PEC 1

Gene García Torres

2024-11-06

Contents

PEC 1

1. Resumen

Este estudio presenta un análisis exploratorio de datos de metabolómica con el objetivo de identificar patrones y correlaciones en los niveles de metabolitos. Utilizando R y el contenedor SummarizedExperiment, se integraron datos de abundancia de metabolitos, información de muestras y detalles específicos de cada metabolito, consolidando así los datos para una exploración robusta.

El análisis incluyó la visualización de distribuciones de abundancia mediante histogramas y boxplots, un mapa de calor de correlaciones para evaluar relaciones entre metabolitos, y un Análisis de Componentes Principales (PCA) que permitió reducir la dimensionalidad y observar agrupamientos de muestras. También se realizó un clustering jerárquico para identificar patrones en las muestras y los metabolitos con perfiles similares.

Los resultados revelaron variaciones notables en las abundancias de ciertos metabolitos, así como correlaciones entre algunos de ellos, sugiriendo relaciones metabólicas significativas. El PCA y el clustering jerárquico destacaron estructuras en los datos, como agrupamientos entre muestras que podrían relacionarse con variables experimentales o biológicas.

2. Objetivos del estudio

Los objetivos dentro del reporte son los siguientes.

- Seleccionar un dataset de metabolómica que podéis obtener de o Este repositorio de github: https://github.com/nutrimetabolomics/metaboData/ o Si lo preferís podéis usar algún dataset del repositorio metabolomicsWorkbench
- Una vez descargados los datos cread un contenedor del tipo SummarizedExperiment que contenga los datos y los metadatos (información acerca del dataset, las filas y las columnas). La clase SummarizedExperiment es una extensión de ExpressionSet y muchas aplicaciones o bases de datos (como metabolomicsWorkbench) lo utilizan en vez de usar expressionSet.
- Llevad a cabo una exploración del dataset que os proporcione una visión general del mismo en la línea de lo que hemos visto en las actividades

3. Materiales y Metodos.

3.1 Origen y Naturaleza de los datos.

Para la exploración de datos del reporte se opto por recurrir al repositorio de github proporcionado por el docente, en donde se eligió un data set con nombre de "2018-MetabotypingPaper", el cual contiene datos utilizados en el articulo "Metabotypes of response to bariatric surgery independent of the magnitude of weight loss".

Estos datos fueron publicados dentro del sitio web del articulo así como en su propio repositorio dentro de github el cual contiene el siguiente contenido:

- DDataInfo_S013.csv: Metadata. Información en cada columna del archivo "DataValues_S013.csv".
- DataValues_S013.csv: Valores clínicos y metabolomicos de 39 pacientes en 5 etapas diferentes.
- AAInformation_S006.csv: Información adicional de metabolitos dentro del archivo "DataValues_S013.csv.

La naturaleza de los datos, es metabolomica y clínica conteniendo las mediciones de abundacion de metabolitos y datos clínicos relevantes en los pacientes.

3.2 Herramientas Informaticas y bioinformaticas.

Para la exploración de los datos se utilizo el programa estadístico de R, y la librería especifica de **"SummarizedExperiment"** para manipular los datos de experimentos.

El modo en que se trabajaron los datos fue **ExpresionSet** o en este caso **SummarizedExperiment** que facilito el manejo y análisis de los datos metabolomicos.

Se emplearon herramientas de análisis estadístico, visualización y pre procesamiento de datos, como heatmap y boxplot, además de procedimientos de imputación de valores faltantes y conversión de variables categóricas a numéricas.

3.3 Procedimiento General de Análisis

Se generaron gráficos de distribución y correlación, tales como histogramas, mapas de calor, y boxplots, para explorar la variabilidad en los datos y eliminar posibles outliers o patrones de ruido.

3.4 Métodos Utilizados

Se utilizó el clustering jerárquico mediante la distancia euclidiana y el método de promedio, representando las agrupaciones en un dendrograma.

También se aplicó un análisis de componentes principales (PCA) para reducir la dimensionalidad y visualizar la variabilidad entre muestras en función de los dos primeros componentes principales.

4. Resultados

4.1 Preparación de datos.

Para poder realizar una exploración de los datos se utilizo la librería **"SummarizedExperiment"** la cual forma parte de Bioconductor.

```
library(SummarizedExperiment)
```

Después se comienza analizando un poco los datos adquiridos del repositorio, en donde observamos primeros los "Data Values", en donde se encuentran los datos clínicos de los pacientes y podemos ver que tiene 39 pacientes y 695 variables.

```
# Leer los datos desde el archivo de valores de metabolitos
data_values <- read.csv("DataValues_S013.csv", row.names = 1)
names(data_values[6:10])
```

```
## [1] "MEDDM_T0"  "MEDCOL_T0" "MEDINF_T0" "MEDHTA_T0" "GLU_T0"
```

Y se muestran un poco el como es la estructura de los datos y en donde encontramos que tenemos valores N.A los cuales se filtraran mas adelante.

```
head(data_values[0:9])
```

```
##   SUBJECTS SURGERY AGE GENDER Group MEDDM_T0 MEDCOL_T0 MEDINF_T0 MEDHTA_T0
## 1        1 by pass  27      F     1        0         0         0         1
## 2        2 by pass  19      F     2        0         0         0         0
## 3        3 by pass  42      F     1        0         0         0         0
## 4        4 by pass  37      F     2        0         0         0         0
## 5        5 tubular  42      F     1        0         0         0         0
## 6        6 by pass  24      F     2        0         0         0         0
```

```
table(is.na(data_values))
```

```
##
## FALSE  TRUE
## 23715  3390
```

Después se leerán los datos dentro del archivo **"DataInfo_S013.csv"**, el cual contiene los meta datos asociados a cada muestra en el conjunto de datos y **"AAInformation_S006"** la cual contiene información adicional de los metabolitos.

```
# Leer metadatos de muestras desde el archivo
sample_info <- read.csv("DataInfo_S013.csv", row.names = 1)
metabolite_info <- read.csv("AAInformation_S006.csv", row.names = 1)
```

En este paso, estamos creando un objeto SummarizedExperiment, una estructura de datos diseñada para integrar y manejar datos experimentales en R, específicamente en análisis de datos ómicos (como metabolómica, transcriptómica, etc.).

```
sample_info <- sample_info[match(colnames(data_values), rownames(sample_info)), ]

metabolite_info <- metabolite_info[match(rownames(data_values), rownames(metabolite_info)), ]
```

Se creo el contenedor con la funcion **SummarizedExperiment**:

```
se <- SummarizedExperiment(
  assays = list(counts = as.matrix(data_values)),
  colData = sample_info,
  rowData = metabolite_info
)
```

Este objeto organiza la información en tres componentes principales:

- **assays:** Aquí es donde almacenamos la matriz de datos (data_matrix) con los valores metabolómicos. Estos valores representan las abundancias de cada metabolito en cada muestra.
- **rowData:** Contiene el data frame metabolite_metadata, el cual guarda información sobre cada metabolito (las filas de data_matrix).

  Este componente facilita la adición de anotaciones o características relevantes para cada metabolito.
- **colData:** Almacena el data frame sample_metadata, que contiene los metadatos de cada muestra (las columnas de data_matrix).

  Esto permite incluir variables como la condición experimental, sexo, tratamiento, entre otros factores relevantes para el análisis.

El uso de SummarizedExperiment permite organizar los datos de una forma que facilita la manipulación y análisis, al tiempo que mantiene toda la información en un solo objeto.

Para verificar el contenedor que se creo, se utilizan las siguientes funciones para evaluar un poco la estructura general, así como las dimensiones.

```
# Verificar el objeto creado
#str(se)          # Estructura general del objeto
dim(se)          # Dimensiones del objeto
```

```
## [1]  39 695
```

Al comenzar con el análisis uilizando la funcion **rowData**, se observa como esta tiene la información sobre las caracteristicas de cada fila de nuestro contenedor.

En este caso, contiene la clase, el metabbolito, platadorma y tipo de dato.

```
rowData(se[1:3])      # Información de los metabolitos
```

```
## DataFrame with 3 rows and 6 columns
##         Class Metabolite.abbreviation  Metabolite    Platform   Data.type
##   <character>             <character> <character> <character> <character>
## 1  aminoacids                     Ala     Alanine    LC-MS/MS  Quantified
## 2  aminoacids                     Arg    Arginine    LC-MS/MS  Quantified
## 3  aminoacids                     Asn  Asparagine    LC-MS/MS  Quantified
##           X
##   <logical>
## 1        NA
## 2        NA
## 3        NA
```

Con la función **colData()**, ahora se puede acceder a los datos sobre las muestras del estudio. En este caso, nos muestra que hay un data frame con 695 filas y 3 columnas, las cuales corresponden a las variables **VarName**, **Vartype** y **Description**.

```
colData(se)      # Información de las muestras
```

```
## DataFrame with 695 rows and 3 columns
##                 VarName      varTpe Description
##             <character> <character> <character>
## SUBJECTS       SUBJECTS     integer    dataDesc
## SURGERY         SURGERY   character    dataDesc
## AGE                 AGE     integer    dataDesc
## GENDER           GENDER   character    dataDesc
## Group             Group     integer    dataDesc
```

```
## ...                 ...         ...         ...
## SM.C18.0_T5 SM.C18.0_T5     numeric    dataDesc
## SM.C18.1_T5 SM.C18.1_T5     numeric    dataDesc
## SM.C20.2_T5 SM.C20.2_T5     numeric    dataDesc
## SM.C24.0_T5 SM.C24.0_T5     numeric    dataDesc
## SM.C24.1_T5 SM.C24.1_T5     numeric    dataDesc
```

Por ultimo, dentro de los tres contenedores, encontramos la funcion **assays()** en donde se encuentra la matriz principal con toda la informacion sobre el estudio. En un primer ejemplo, se muestran solamente los primeros cinco valores de esta, para poder ver como es que se encuentran organizados los datos.

```
assays(se)[[1]][1:5, 1:5]
```

```
##   SUBJECTS SURGERY   AGE  GENDER Group
## 1 " 1"     "by pass" "27" "F"    "1"
## 2 " 2"     "by pass" "19" "F"    "2"
## 3 " 3"     "by pass" "42" "F"    "1"
## 4 " 4"     "by pass" "37" "F"    "2"
## 5 " 5"     "tubular" "42" "F"    "1"
```

4.2 Analisis de datos

Para analizar los datos de manera general se filtro el contenido de la matriz mediante la función **"assay()"** en donde, se decidio utilizar solo las columnas que contienen informacion sobre metabolitos.

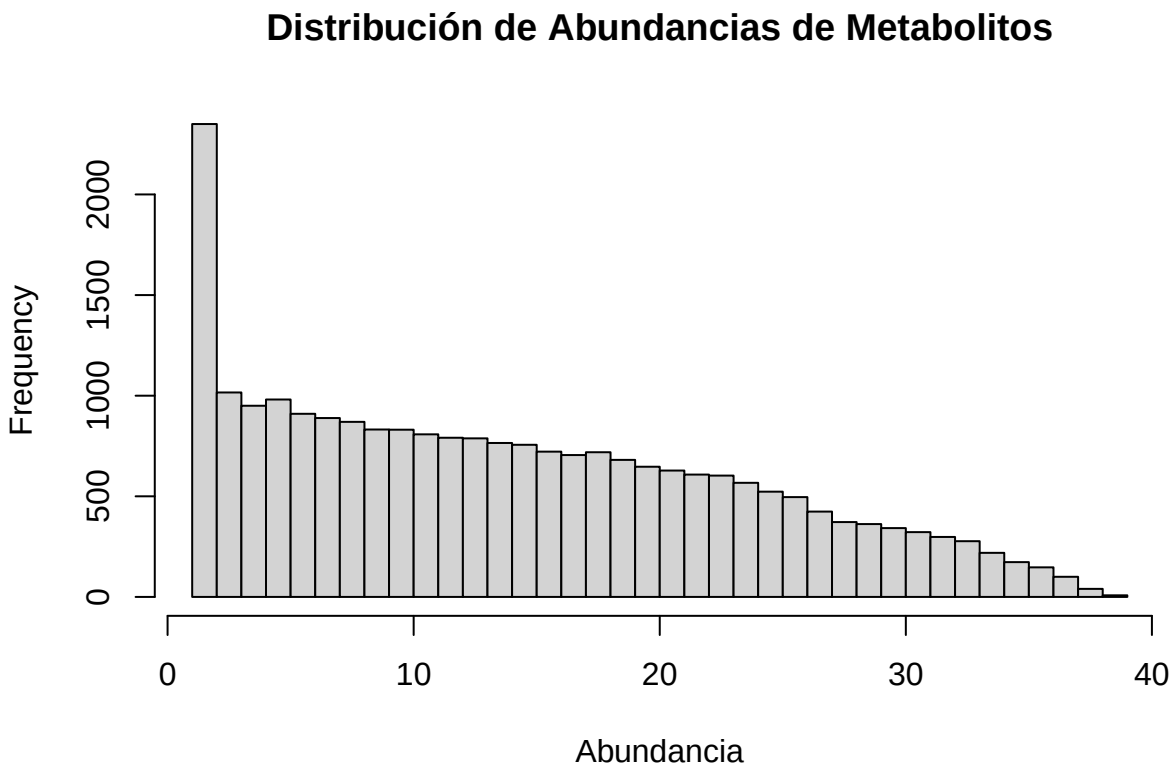
```
# Seleccionar solo las columnas de metabolitos que contienen valores numéricos
metabolites_data <- assay(se)[, 6:ncol(assay(se))]

metabolites_data <- as.data.frame(metabolites_data)  # Convertir a data.frame
metabolites_data <- data.matrix(metabolites_data)    # Convertir todas las columnas a matriz numérica
```

Tambien se realizo un filtrado de datos, para eliminar todo los valores **N.A** que se encontraron.

```
metabolite_clean <- metabolites_data[, colSums(is.na(metabolites_data)) == 0]
```

4.2.1 Histograma Para comenzar con un análisis mas a profundidad, se realizo un histograma con los datos de los metabolitos presentes dentro de la matriz, en donde el eje x nos muestra la abundancia de estos y el eje y la frecuencia con la que ocurren diferentes valores de esta.

```
# Graficar el histograma de abundancia de metabolitos
hist(
  as.vector(metabolites_data),
  main = "Distribución de Abundancias de Metabolitos",
  xlab = "Abundancia",
  breaks = 30
)
```



Conforme el gráfico avanza hacia la derecha, la frecuencia disminuye indicando que los metabolismos con abundancias mas altas son menos comunes por lo que sugiere que en concentraciones mas altas podrian representar compuestos abundantes en el organismo.

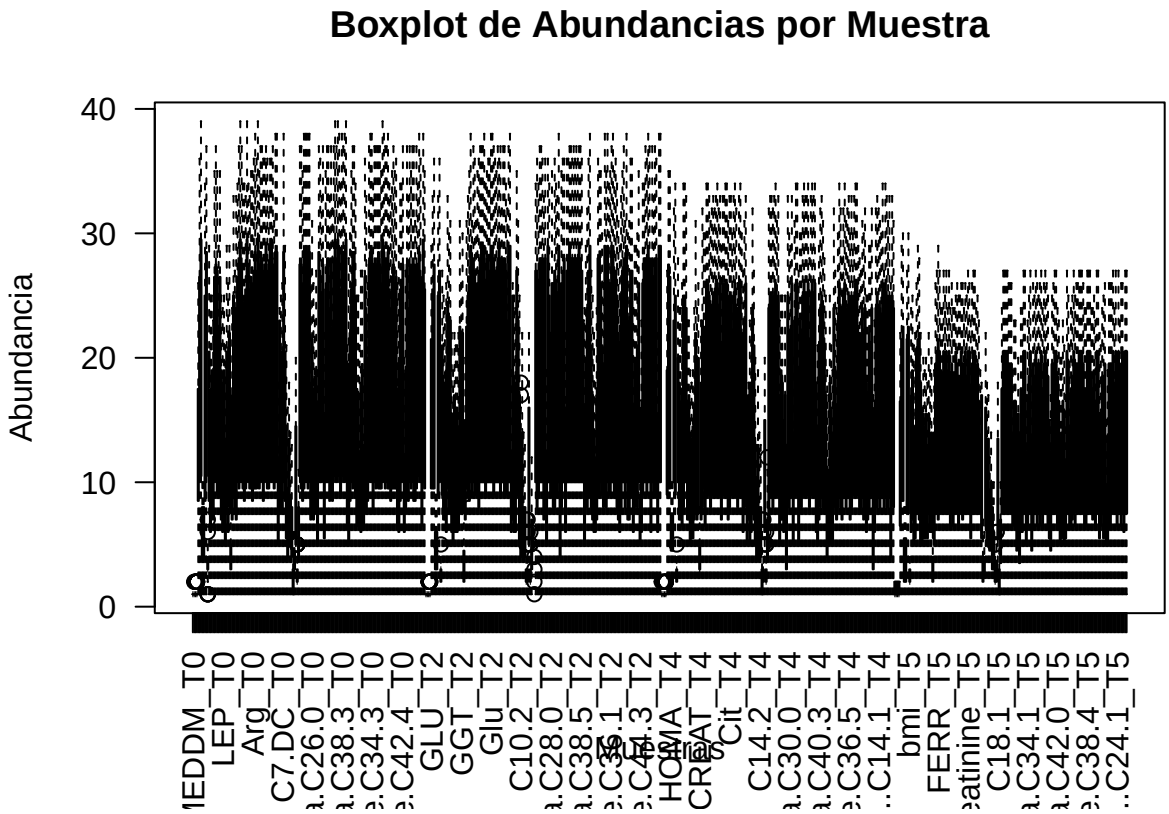
Ademas, que la distribución se encuentre de esta manera puede ser común dentro de datos biológicos al existir una gran diversidad en la expresión o por el contrario que necesite de una transformación de datos para un mejor análisis estadístico.

4.2.2 Boxplot de metabolitos por muestra El siguiente análisis, es el de un diagrama de cajas para visualizar la distribución de metabolitos en cada muestra.

Como se vio en el histograma, la mayoría de los metabolitos tienen concentraciones pequeñas, sin embargo existen outliners lo que indica la presencia de algunos metabolitos con abundancia significativamente mayor que las demás muestras.

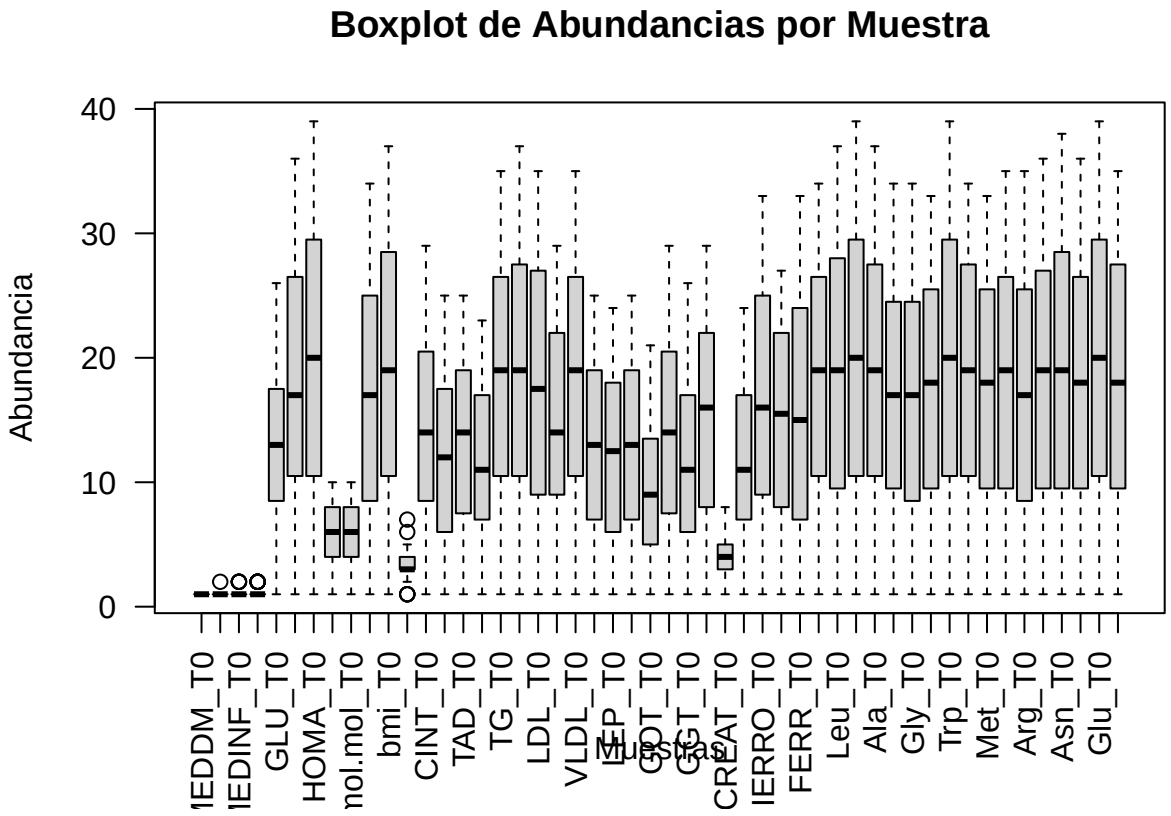
Para fines prácticos del análisis, se realizaron dos exploraciones por boxplot, la primera en donde solo se consideran todos los elementos de nuestro dataset. En donde se observa como es la distribución general de los metabolitos, sin embargo, para fines de análisis es mejor analizar en pequeños grupos para evaluar de mejor manera los cambios dentro de los boxplots.

```
boxplot(
  metabolites_data,
  main = "Boxplot de Abundancias por Muestra",
  xlab = "Muestras",
  ylab = "Abundancia",
  las = 2
)
```



Como se observa en el segundo boxplot, al seccionar los primeros 50 elementos, se pueden apreciar de mejor manera los datos, en donde se observan algunos outliners los cuales pueden ser de interés par algún analisis adicional.

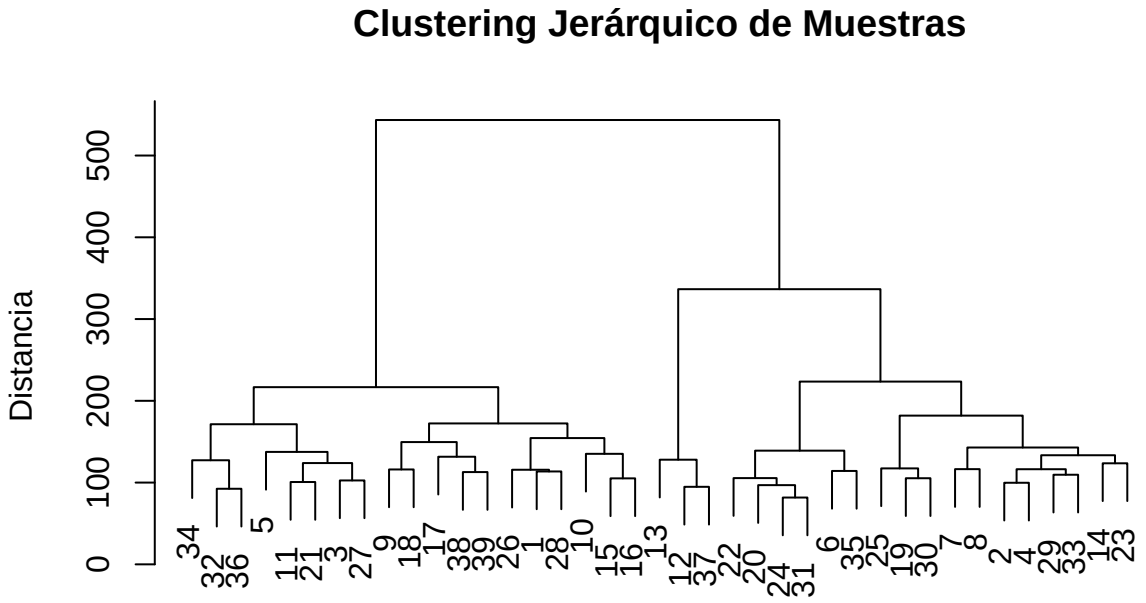
```
# Boxplot de las abundancias de los primeros 50 metabolitos por muestra

boxplot(
  metabolites_data[, 1:50],
  main = "Boxplot de Abundancias por Muestra",
  xlab = "Muestras",
  ylab = "Abundancia",
  las = 2
)
```



4.2.3 Cluster El tercer analisis representa un clustering jerarquico de las muestras basado en la abundancia de los metabolitos. Este cuantifica la similitud o diferencia de la abundancia entre las muestras, el clusterin minimiza la varianza dentro de cada grupo de muestras lo cual brinda una estructura jerarquica qen donde la cohesion de los grupos es mayor y por ultimo un dendograma el cual es una representación visual donde las ramas representan los niveles de agrupamiento.

El tener el dendograma final, nos permite identificar dentro de los grupos observados indicativos de similitudes biologicas o experimentales, mediante la identificacion de patrones y relaciones de acuerdo a la variabilidad y similitud entre muestras.

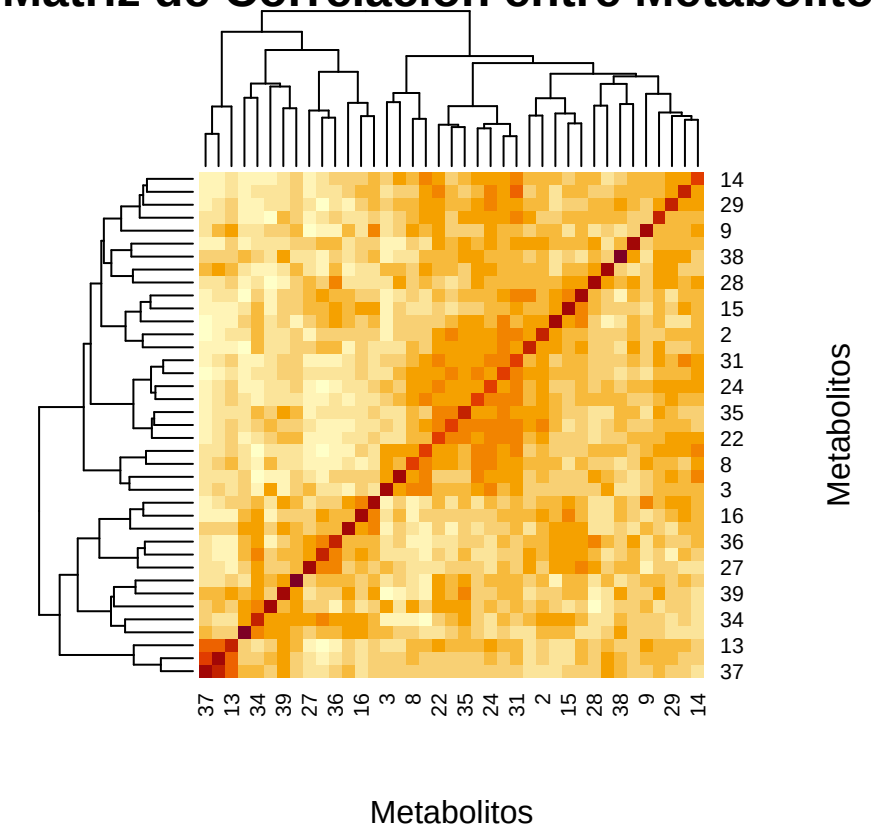
```
# Asumiendo que tus datos están en un data frame o matriz llamado metabolite_clean
# y que las muestras están en las filas y los metabolitos en las columnas.

# 1. Calcular la matriz de distancias
dist_matrix <- dist(metabolite_clean, method = "euclidean")

# 2. Realizar el clustering jerárquico
hc <- hclust(dist_matrix, method = "ward.D2")

# 3. Graficar el dendrograma
plot(hc, main = "Clustering Jerárquico de Muestras", xlab = "", sub = "", ylab = "Distancia")
```



4.2.4 Matriz de correlación Para obtener una visión general de las relaciones entre los metabolitos, se realizó una matriz de correlación, visualizada en un mapa de calor. Este tipo de gráfico permite identificar patrones de correlación entre diferentes metabolitos. Si dos metabolitos muestran una alta correlación, es posible que estén involucrados en procesos biológicos o rutas metabólicas similares.

En el mapa de calor, los colores oscuros representan valores de correlación más cercanos a 1, lo que indica una fuerte correlación positiva entre los metabolitos. Por el contrario, los colores más claros indican valores de correlación cercanos a 0 o negativos, reflejando una correlación baja o negativa. Este esquema de color facilita la identificación visual de los metabolitos que podrían estar estrechamente relacionados en términos de su actividad biológica.

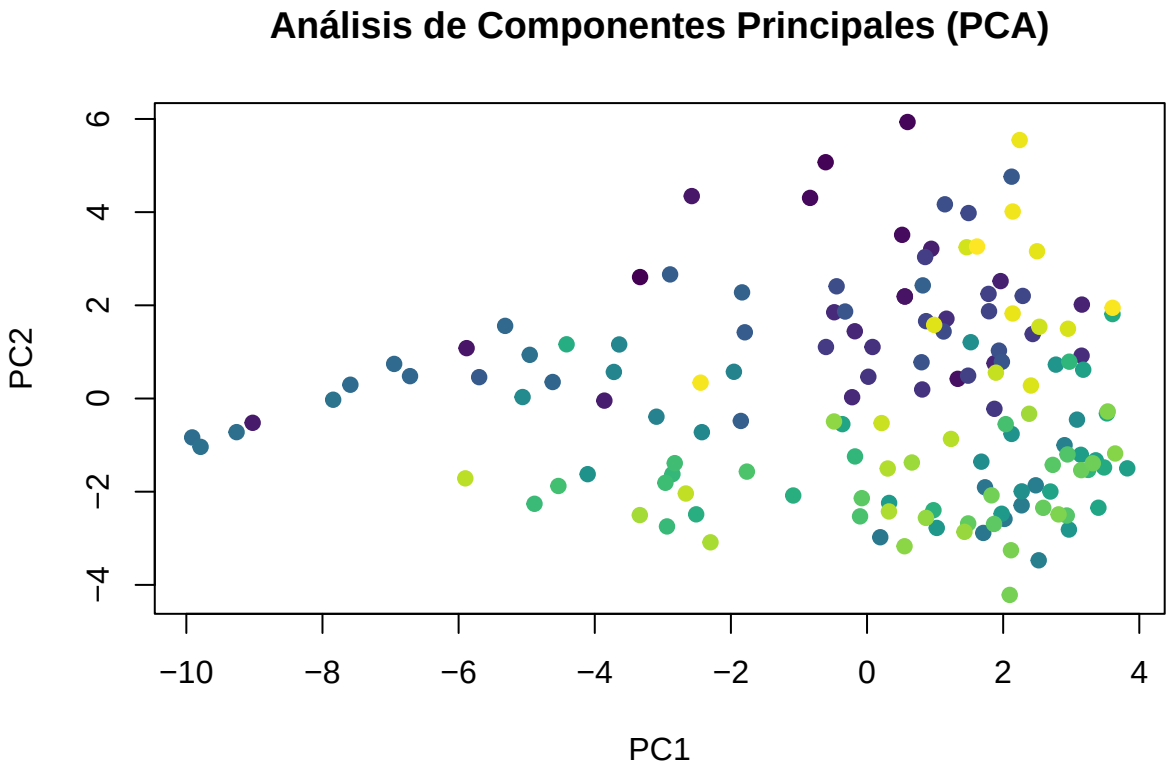
```
cor_matrix <- cor(t(metabolites_data), use = "complete.obs")
heatmap(
  cor_matrix,
  main = "Matriz de Correlación entre Metabolitos",
  xlab = "Metabolitos",
  ylab = "Metabolitos"
)
```



4.2.5 PCA Este PCA nos ayuda a transformar los datos originales de alta dimensional en un grupos que capturan la estructura principal de los datos mostrándonos las relaciones entre las diferentes muestras de forma simplificada, permitiéndonos también identificar patrones o agrupamientos.

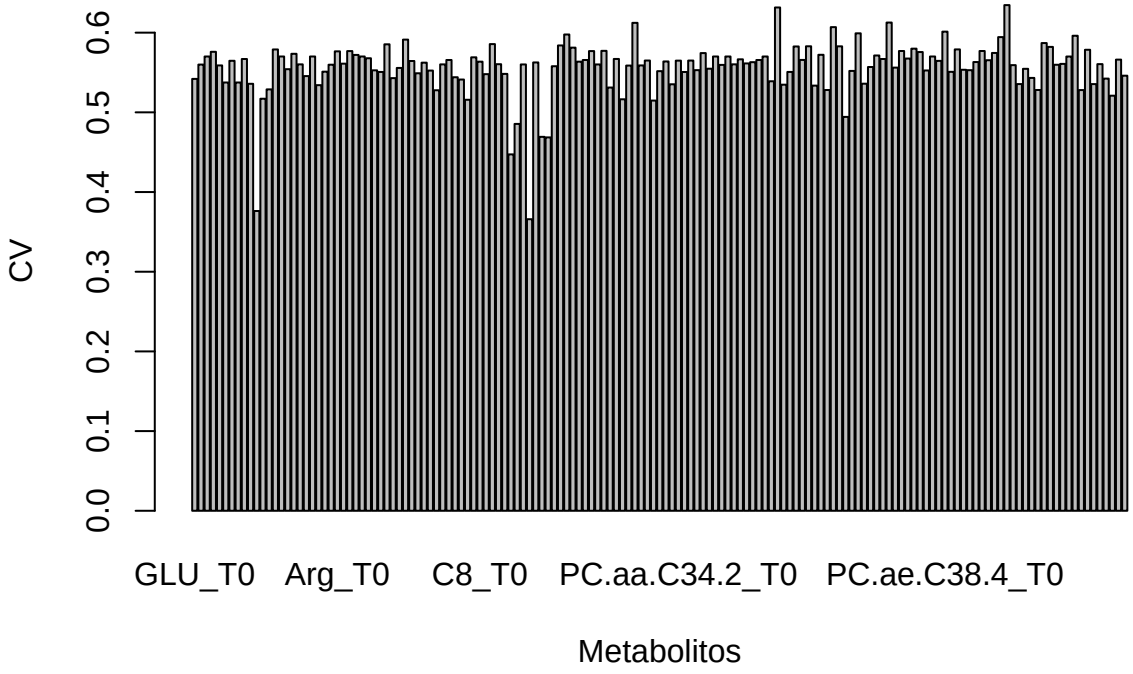
```
library("viridis")
```

```
# Realizar PCA
pca <- prcomp(t(metabolite_clean), scale. = TRUE)

# Graficar los primeros dos componentes con colores por grupo

colors <- viridis::viridis(n = nrow(pca$x))

plot(pca$x[, 1], pca$x[, 2],
     col = colors,
     pch = 19, # Cambia el estilo del punto si lo deseas
     xlab = "PC1", ylab = "PC2",
     main = "Análisis de Componentes Principales (PCA)")
```



4.2.6 Analisis de coeficientes Este analisis es una medida que indica la variabilidad relativa de cada metabolito en comparación con su media. En este grafico se observa como los metabolitos tienen un CV alto y la barras muestran una altura similar lo que sugiere que la vairabilidad relativa de los metabolitos es relativamente homogénea ya que no se presentan outliners notables. Por lo que se puede concluir que es un sistema bien controlado.

```
## Analisis de coeficiente

cv <- apply(metabolite_clean, 2, function(x) sd(x, na.rm = TRUE) / mean(x, na.rm = TRUE))
barplot(cv, main = "Coeficiente de Variación de Metabolitos", ylab = "CV", xlab = "Metabolitos")
```



5 Discusión y Limitaciones

5.1 Discusión

Este análisis exploratorio de datos de metabolómica ha permitido observar patrones de abundancia, correlaciones y agrupamientos entre metabolitos, contribuyendo a una comprensión preliminar de las relaciones en el perfil metabolómico de las muestras.

La implementación de histogramas y boxplots proporcionó una visión general de la distribución y variabilidad de las abundancias, mientras que el análisis de correlación y el mapa de calor revelaron interrelaciones entre metabolitos que podrían tener implicaciones biológicas importantes. El PCA y el clustering jerárquico fueron herramientas efectivas para reducir la dimensionalidad y visualizar agrupamientos en los datos, ayudando a identificar patrones que pueden estar relacionados con factores experimentales.

Sin embargo, este estudio presenta algunas limitaciones. La presencia de datos faltantes y valores no numéricos limitó la cantidad de datos utilizables, lo cual pudo afectar la interpretación de los resultados.

Además, debido a que se trata de un análisis exploratorio, los resultados identificados deben interpretarse con cautela; se requiere validación experimental adicional para confirmar los hallazgos.

Otro aspecto a considerar es que el análisis de clustering y PCA depende de la estandarización previa de los datos, y posibles variaciones en esta etapa podrían impactar los resultados obtenidos.

5.2 Limitaciones

El análisis exploratorio ha demostrado ser un paso valioso para la identificación de tendencias iniciales en datos metabolómicos. Los hallazgos sugieren la existencia de relaciones potencialmente relevantes entre ciertos metabolitos y grupos de muestras. Este estudio sienta una base para futuras investigaciones que podrían explorar más a fondo las correlaciones observadas y examinar metabolitos específicos en contextos biológicos o experimentales relevantes.

La integración de diferentes enfoques estadísticos y visuales en el análisis de datos complejos, como los de metabolómica, ofrece perspectivas prometedoras para avanzar en la comprensión de los sistemas metabóli-

cos. Se recomienda realizar estudios de seguimiento que incluyan mayor validación y refinamiento en la metodología para robustecer los resultados y profundizar en las posibles implicaciones biológicas de los patrones identificados.

6. Repositorio

Para fines del reporte se genero un repositorio en github, en donde se encuentra la información necesaria para replicar el análisis exploratorio.

El enlace se encuentra a continuación: **https://github.com/GeneGarcia96/PEC1_Gene_Garc-a_Datos_Omicos.git**