UOC

Análisis de Datos Ómicos

Marco Antonio Ramírez Rueda

2024-11-05

Contents

METABOLÓMICA **2**
1.- Abstract . . . 2
2.- Objetivos de estudio . . . 2
2.1.- Objetivo General . . . 2
2.2.- Objetivos Específicos . . . 2
3.- Materiales y Métodos . . . 2
4.- Resultados . . . 2
4.1.- Subida de la base de datos . . . 2
4.2.- Primera visuaización mediante descriptivos de la base de datos. . . . 3
4.3.- Gráfico de Correlación . . . 4
4.4.- Histogramas de las variables . . . 5
4.5.- Gráficos de caja y bigote . . . 6
4.6.- Análisis de Cluster . . . 8
4.7.- Enlace simple . . . 8
6.- Discusión y limitaciones . . . 15
7.- Conclusiones del Estudio . . . 16

APÉNDICE CÓDIGO R **17**

DIRECCIÓN URL DEL REPOSITORIO **21**

METABOLÓMICA

1.- Abstract

El factor inducible por hipoxia (HIF)-1a se produce y se degrada de manera continua en condiciones de normoxia, mientras que durante la hipoxia, su estabilización limita la utilización de oxígeno en las células. Sin embargo, se conoce menos sobre el papel de HIF1a(s) y los efectos específicos del sexo en el estado basal durante la normoxia. Dado que el músculo esquelético es el principal reservorio de proteínas en los mamíferos y su homeostasis requiere una gran cantidad de energía, investigamos la función de HIF1a al inicio de la normoxia en el músculo esquelético utilizando metabolómica no dirigida. Se recolectaron y analizaron muestras de músculo esquelético de ratones, tanto con como sin deleción de HIF1a, mediante métodos metabolómicos no dirigidos, incluyendo cromatografía de interacción hidrofílica (HILIC) y cromatografía líquida de fase inversa (RPLC). Identificamos metabolitos que se expresan de manera diferencial en las rutas del ciclo glucolítico y del ciclo de ácido cítrico (TCA), la implementación de bombas osmóticas en ratones de 10-12 semanas, que se sacrificaron 28 días después de la implantación, aunque las muestras tratadas con amoníaco no se incluyeron en este análisis inicial.

2.- Objetivos de estudio

2.1.- Objetivo General

Realizar una exploración inicial y un análisis estadístico del conjunto de datos metabolómicos de muestras de músculo esquelético de ratones con y sin deleción de HIF1a, para identificar posibles tendencias metabólicas relacionadas con la fisiología de HIF1a.

2.2.- Objetivos Específicos

- 1) Implementar métodos estadísticos univariantes y bivariantes, utilizando visualizaciones como boxplots, histogramas y gráficos de correlación, para analizar la variabilidad y distribución de los metabolitos en las muestras.
- 2) Evaluar las relaciones entre las variables del conjunto de datos, con el fin de identificar tendencias metabólicas que puedan estar asociadas a la deleción de HIF1a.
- 3) Desarrollar gráficos de los perfiles metabólicos entre las muestras con y sin deleción de HIF1a, para visualizar el impacto y entender de una manera mas sencilla el metabolismo del músculo esquelético.

3.- Materiales y Métodos

Como primer paso se ha seleccionado un dataset de metabolómica en el repositorio metabolomicsworkbench, la cual trata de **Fisiología del factor inducible por hipoxia-1a (HIF1a) en el músculo esquelético del ratón**, esta base de datos se encuentra en formato JSON.

En cuanto a los métodos estadísticos, podemos diferenciar,análisis univariante y bivariante de los datos, mediante boxplots y/o histogramas y gráficos de correlación para ver la relación de las variables y tendencías.

4.- Resultados

4.1.- Subida de la base de datos

La base de datos en formato JSON sobre **Fisiología del factor inducible por hipoxia-1a (HIF1a) en el músculo esquelético del ratón**, contiene 8003 observaciones con 25 variables.

Una vez descargados los datos crear un contenedor del tipo *SummarizedExperiment*, para ello se crea una matriz que contenga diversar expresiones de 0, así se completa el número de filas de la base, esto es necesario ya que es un requisito para crear el contenedor, en caso de no ser necesario, solamente se crea el contenedor.

Se realiza la asignación de los metadatos, para la creación del contenedor a partir de la base de datos en este caso llamada df = "Fisiología del factor inducible por hipoxia-1a (HIF1a) en el músculo esquelético del ratón"

```
# Crear metadatos para las columnas
col_data <- data.frame(
  METABOLOMICS.WORKBENCH.STUDY_ID = rep(df$M_T[1], ncol(df) - 1),
  PROJECT.PROJECT_TITLE = rep("Fisiología del factor inducible por hipoxia-1a (HIF1a)",
                              ncol(df) - 1),
  SUBJECT.SUBJECT_TYPE = rep("Mouse", ncol(df) - 1)
)
```

Creación del contenedor que contenga los datos y los metadatos (información acerca del dataset, las filas y las columnas

```
library(IRanges)
library(SummarizedExperiment)
library(S4Vectors)
library(MatrixGenerics)
library(matrixStats)
library(GenomicRanges)
library(stats4)
library(Biobase)
library(MatrixGenerics)
library(matrixStats)

# Convertir los datos en un objeto de tipo SummarizedExperiment
contenedor <- SummarizedExperiment(
  assays = list(counts = as.matrix(df[, -1])),
  rowData = DataFrame(Gene_ID = rownames(df)),
  colData = col_data
)
```

4.2.- Primera visuaización mediante descriptivos de la base de datos.

Para tener una visualización de la base de datos y extraer los estadísticos descriptivos se toma en consideración la realización de los siguientes pasos:

- 1) Se debe extraer la matriz de expresión o usar el dataframe original
- 2) Extraer los Estadísticos descriptivos para ver patrones, tendencias que permitan responder el porque los metabotipos de respuesta a la cirugía bariátrica son independientes de la magnitud de la pérdida de peso?

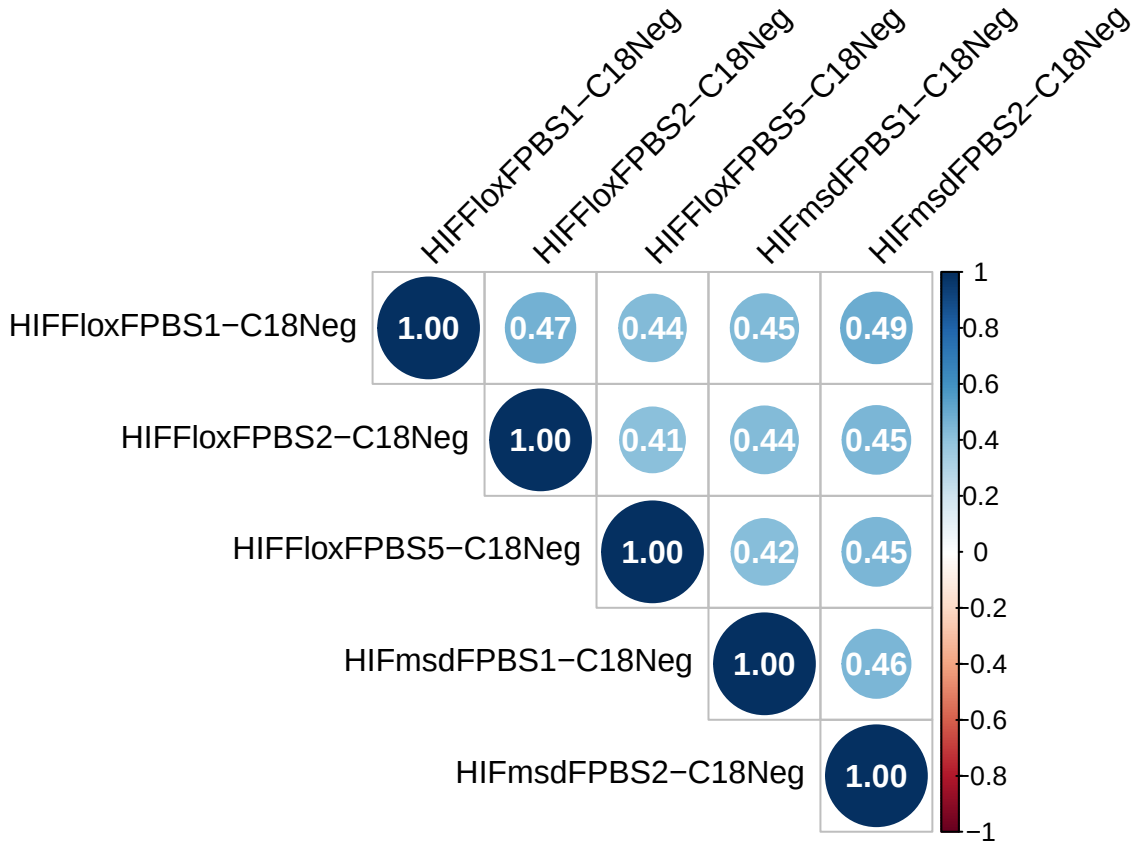
```
library(dplyr)

# Definir las variables de interés
variables_interes <- c("HIFFloxFPBS1-C18Neg",
                       "HIFFloxFPBS2-C18Neg",
                       "HIFFloxFPBS5-C18Neg",
                       "HIFmsdFPBS1-C18Neg",
                       "HIFmsdFPBS2-C18Neg")

# Crear un subconjunto del data frame solo con las variables seleccionadas
subset_df <- df %>%
  select(M_T, all_of(variables_interes))
head(subset_df)
```

```
## # A tibble: 6 x 6
##   M_T          `HIFFloxFPBS1-C18Neg` `HIFFloxFPBS2-C18Neg` `HIFFloxFPBS5-C18Neg`
##   <chr>                        <dbl>                 <dbl>                 <dbl>
## 1 241.0124544~             1241446572            1579544809             116161714
## 2 294.9847554~             4195726239            3824543696            3638209158
## 3 293.1785827~              114480084            8008830856            9394933619
## 4 85.00489042~             4120806546            4094512311            4647959283
## 5 248.9986354~             2391445776            2153411982             223043543
## 6 249.1534636~             5913997663            5039972324            8489573804
## # i 2 more variables: `HIFmsdFPBS1-C18Neg` <dbl>, `HIFmsdFPBS2-C18Neg` <dbl>
```

4.3.- Gráfico de Correlación

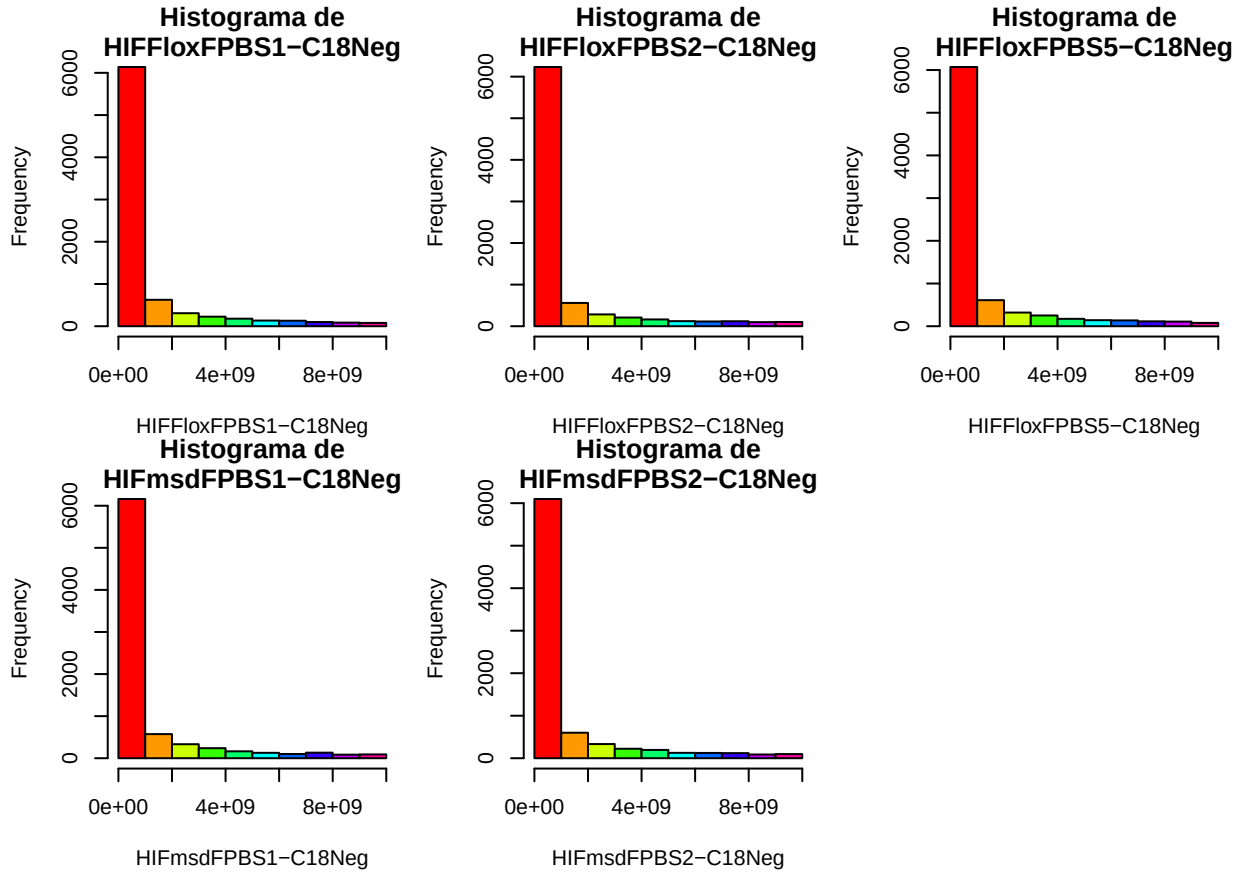
El gráfico de correlación para ver la relación existente entre las variables del estudio

```
library(corrplot)
numeric_subset <- subset_df %>% select(where(is.numeric))
cor_matrix <- cor(numeric_subset, use = "complete.obs")
corrplot(cor_matrix, method = "circle", type = "upper",
         tl.col = "black", tl.srt = 45,addCoef.col = "white")
```



4.4.- Histogramas de las variables

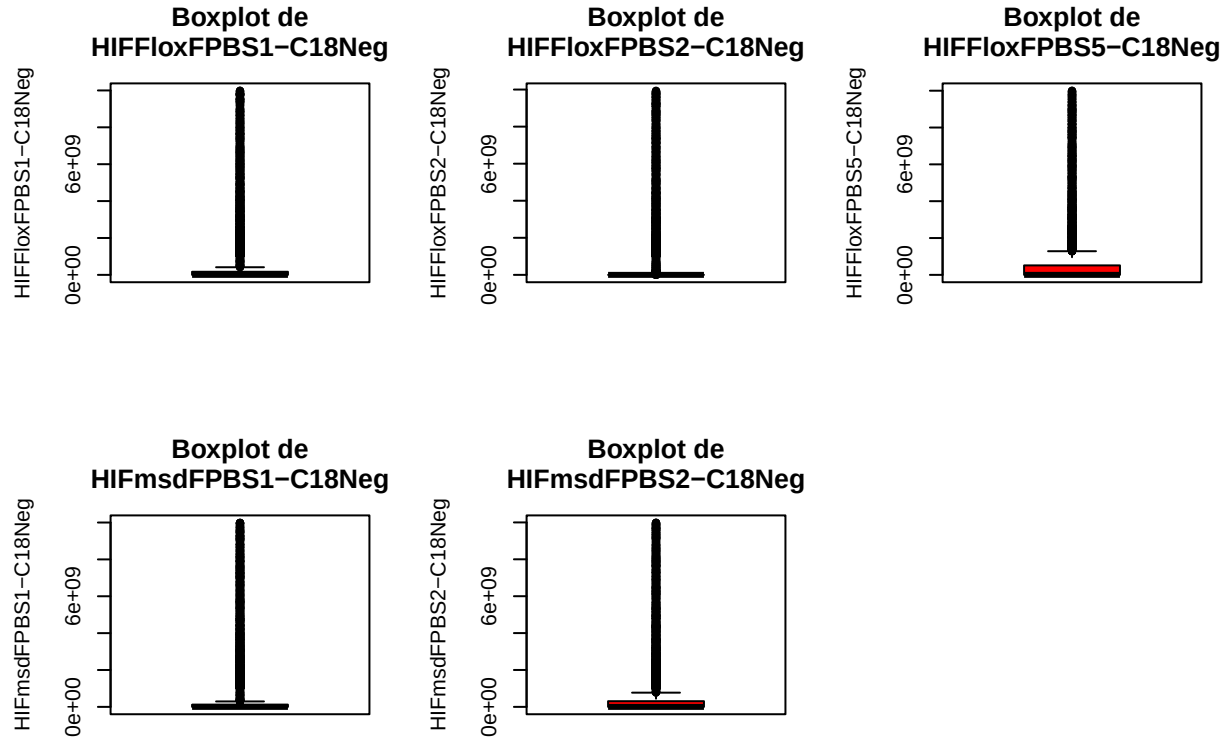
```
par(mfrow = c(2, 3), mar = c(4, 4, 2, 1))
for (var in variables_interes) {
  hist(numeric_subset[[var]],
       include.lowest = TRUE,
       col = rainbow(10),
       main = paste("Histograma de", var, sep = "\n"), xlab = var)}
```



Interpretación: La información sobre los ratones "Floxed Control" (FC) se relaciona con la genética y la manipulación de genes en el contexto de investigaciones biomédicas. En este caso, se refiere a ratones que tienen alelos con secuencias de ADN específicas que están flanqueadas por sitios loxP. Estos sitios permiten la recombinación genética controlada, lo que facilita la eliminación de genes específicos en ciertas condiciones experimentales, al observar el histograma, que muestra una concentración de valores en el primer cuadrante (de 0 a 2e+9), podemos interpretar que la manipulación genética ha influido en la producción de metabolitos, lo que puede ser relevante para entender la función de HIF1a en la respuesta metabólica.

4.5.- Gráficos de caja y bigote

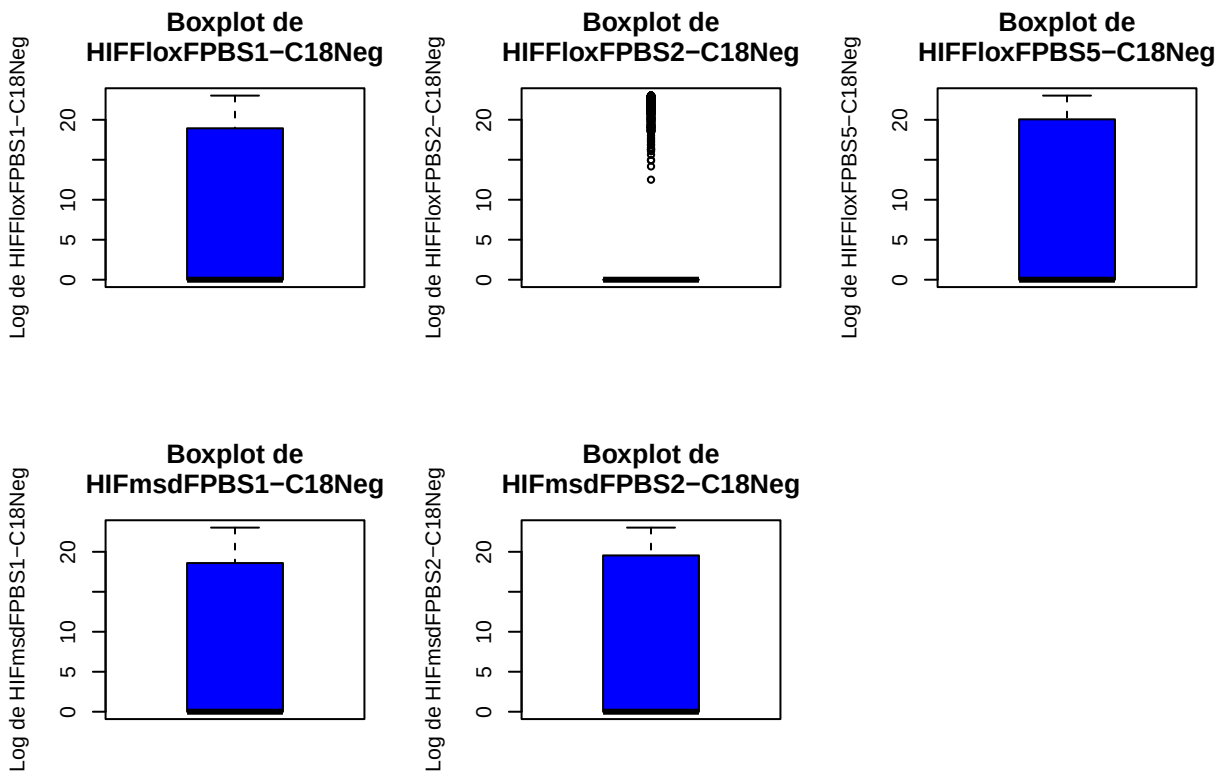
```
par(mfrow = c(2, 3))
for (var in variables_interes) {
  boxplot(numeric_subset[[var]],
          main = paste("Boxplot de", var, sep = "\n"),
          ylab = var,
          col = rainbow(10))
}
```



Interpretación: Como se puede observar, las variables **Floxed Control**, tiene presencia de **outliers**, se puede considerar por tanto, las condiciones experimentales, de los alelos tiene una manipulación poco considerable en la pérdida de peso del ratón despúes de la ciugía experimental bariátrica

Los datos son claramente asimétricos, lo que sugiere que puede tener sentido trabajar con los mismos datos en escala logarítmica.

```
par(mfrow = c(2, 3))  # Configurar la ventana gráfica para mostrar 6 gráficos

for (var in variables_interes) {
  # Aplicar la transformación logarítmica (asegurando que no hay valores <= 0)
  log_variable <- log(numeric_subset[[var]] + 1)  # Se suma 1 para evitar log(0)

  boxplot(log_variable,
          main = paste("Boxplot de", var, sep = "\n"),
          ylab = paste("Log de", var),
          col = "blue")
}
```



Interpretación: Los gráfico de caja y bigotes asimétrico, incluso después de transformaciones, sugiere que hay características importantes en la distribución de los datos que deben ser exploradas más a fondo para entender las implicaciones biológicas o experimentales, en nuestro estudio a pesar que son variables numéricas y se les aplico un tratamiento estadistico siguen siendo muy dificiles de graficarlas

La asimetría puede ser el resultado de la presencia de valores atípicos que influyen en la forma de la distribución. Los valores extremos pueden alterar la mediana y ampliar el rango intercuartílico, lo que se refleja en un gráfico de caja desbalanceado.

4.6.- Análisis de Cluster

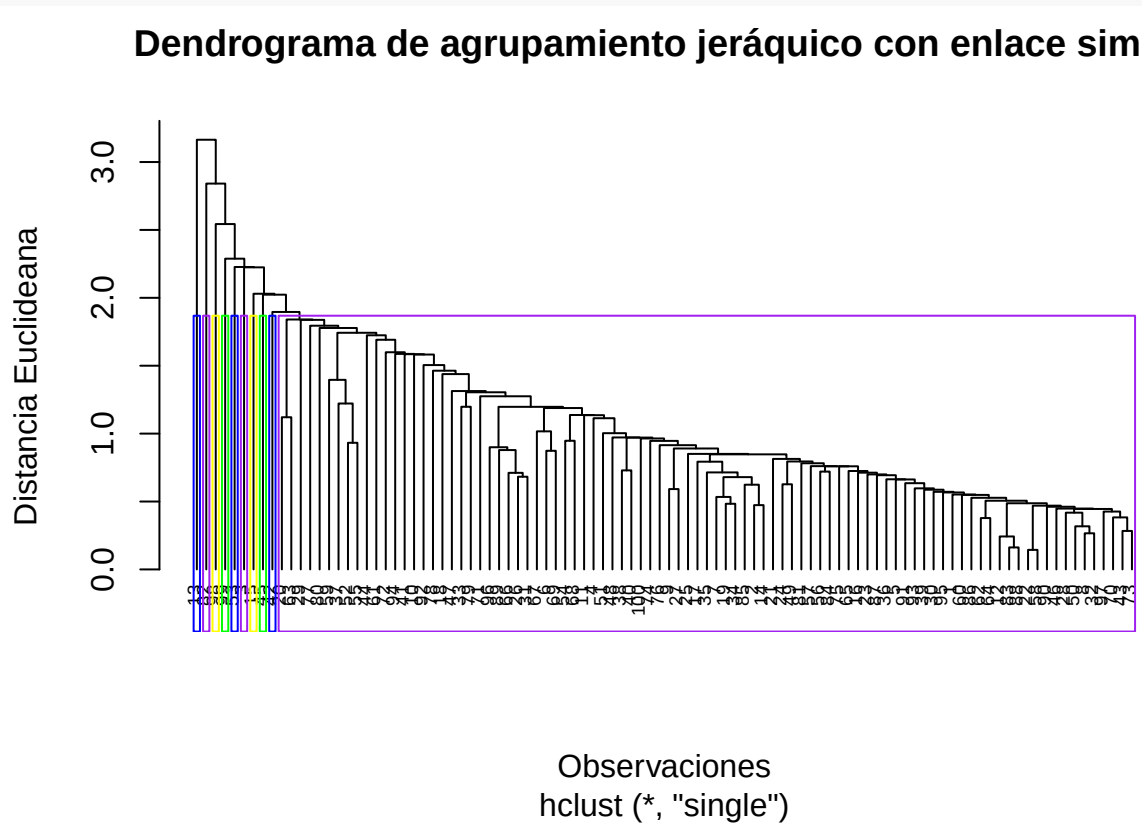
El análisis de conglomerado es una técnica primitiva, en el sentido que no se hace suposiciones sobre el número de grupos o la estructura del grupo. La agrupación se realiza sobre la base de similitudes o distancias, Según [@johnson_applied_2014] "el AC busca dividir un conjunto de objetos (observaciones), en grupos, formados son homogéneos dentro y son heterogéneos entre ellos"

```
library(purrr)
library(cluster)

# Seleccionar las primeras 100 observaciones
subset_numeric <- numeric_subset[1:100, ]

# Estandarizar variables
clusters <- scale(subset_numeric)
# Matriz de similaridades
simil <- dist(x = clusters, method = "euclidean", diag = TRUE)
```

4.7.- Enlace simple

Agrupamiento usando el enlace simple o el vecino más cercano

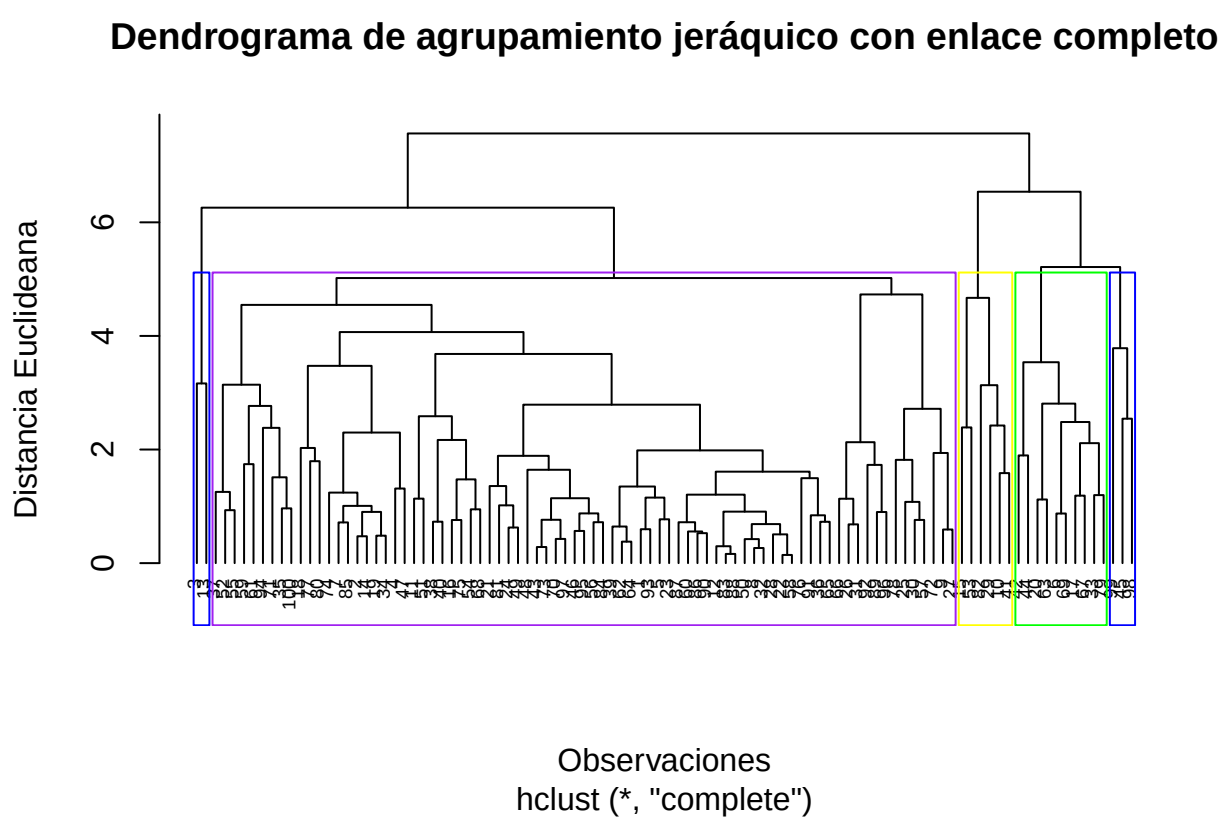
```
enla.simple<-hclust(d = simil, method = "single")
#gráfico del dendrograma
plot(enla.simple,cex=0.6,hang=-1,main = "Dendrograma de agrupamiento jeráquico con enlace simple",xlab=

rect.hclust(enla.simple, k=10,
            border = c("blue","purple","yellow","green"))
```



Podemos seguir con los distintos métodos de enlace

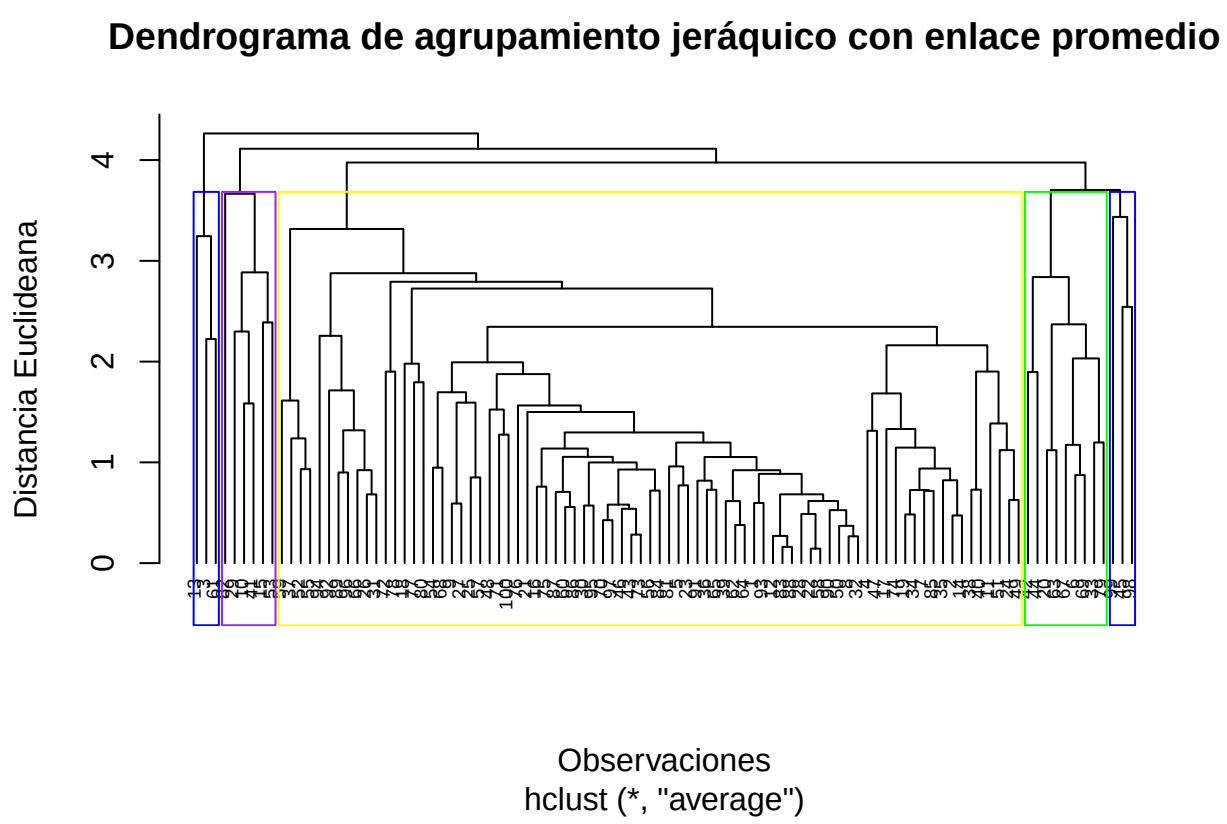
Enlace Completo Agrupamiento usando el enlace completo o el vecino más lejano

```
enla.completo<-hclust(d = simil, method = "complete")
#gráfico del dendrograma
plot(enla.completo,
     cex=0.6,
     hang=-1,
     main = "Dendrograma de agrupamiento jeráquico con enlace completo",
     xlab="Observaciones", ylab="Distancia Euclideana")
rect.hclust(enla.completo, k=5,
            border = c("blue","purple","yellow","green"))
```



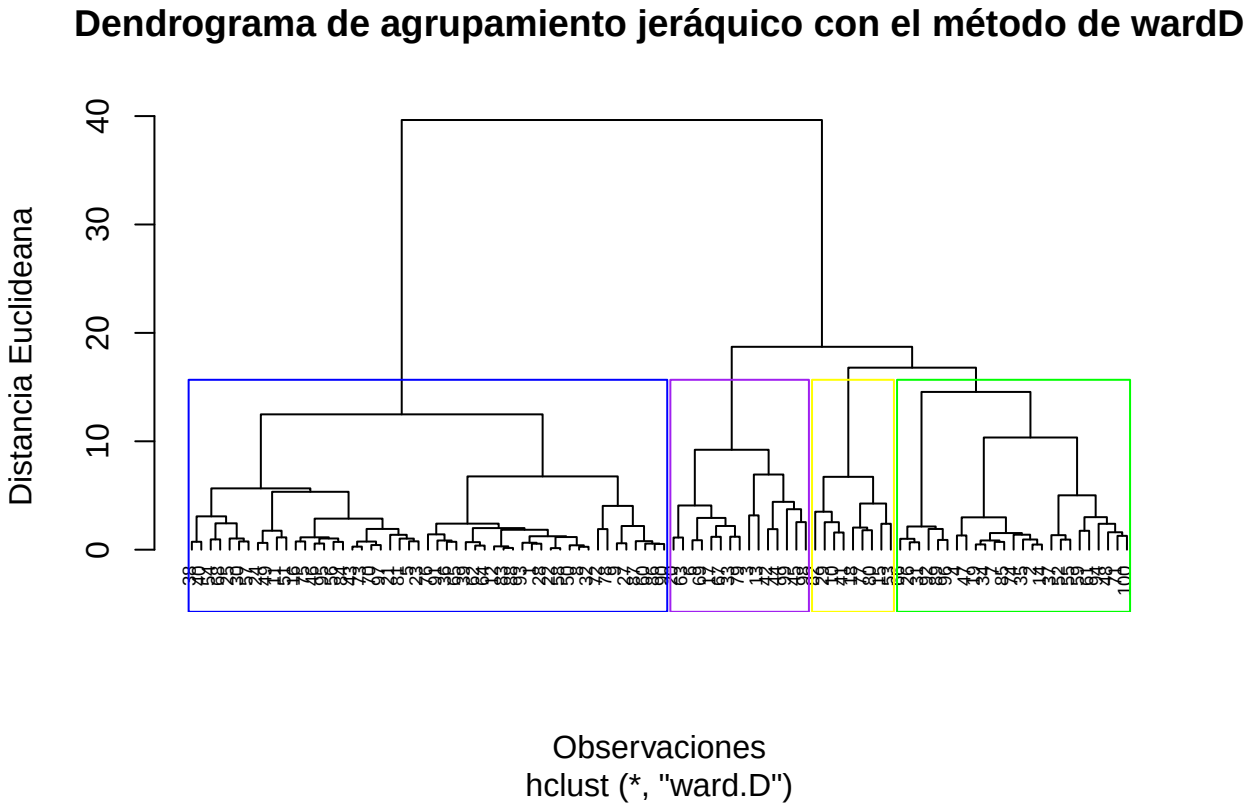
Enlace promedio Agrupamiento usando el enlace completo o el vecino más lejano

```
enla.promedio<-hclust(d = simil, method = "average")
#gráfico del dendrograma
plot(enla.promedio,
     cex=0.6,
     hang=-1,
     main = "Dendrograma de agrupamiento jeráquico con enlace promedio",
     xlab="Observaciones", ylab="Distancia Euclideana")
rect.hclust(enla.promedio, k=5,
            border = c("blue","purple","yellow","green"))
```



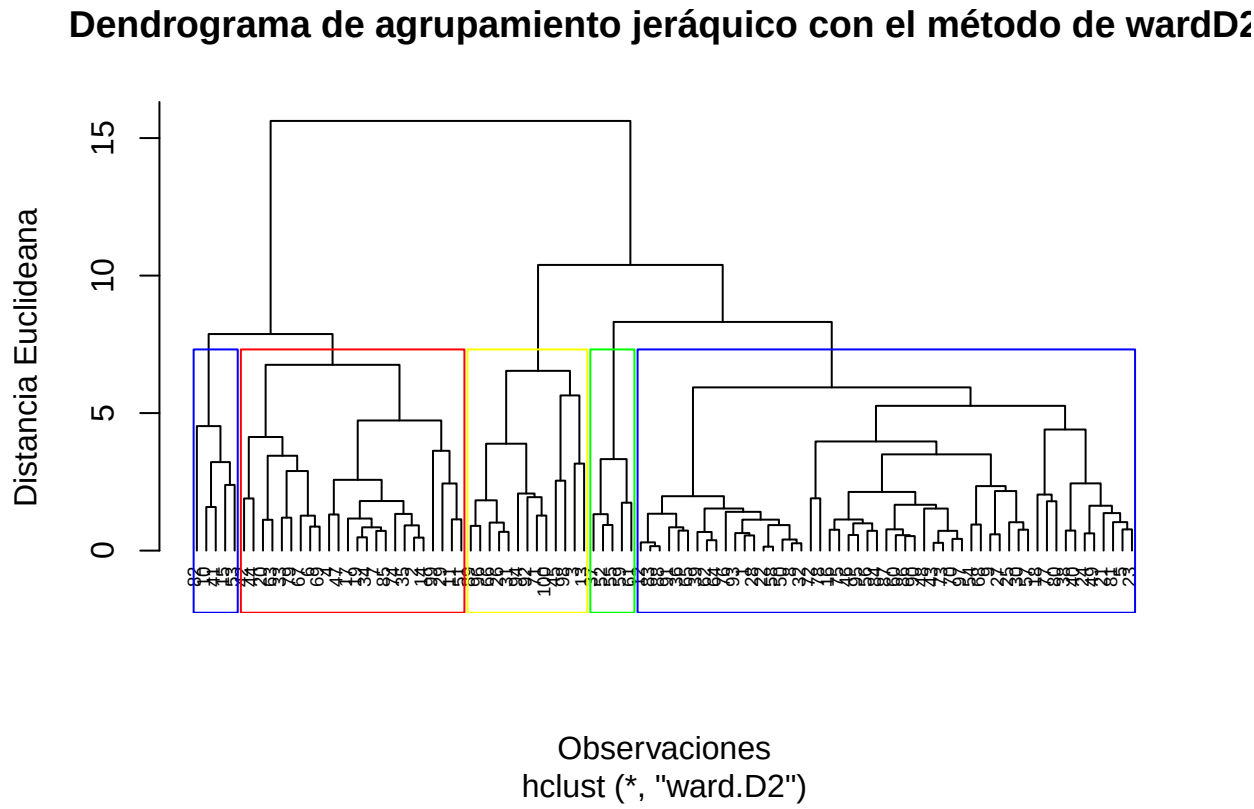
Ward Agrupamiento usando Ward

```
wardD<-hclust(d = simil, method = "ward.D")
#gráfico del dendrograma
plot(wardD,
     cex=0.6,
     hang=-1,
     main = "Dendrograma de agrupamiento jeráquico con el método de wardD",
     xlab="Observaciones", ylab="Distancia Euclideana")
rect.hclust(wardD, k=4,
            border = c("blue","purple","yellow","green"))
```


Dendrograma de agrupamiento jeráquico con el método de wardD
Distancia Euclideana
0
10
20
30
40
Observaciones
hclust (*, "ward.D")

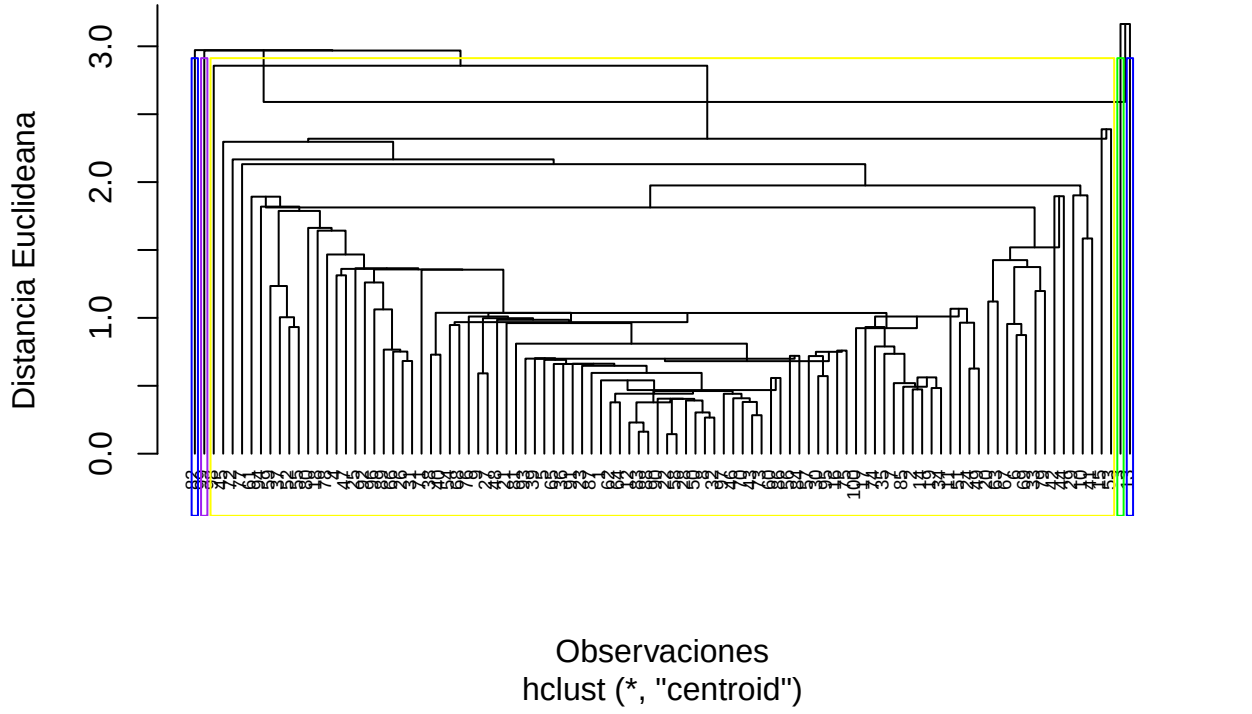
```
wardD2<-hclust(d = simil, method = "ward.D2")
#gráfico del dendrograma
plot(wardD2,
     cex=0.6,
     hang=-1,
     main = "Dendrograma de agrupamiento jeráquico con el método de wardD2",
     xlab="Observaciones", ylab="Distancia Euclideana")
rect.hclust(wardD2, k=5,
            border = c("blue","red","yellow","green"))
```

Agrupamiento usando Ward2



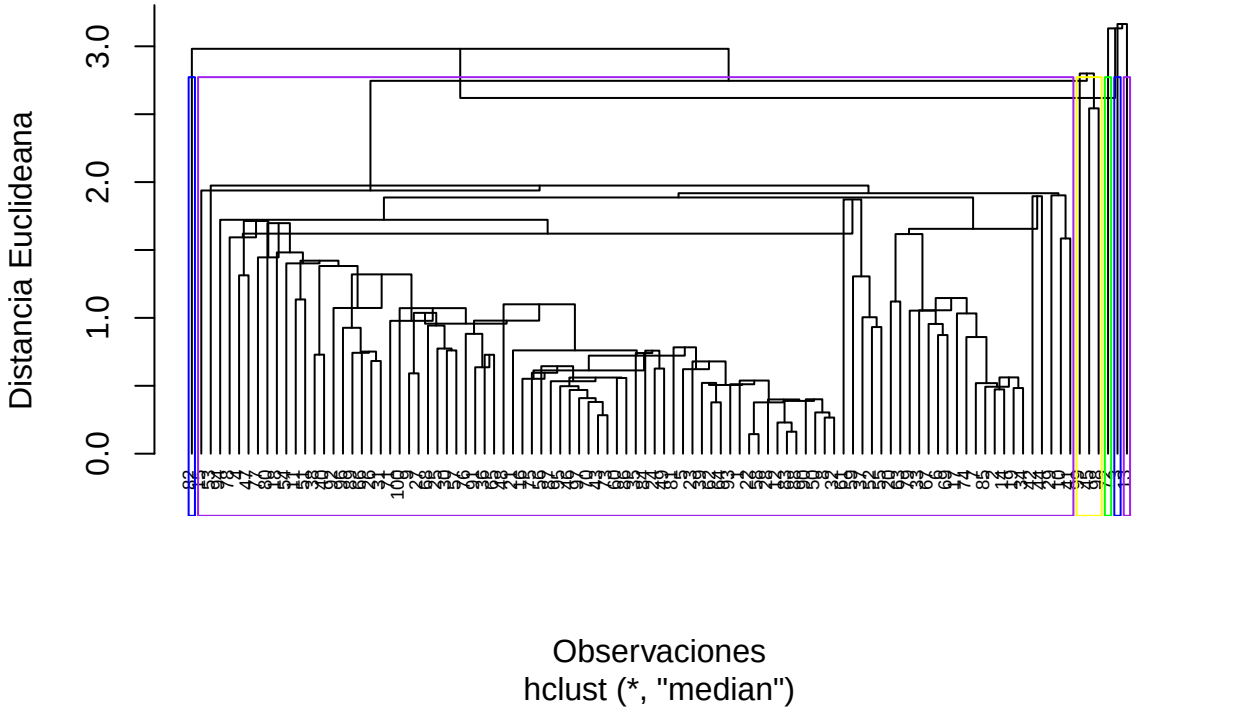
Método de centroide Agrupamiento usando el centroide

```
centroide<-hclust(d = simil, method = "centroid")
#gráfico del dendrograma
plot(centroide,
     cex=0.6,
     hang=-1,
     main = "Dendrograma de agrupamiento jeráquico con el método del centroide",
     xlab="Observaciones", ylab="Distancia Euclideana")
rect.hclust(centroide, k=5,
            border = c("blue","purple","yellow","green"))
```



Método de la mediana Agrupamiento usando mediana

```
mediana<-hclust(d = simil, method = "median")
#gráfico del dendrograma
plot(mediana,
     cex=0.6,
     hang=-1,
     main = "Dendrograma de agrupamiento jeráquico con el método de la mediana",
     xlab="Observaciones", ylab="Distancia Euclideana")
rect.hclust(mediana, k=6,
            border = c("blue","purple","yellow","green"))
```


Dendrograma de agrupamiento jeráquico con el método de la media
Distancia Euclideana
0.0
1.0
2.0
3.0
Observaciones
hclust (*, "median")

6.- Discusión y limitaciones

El factor inducible por hipoxia (HIF)-1a en el músculo esquelético de ratones proporciona valiosos conocimientos sobre la interacción entre la manipulación genética y los procesos metabólicos. Utilizando metabolómica no dirigida, se ha podido identificar cómo la eliminación de HIF1a influye en la expresión de metabolitos, afectando rutas metabólicas esenciales como el ciclo glucolítico y el ciclo de ácido cítrico (TCA). Los resultados sugieren que HIF1a actúa como un regulador crucial en la fisiología del músculo esquelético, con implicaciones importantes para la adaptación metabólica en la cirugía bariátrica y la manipulación en el ADN del ratón.

El análisis visual de los datos mediante boxplots y histogramas ha revelado una notable concentración de valores en el rango de 0 a 2e+9 en las variables analizadas, lo que indica una producción elevada de ciertos metabolitos en los ratones "Floxed Control", sin embargo, la presencia de outliers sugiere una variabilidad significativa en la respuesta metabólica entre los ratones, lo que podría estar relacionado con la manipulación genética y las condiciones experimentale, por otro lado, la implementación de bombas osmóticas en ratones de 10-12 semanas, que se sacrificaron 28 días después de la implantación, añade un contexto experimental relevante al estudio, aunque las muestras tratadas con amoníaco no se incluyeron en este análisis inicial.

La variabilidad observada plantea preguntas sobre la eficacia de la manipulación genética en la regulación del metabolismo y la pérdida de peso post-cirugía bariátrica. Aunque se esperaría que la eliminación de HIF1a afecte de manera uniforme la producción de metabolitos, la variabilidad sugiere que otros factores, como el contexto ambiental y las interacciones genéticas, también son importantes. La comparación de muestras no tratadas con futuros análisis de muestras tratadas con amoníaco será crucial para comprender completamente el impacto de la manipulación de HIF1a y la exposición a amoníaco en la metabolómica del músculo esquelético.

7.- Conclusiones del Estudio

La HIF1a en el músculo esquelético de ratones revela su papel crucial en la regulación del metabolismo, destacando la variabilidad en la respuesta metabólica debido a la manipulación genética y condiciones experimentales, la futura comparación con muestras tratadas con amoníaco permitirá profundizar en la comprensión de estos mecanismos metabólicos.

APÉNDICE CÓDIGO R

```
## Cargamos la base y las librerias
library(readr)
# Cargar el archivo de texto
library(readr)
df <- read_delim("metabolito.csv",
    delim = ";", escape_double = FALSE, trim_ws = TRUE)

##-------------------------------------------------------------------------------------

## Crear metadatos para las columnas
col_data <- data.frame(
  METABOLOMICS.WORKBENCH.STUDY_ID = rep(df$M_T[1], ncol(df) - 1),
  PROJECT.PROJECT_TITLE = rep("Fisiología del factor inducible por hipoxia-1a (HIF1a)",
                              ncol(df) - 1),
  SUBJECT.SUBJECT_TYPE = rep("Mouse", ncol(df) - 1)

##-------------------------------------------------------------------------------------

library(IRanges)
library(SummarizedExperiment)
library(S4Vectors)
library(MatrixGenerics)
library(matrixStats)
library(GenomicRanges)
library(stats4)
library(Biobase)
library(MatrixGenerics)
library(matrixStats)

# Convertir los datos en un objeto de tipo SummarizedExperiment
contenedor <- SummarizedExperiment(
  assays = list(counts = as.matrix(df[, -1])),
  rowData = DataFrame(Gene_ID = rownames(df)),
  colData = col_data

##-------------------------------------------------------------------------------------

library(dplyr)

# Definir las variables de interés
variables_interes <- c("HIFFloxFPBS1-C18Neg",
                       "HIFFloxFPBS2-C18Neg",
                       "HIFFloxFPBS5-C18Neg",
                       "HIFmsdFPBS1-C18Neg",
                       "HIFmsdFPBS2-C18Neg")

# Crear un subconjunto del data frame solo con las variables seleccionadas
subset_df <- df %>%
  select(M_T, all_of(variables_interes))
head(subset_df)
```

```
##-----------------------------------------------------------------------------------------

library(corrplot)
numeric_subset <- subset_df %>% select(where(is.numeric))
cor_matrix <- cor(numeric_subset, use = "complete.obs")
corrplot(cor_matrix, method = "circle", type = "upper",
         tl.col = "black", tl.srt = 45,addCoef.col = "white")

##-----------------------------------------------------------------------------------------

par(mfrow = c(2, 3), mar = c(4, 4, 2, 1))
for (var in variables_interes) {
  hist(numeric_subset[[var]],
       include.lowest = TRUE,
       col = rainbow(10),
       main = paste("Histograma de", var, sep = "\n"), xlab = var)}

##-----------------------------------------------------------------------------------------

par(mfrow = c(2, 3))
for (var in variables_interes) {
  boxplot(numeric_subset[[var]],
          main = paste("Boxplot de", var, sep = "\n"),
          ylab = var,
          col = rainbow(10))
}
##-----------------------------------------------------------------------------------------

library(purrr)
library(cluster)

# Seleccionar las primeras 100 observaciones
subset_numeric <- numeric_subset[1:100, ]

# Estandarizar variables
clusters <- scale(subset_numeric)
# Matriz de similaridades
simil <- dist(x = clusters, method = "euclidean", diag = TRUE)

##-----------------------------------------------------------------------------------------
enla.simple<-hclust(d = simil, method = "single")
#gráfico del dendrograma
plot(enla.simple,cex=0.6,hang=-1,main = "Dendrograma de agrupamiento jeráquico con enlace simple",xlab="

rect.hclust(enla.simple, k=10,
            border = c("blue","purple","yellow","green"))

##-----------------------------------------------------------------------------------------

enla.completo<-hclust(d = simil, method = "complete")
#gráfico del dendrograma
plot(enla.completo,
     cex=0.6,
```

```
     hang=-1,
     main = "Dendrograma de agrupamiento jeráquico con enlace completo",
     xlab="Observaciones", ylab="Distancia Euclideana")
rect.hclust(enla.completo, k=5,
            border = c("blue","purple","yellow","green"))
##-------------------------------------------------------------------------------------
enla.promedio<-hclust(d = simil, method = "average")
#gráfico del dendrograma
plot(enla.promedio,
     cex=0.6,
     hang=-1,
     main = "Dendrograma de agrupamiento jeráquico con enlace promedio",
     xlab="Observaciones", ylab="Distancia Euclideana")
rect.hclust(enla.promedio, k=5,
            border = c("blue","purple","yellow","green"))

##-------------------------------------------------------------------------------------
wardD<-hclust(d = simil, method = "ward.D")
#gráfico del dendrograma
plot(wardD,
     cex=0.6,
     hang=-1,
     main = "Dendrograma de agrupamiento jeráquico con el método de wardD",
     xlab="Observaciones", ylab="Distancia Euclideana")
rect.hclust(wardD, k=4,
            border = c("blue","purple","yellow","green"))
##-------------------------------------------------------------------------------------
wardD2<-hclust(d = simil, method = "ward.D2")
#gráfico del dendrograma
plot(wardD2,
     cex=0.6,
     hang=-1,
     main = "Dendrograma de agrupamiento jeráquico con el método de wardD2",
     xlab="Observaciones", ylab="Distancia Euclideana")
rect.hclust(wardD2, k=5,
            border = c("blue","red","yellow","green"))

##-------------------------------------------------------------------------------------
centroide<-hclust(d = simil, method = "centroid")
#gráfico del dendrograma
plot(centroide,
     cex=0.6,
     hang=-1,
     main = "Dendrograma de agrupamiento jeráquico con el método del centroide",
     xlab="Observaciones", ylab="Distancia Euclideana")
rect.hclust(centroide, k=5,
            border = c("blue","purple","yellow","green"))
##-------------------------------------------------------------------------------------
mediana<-hclust(d = simil, method = "median")
#gráfico del dendrograma
plot(mediana,
     cex=0.6,
     hang=-1,
```

```
     main = "Dendrograma de agrupamiento jeráquico con el método de la mediana",
     xlab="Observaciones", ylab="Distancia Euclideana")
rect.hclust(mediana, k=6,
            border = c("blue","purple","yellow","green"))
```

DIRECCIÓN URL DEL REPOSITORIO

https://github.com/marco673/Ramirez-Rueda-Marco-PEC1