplot_distribucion_metabolitos_baseR.svg", width = 14, height = 10)

#boxplot
boxplot(Value ~ Metabolite, data = df_long,
        main = "Distribución de los metabolitos",
        xlab = "Metabolito", ylab = "Valor de la medición",
        col = "skyblue",
        las = 2,  #angulo de etiquetas eje x
        cex.axis = 0.8,  #fuente de ejes
        cex.main = 1.2,  #tamaño titulo
        cex.lab = 1.0)   #tamaño etiquetas eje

dev.off()  #cierra SVG

#ver el gráfico
boxplot(Value ~ Metabolite, data = df_long,
        main = "Distribución de los metabolitos",
        xlab = "Metabolito", ylab = "Valor de la medición",
        col = "skyblue",
        las = 2,
        cex.axis = 0.6,
        cex.main = 1.2,
        cex.lab = 0.8)

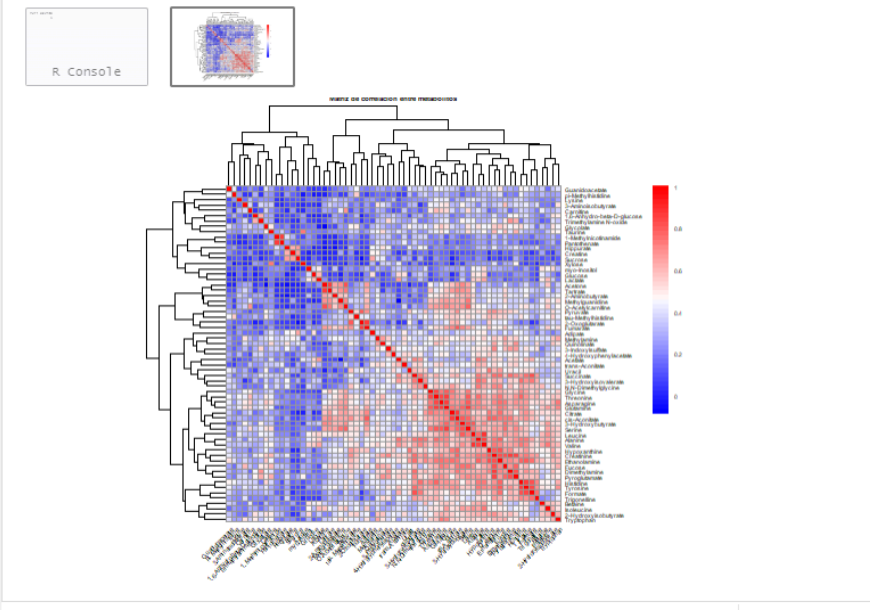
```
```



Matriz de Correlación

La matriz de correlación se genera para relacionar o intentar identificar las relaciones que pueden existir entre los distintos metabolitos. Los resultados se representan mediante un mapa de calor que permite visualizar rápidamente las asociaciones que tienen los metabolitos entre sí, tanto de forma positiva como negativa. El hecho de que ciertos metabolitos esten positivamente relacionados entre si puede indicar que estan involucrados en rutas metabólicas similares y/o procesos fisiológicos relacionados.

```{r Matriz de Correlación}
#matriz de correlacion entre metabolitos
cor_matrix <- cor(df, use = "complete.obs")  #eliminar valores nulos

#SVG
svg("matriz_correlacion_metabolitos.svg")

pheatmap(cor_matrix,
         cluster_rows = TRUE,
         cluster_cols = TRUE,
         color = colorRampPalette(c("blue", "white", "red"))(50),
         main = "Matriz de correlación entre metabolitos",
         fontsize = 4,  #fuente
         cellwidth = 4,   #ancho celda
         cellheight = 4,  #alto celda
         angle_col = 45,  #angulo columna
         angle_row = 0)   #angulo fila

dev.off()  #cierra SVG

#ver el gráfico
pheatmap(cor_matrix,
         cluster_rows = TRUE,
         cluster_cols = TRUE,
         color = colorRampPalette(c("blue", "white", "red"))(50),
         main = "Matriz de correlación entre metabolitos",
         fontsize = 3.4,
         cellwidth = 3.5,
         cellheight = 3.5,
         angle_col = 45,
         angle_row = 0)
```

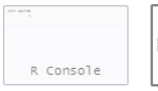


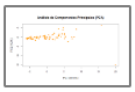
Análisis de Componentes Principales

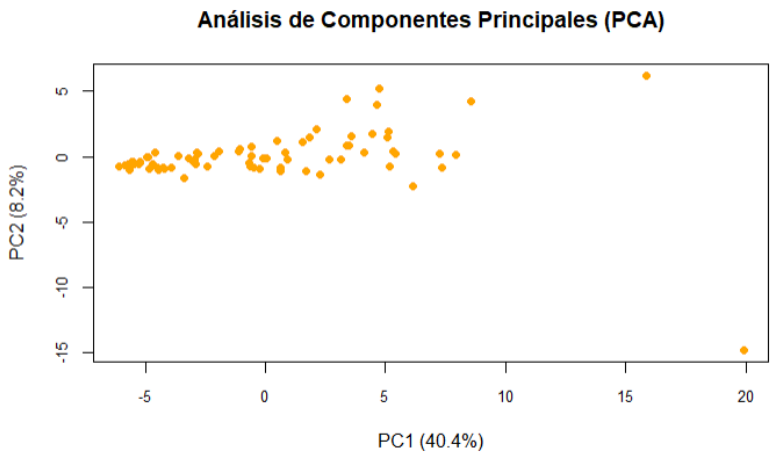
El PCA se realiza para simplificar la dimensionalidad del conjunto de datos y descubrir patrones generales.En este caso se grafican los dos primeros componentes principales.

Al observar el gráfico generado para el PCA se ve que no se forman cluster definidos sino que se ve una semi nube de puntos que tienden a una distribución común. Esto indica que aunque los metabolitos analizados muestran variabilidad, no existe una separación definida que muestre una clasificación clara dentro del conjunto de datos.

```{r PCA}
pca <- prcomp(df, center = TRUE, scale. = TRUE)

#SVG
svg("pca_metabolitos.svg", width = 14, height = 10)

plot(pca$x[, 1:2],
     col = "orange",
     pch = 16, #tipo de marcador (circulos)
     main = "Análisis de Componentes Principales (PCA)",
     xlab = paste("PC1 (", round(100 * summary(pca)$importance[2,1], 1), "%)", sep = ""),
     ylab = paste("PC2 (", round(100 * summary(pca)$importance[2,2], 1), "%)", sep = ""),
     cex.axis = 0.8, #tamaño fuente ejes
     cex.main = 1.2, #tamaño titulo
     cex.lab = 1.0)  #tamaño etiquetas eje

dev.off()

#ver grafico
plot(pca$x[, 1:2],
     col = "orange",
     pch = 16,
     main = "Análisis de Componentes Principales (PCA)",
     xlab = paste("PC1 (", round(100 * summary(pca)$importance[2,1], 1), "%)", sep = ""),
     ylab = paste("PC2 (", round(100 * summary(pca)$importance[2,2], 1), "%)", sep = ""),
     cex.axis = 0.8,
     cex.main = 1.2,
     cex.lab = 1.0)
```







Discusión

Bloxpot

El boxplot permite visualizar la distribución de cada uno de los metabolitos. Se destacan la creatina presentando valores significativamente más altos que el resto de los metabolitos. También podemos ver otros como el citrato entre otros son más elevados que la mayoría.

La presencia de estos valores altos podría estar relacionada con el metabolismo muscular y la energética celular, aunque necesitaría de un análisis adicional para confirmar esta hipótesis.

Matriz de Correlación

La matriz de correlación muestra relaciones significativas entre algunos metabolitos, lo que puede indicar que participan en las mismas vías metabólicas o que su regulación está relacionada.

El hecho de que se puedan observar correlaciones fuertes sugiere que ciertos grupos de metabolitos pueden estar influenciados por factores comunes, como la actividad enzimática o la disponibilidad de sustratos. Sin embargo, también se observan metabolitos con baja correlación entre sí, lo que puede significar que su comportamiento es independiente.

Análisis de Componentes Principales

El gráfico de PCA muestra una nube de puntos dispersa alrededor del eje Y = 0, con valores en el eje X que van desde aproximadamente -5 hasta 7. Esto indica que no existe una estructura que pueda diferenciar claramente a los pacientes según la información dada. Esto puede deberse a muchos factores que no se han considerado en este informe.

Conclusiones

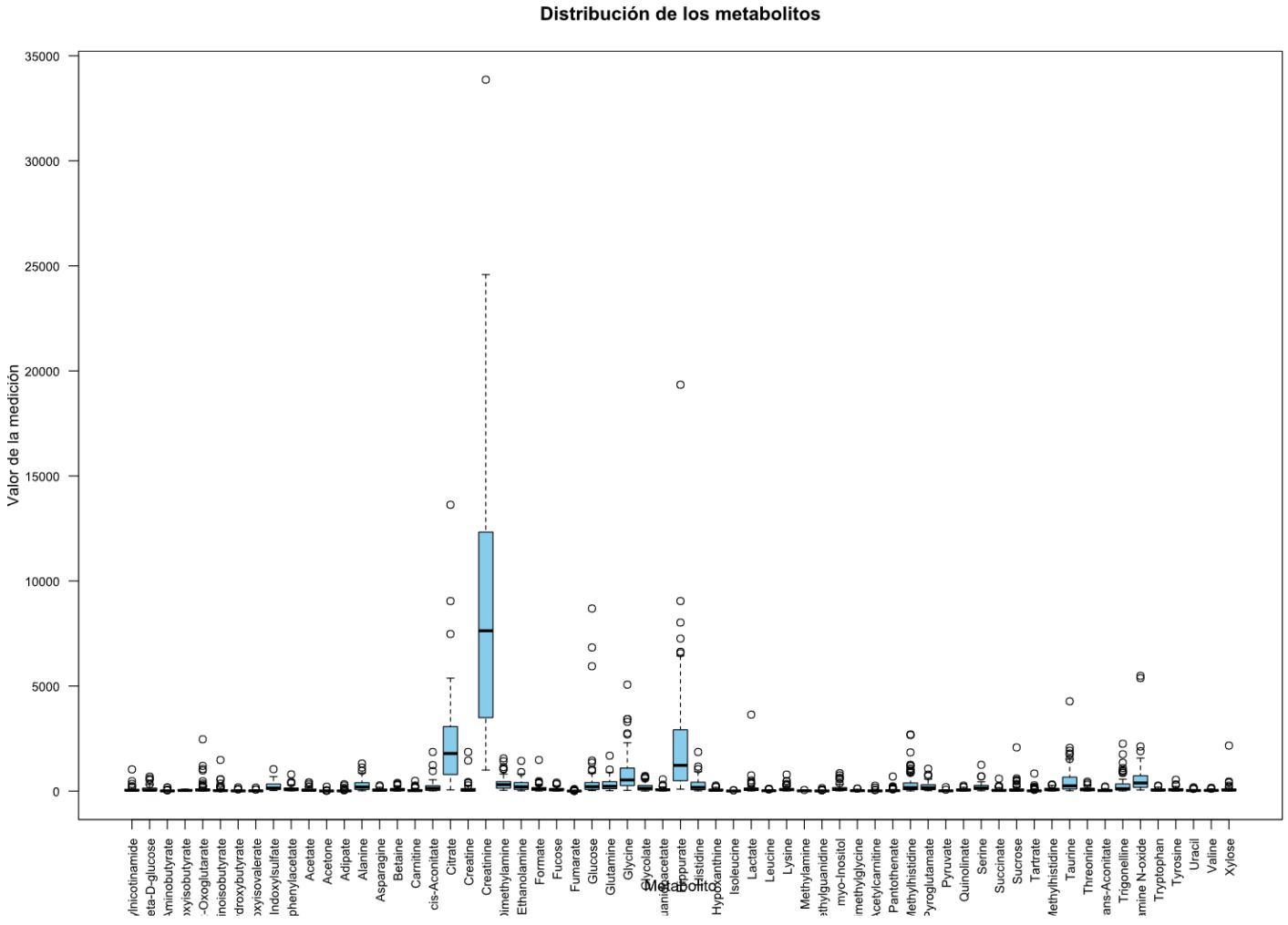
En este informe se exploraron los datos metabólicos en pacientes con cachexia. El boxplot nos mostró que la creatina es uno de los metabolitos con valores más altos, lo que podría indicar ser una sustancia relevante en esta condición. También se ve que hay algunos metabolitos bastante relacionados entre sí, lo que podría sugerir bastantes relaciones relevantes a la hora de investigar esta condición. Sin embargo, por su parte el PCA no muestra una estructura diferenciada que pueda usarse para identificar a los pacientes.

En conclusión este analisis ha permitido extraer algo de información acerca de los datos que pueden ser relevantes en los próximos estudios acerca de esta condiciones. Por la parte de este informe, tan solo se ha pretendido mostrar de manera general la importancia de seguir investigando acerca de los posibles factores que tienen relevancia en esta condición.

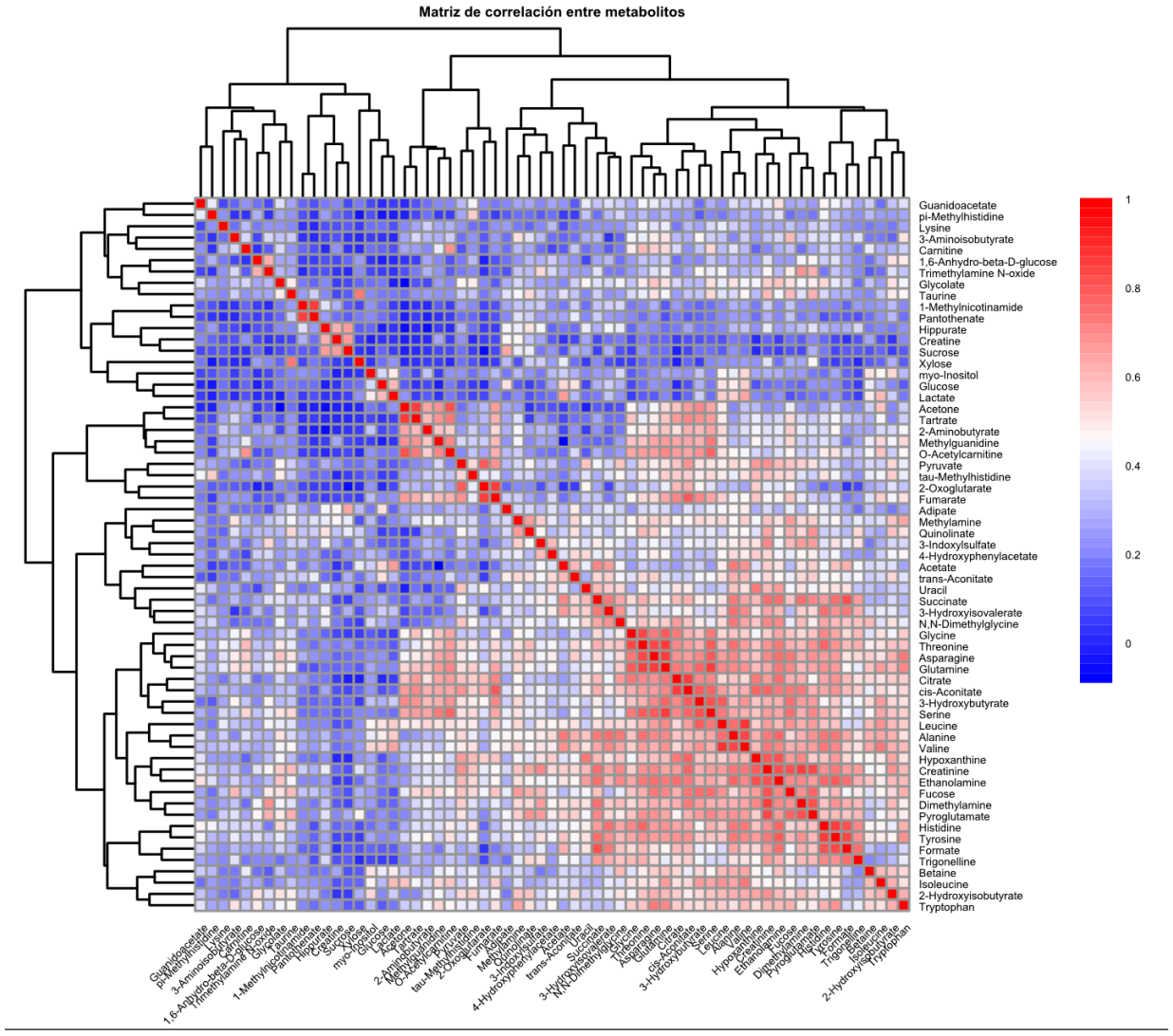
Referencias

- Los materiales proporcionados tanto en la PEC como los que se pueden encontrar a disposición de los alumnos en los contenidos de la UOC para esta asignatura.
- https://github.com/GiulianaUOC/Ricco-Giuliana-PEC1.git

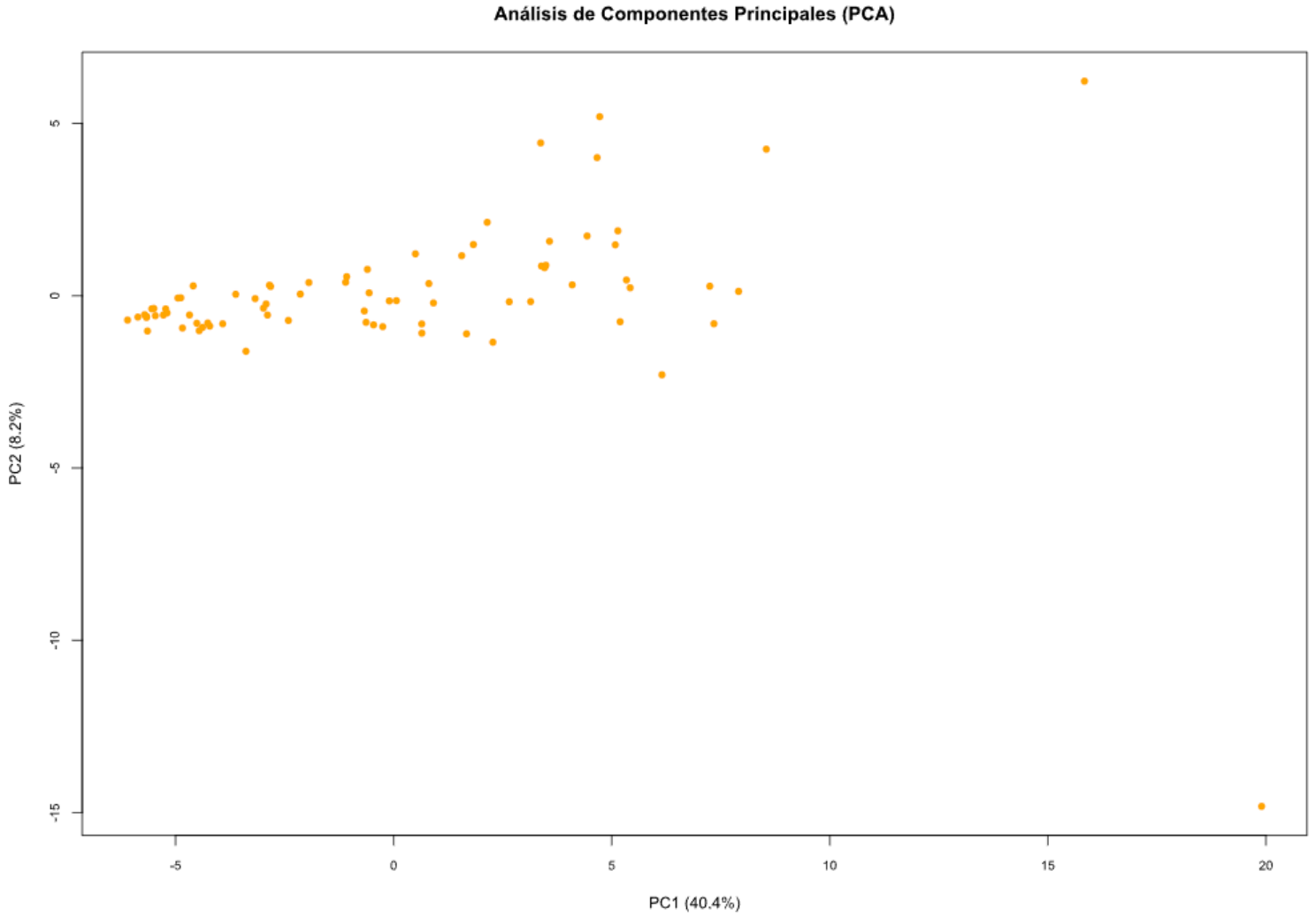
Anexo 1


Distribución de los metabolitos
Valor de la medición
35000
30000
25000
20000
15000
10000
5000
0
Metabolito
Acetate
Acetone
Adipate
Alanine
Asparagine
Betaine
Carnitine
Citrate
Creatine
Creatinine
Ethanolamine
Formate
Fucose
Fumarate
Glucose
Glutamine
Glycine
Hippurate
Histidine
Hypoxanthine
Isoleucine
Lactate
Leucine
Lysine
Methylamine
Pantothenate
Pyruvate
Quinolinate
Serine
Succinate
Sucrose
Tartrate
Taurine
Threonine
Trigonelline
Tryptophan
Tyrosine
Uracil
Valine
Xylose

Anexo 2


Matriz de correlación entre metabolitos
Guanidoacetate
pi-Methylhistidine
Lysine
3-Aminoisobutyrate
Carnitine
1,6-Anhydro-beta-D-glucose
Trimethylamine N-oxide
Glycolate
Taurine
1-Methylnicotinamide
Pantothenate
Hippurate
Creatine
Sucrose
Xylose
myo-Inositol
Glucose
Lactate
Acetone
Tartrate
2-Aminobutyrate
Methylguanidine
O-Acetylcarnitine
Pyruvate
tau-Methylhistidine
2-Oxoglutarate
Fumarate
Adipate
Methylamine
Quinolinate
3-Indoxylsulfate
4-Hydroxyphenylacetate
Acetate
trans-Aconitate
Uracil
Succinate
3-Hydroxyisovalerate
N,N-Dimethylglycine
Glycine
Threonine
Asparagine
Glutamine
Citrate
cis-Aconitate
3-Hydroxybutyrate
Serine
Leucine
Alanine
Valine
Hypoxanthine
Creatinine
Ethanolamine
Fucose
Dimethylamine
Pyroglutamate
Histidine
Tyrosine
Formate
Trigonelline
Betaine
Isoleucine
2-Hydroxyisobutyrate
Tryptophan
1
0.8
0.6
0.4
0.2
0

Anexo 3


Análisis de Componentes Principales (PCA)
PC2 (8.2%)
PC1 (40.4%)
5
0
-5
-10
-15
-5
0
5
10
15
20