Informe de PEC 1

Julio David Soto López

Resumen:

Los resultados del análisis de los datos de cáncer gástrico permitieron mostrar la construcción de un objeto SummarizedExperiment a partir de un objeto ExpressionSet. Este objeto permitió acceder a los datos de concentración de metabolitos del conjunto de datos de cáncer gástrico para poder manipularlos y explorarlos por medio de una gráfica de cajas, un análisis de componentes principales y un dendograma. Por último, los datos fueron almacenados en un repositorio de Github.

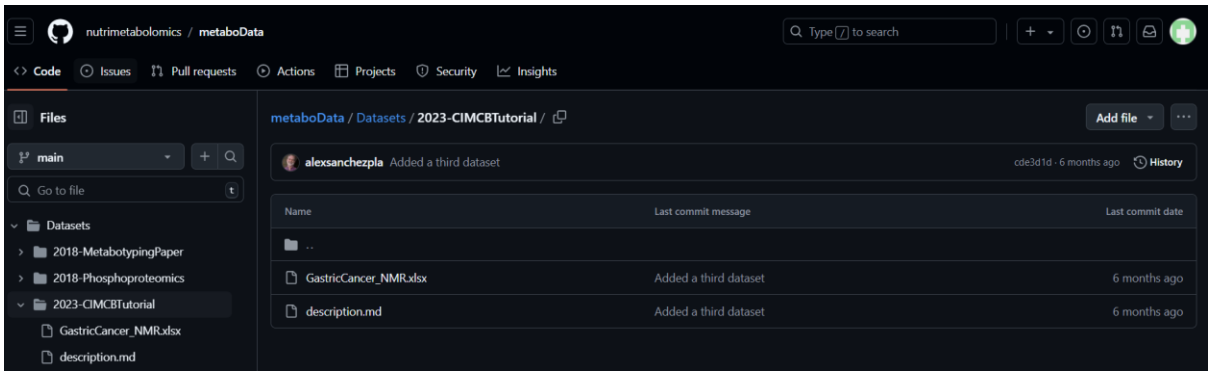
Objetivos:

Construir un objeto SummarizedExperiment, explorar los datos que intentan establecer las diferencias en la concentración de los metabolitos medidos en los 140 pacientes evaluados entre grupos con cáncer, con tumores benignos e individuos sanos. Por último, crear un repositorio Github.

Materiales y métodos:

Los datos fueron obtenidos del repositorio Hithub: https://github.com/nutrimetabolomics/metaboData/. Son datos publicados previamente como un artículo de acceso abierto de Chan et al. (2016), en el British Journal of Cancer, y el archivo de datos deconvolucionados y anotados se depositó en el repositorio de datos de Metabolomics Workbench (ID de proyecto PR000699). Se puede acceder a los datos directamente a través de su proyecto DOI:10.21228/M8B10B.

Los archivos descargados fueron dos, un archivo en formato xlsx y una descripción muy somera de los mismos en un archivo markdown.



Los datos primero fueron filtrados ya que en el documento original fueron conservados los metabolitos con un QC-RSD inferior al 20 % y que tuvieran datos faltantes menor al 10 % de los valores.

```r
# Data set: Gastric_cancer

# Cargando paquetes
library(readxl)
library(Biobase)
library(SummarizedExperiment)

# recuperando los datos

GastricCancer_NMR <- read_excel("GastricCancer_NMR.xlsx")
Peak <- read_excel("GastricCancer_NMR.xlsx", sheet = "Peak")

Peak$filtrado <- subset(Peak[,4] <= 10 & Peak[,5] <= 20) #filtrando
Peak_filtrado <- Peak[Peak$filtrado == TRUE,]
GastricCancer <- GastricCancer_NMR[GastricCancer_NMR$Idx %in%
                                   Peak_filtrado$Idx,]
```

Con estos valores se procedió a construir un objeto SummarizedExperiment a partir de un objeto ExpressionSet. Primero se construyeron los objetos por separado para unirlos en un objeto Phenodata. Luego agregué la información general del experimento y una breve descripción de este a otros objetos:

```r
# datos de la expresion a matriz traspuesta
Expresiongenes <- t(as.matrix(GastricCancer[,5:153]))
colnames(Expresiongenes) <- paste0("sample_",
                                   1:48)# agrego nombre a las columnas

# phenodata
targets <- data.frame(sampleNames = paste0("sample_", 1:48),
                      sampleType = GastricCancer$SampleType,
                      sampleIdx = GastricCancer$Idx,
                      sampleClass = GastricCancer$Class,
                      row.names = 1)

columnDesc <- data.frame(labelDescription= c("Sample/QC",
                                             "Identificador",
                                             "Subclass (QC: Control de calidad,
                                             GC: Cancer gastrico,
                                             BN: Tumor benigno,
                                             HE: Control saludable)"))
myAnnotDF <- new("AnnotatedDataFrame", data=targets, varMetadata= columnDesc)

# agrego nombre de genes
mymetabolitos <- paste0("M", 1:149)
```

```r
# informacion minima del experimento
myInfo=list(myName="Julio Soto",
            myLab="Analisisd e datos omicos",
            myContact="jdjulio@uoc.es",
            myTitle="PEC1")

myDesc <- new("MIAME", name = myInfo[["myName"]],
              lab = myInfo[["myLab"]],
              contact = myInfo[["myContact"]] ,
              title = myInfo[["myTitle"]],
              url = "https://cimcb.github.io/MetabWorkflowTutorial/Tutorial1.html"
              abstract = "Dataset used in the CIMBC tutorial on Basic
              Metabolomics Data Analysis Workflow
              The tutorial describes the data as follows:
              - The study used in this tutorial has been previously published
              as an open access article Chan et al. (2016), in the British
              Journal of Cancer. - The deconvolved and annotated data file have
              been deposited at the Metabolomics Workbench data repository
              (Project ID PR000699). - The data can be accessed directly via
              its project DOI:10.21228/M8B10B - 1H-NMR spectra were acquired
              at Canada's National High Field Nuclear Magnetic Resonance Centre
              (NANUC) using a 600 MHz Varian Inova spectrometer. - Spectral
              deconvolution and metabolite annotation was performed using
              the Chenomx NMR Suite v7.6. ")
```

Luego construí un objeto ExpresionSet y este lo transformé en un RangedSummarizedExperiment:

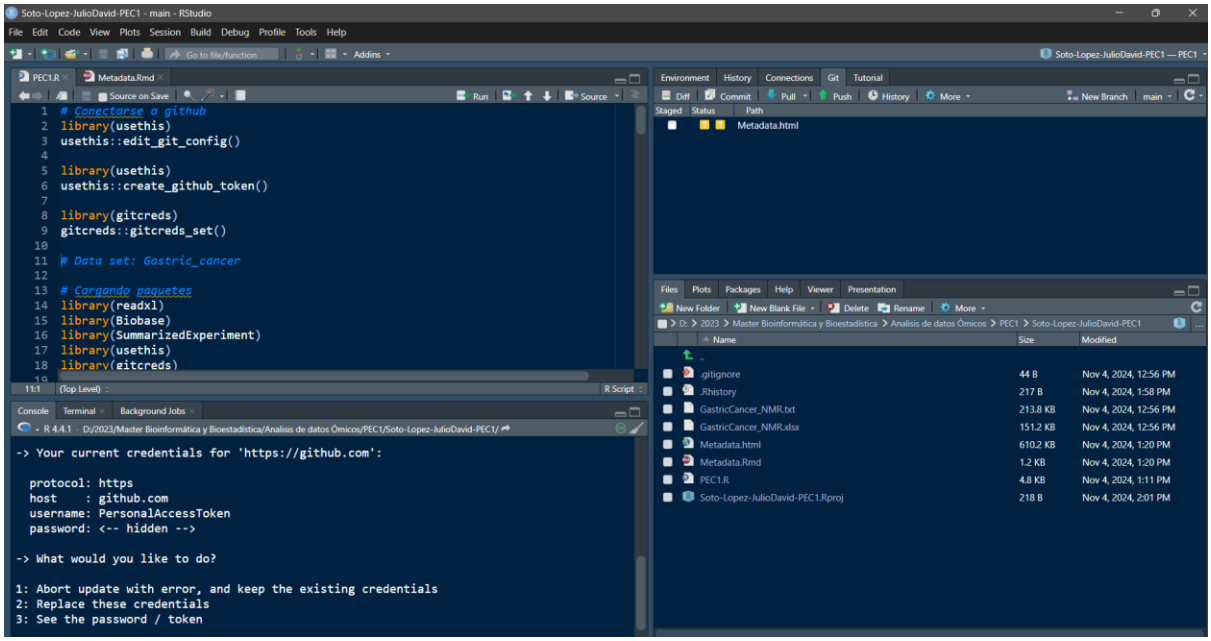
```r
# Datos a tipo ExpressionSet
myEset <- ExpressionSet(assayData = Expresiongenes,
                        phenoData = myAnnotDF,
                        featureNames = mymetabolitos,
                        experimentData = myDesc)

# SummarizedExperiment
eset <- makeSummarizedExperimentFromExpressionSet(from = myEset,
                                                  mapFun = naiveRangeMapper)
# resumen del objeto sumarizedExperiment
class(eset)
show(eset)

```

El objeto resultante fue un objeto RangedSummarizedExperiment (SummarizedExperiment) que sirvió luego para explorar los datos con gráficos boxplot, análisis de componentes principales y una cladograma resultante de un algoritmo de vecino más cercano.

```r
# visualizacion

colores <- c("red", "blue", "yellow", "green")
gctotal <- colores[as.factor(targets$sampleClass)]

boxplot(log10(assays(eset)$exprs),
        main="Concentracion de metabolitos para todas las muestras",
        xlab="Muestra", col = gctotal,
        ylab="Log 10 Concentracion", las=2, cex.axis=0.5, cex.main=0.9)
```

```r
# PCA
library(FactoMineR)
library(factoextra)
pca <- PCA(X = Xknn, scale.unit = TRUE, graph = FALSE)
head(pca$eig)
fviz_pca_ind(pca, geom.ind = "point",
             col.ind = "#FC4E07",
             axes = c(1, 2),
             pointsize = 1.5)

# Cluster
clust.euclid.average <- hclust(dist(t(Xknn)),method="average")
plot(clust.euclid.average, hang=-1)
```

Por último, se construyó un repositorio en HitHub que contiene el informe presentado, el objeto contenedor con los datos y los metadatos en formato binario (.Rda), el código R para la exploración de los datos (por medio de un push up desde Rstudio), los datos en formato texto y los metadatos acerca del dataset en un archivo markdown. Para esto se creo una cuenta de Hithub utilizando un Two-Factor authenticator. El proyecto fue creado primero en Hithub de acuerdo con las indicaciones del guión para la PEC1. Utilizando la dirección de URL del proyecto se creo un proyecto con versión de cambios en Rstudio utilizando la URL previamente copiada. Al nuevo proyecto se le copio el código de R utilizado para la exploración de los datos, se comprometió el archivo y se empujó al proyecto de Hithub. A continuación, se creó un archivo markdown y se incluyó en este la información sobre los datos subidos al proyecto en formato xlsx. Este documento se comprometió y se empujo al proyecto en línea. El archivo html resultante no fue incluido debido a una advertencia de Rstudio.



Resultados:

El objeto resultante RangedSummarizedExperiment (SummarizedExperiment) pudo ser construido gracias a la función makeSummarizedExperimentFromExpressionSet de la librería SummarizedExperiment:

```
> # resumen del objeto sumarizedExperiment
> class(eset)
[1] "RangedSummarizedExperiment"
attr(,"package")
[1] "SummarizedExperiment"
> show(eset)
class: RangedSummarizedExperiment
dim: 149 48
metadata(3): experimentData annotation protocolData
assays(1): exprs
rownames(149): M1 M2 ... M148 M149
rowData names(0):
colnames(48): sample_1 sample_2 ... sample_47 sample_48
colData names(3): sampleType sampleIdx sampleClass
```

Al explorar el objeto resultante podemos observar que tiene la matriz de concentraciones de metabolitos, el Phenodata y la metadata se encuentran almacenados como una instancia del objeto SummarizedExperiment:

```
83  # explorando el objeto sumarizedExperiment
84  assays(eset)$exprs[1:7,1:7]
85  colData(eset)
86  metadata(eset)
87
```

```
91:1  (Top Level)
```

```
Console  Terminal  Background Jobs
R 4.4.1 · D:/2023/Master Bioinformática y Bioestadística/Analisis de datos Ómicos/PEC1/
```

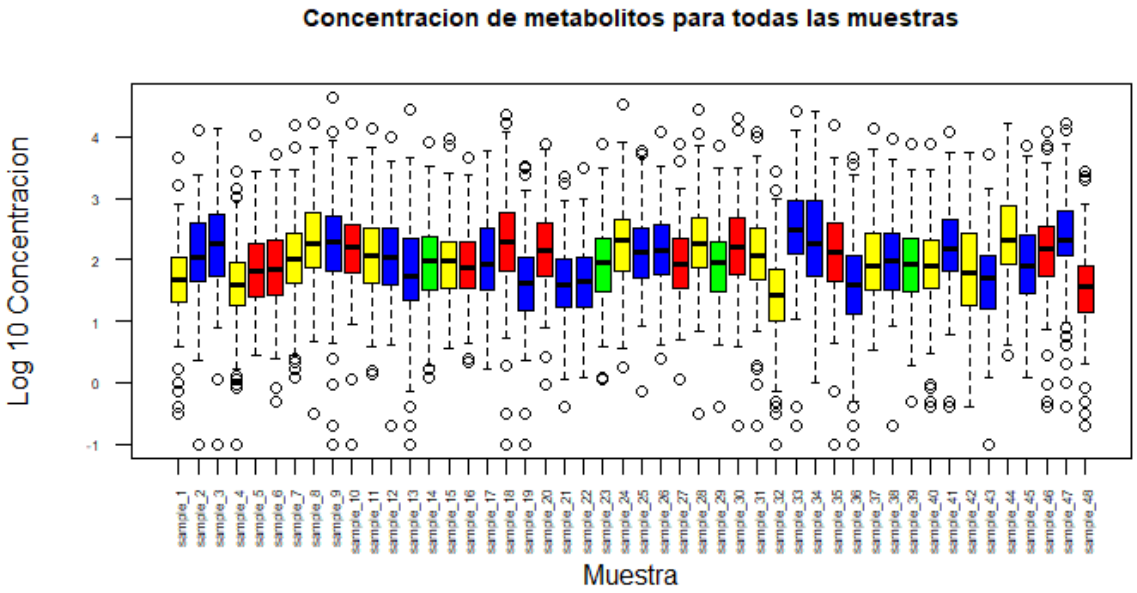
```
> # explorando el objeto sumarizedExperiment
> assays(eset)$exprs[1:7,1:7]
   sample_1 sample_2 sample_3 sample_4 sample_5 sample_6 sample_7
M1     31.6     81.9     45.5     91.0     36.5     52.1       NA
M2     59.7    258.7    190.4    231.9    190.1     94.7    250.3
M3     86.4    315.1     32.0    212.5    153.1     68.2     59.1
M4     14.0      8.7       NA     18.2     47.4     26.0     70.6
M5     88.6    243.2    362.7     72.5    146.5     95.5     65.4
M6     10.3     18.4     35.7      6.7     26.9      9.0     20.5
M7    170.3    349.4     59.6     15.3     20.6     10.4     26.2
> colData(eset)
DataFrame with 48 rows and 3 columns
          sampleType sampleIdx sampleClass
         <character> <numeric> <character>
sample_1       Sample         4          HE
sample_2       Sample         5          GC
sample_3       Sample         7          GC
sample_4       Sample         8          HE
sample_5       Sample        11          BN
...               ...       ...         ...
sample_44      Sample       129          HE
sample_45      Sample       130          GC
sample_46      Sample       134          BN
sample_47      Sample       137          GC
sample_48      Sample       138          BN
```

```
> metadata(eset)
$experimentData
Experiment data
  Experimenter name: Julio Soto
  Laboratory: Analisisd e datos omicos
  Contact information: jdjulio@uoc.es
  Title: PEC1
  URL: https://cimcb.github.io/MetabWorkflowTutorial/Tutorial1.html
  PMIDs:

  Abstract: A 278 word abstract is available. Use 'abstract' method.

$annotation
character(0)

$protocolData
An object of class 'AnnotatedDataFrame': none
```

Al explorar los datos con estadística univariada de las primeras 7 muestras se puede ver que las concentraciones son muy estables por muestra, con una media alrededor de 55, excepto de la muestra 2. Esto muestra que lo mejor es observar todas las muestras a la vez para establecer que grupos se diferencian.

```
> summary(assays(eset)$exprs[1:7,1:7])
    sample_1         sample_2         sample_3         sample_4         sample_5
 Min.   : 10.30   Min.   :  8.70   Min.   : 32.00   Min.   :  6.70   Min.   : 20.60
 1st Qu.: 22.80   1st Qu.: 50.15   1st Qu.: 38.15   1st Qu.: 16.75   1st Qu.: 31.70
 Median : 59.70   Median :243.20   Median : 52.55   Median : 72.50   Median : 47.40
 Mean   : 65.84   Mean   :182.20   Mean   :120.98   Mean   : 92.59   Mean   : 88.73
 3rd Qu.: 87.50   3rd Qu.:286.90   3rd Qu.:157.70   3rd Qu.:151.75   3rd Qu.:149.80
 Max.   :170.30   Max.   :349.40   Max.   :362.70   Max.   :231.90   Max.   :190.10
                                   NA's   :1
    sample_6         sample_7
 Min.   : 9.00   Min.   : 20.50
 1st Qu.:18.20   1st Qu.: 34.42
 Median :52.10   Median : 62.25
 Mean   :50.84   Mean   : 82.02
 3rd Qu.:81.45   3rd Qu.: 69.30
 Max.   :95.50   Max.   :250.30
                 NA's   :1
```

En la gráfica de cajas se pueden observar todas las muestras y con un color distinto cada grupo de experimentación: QC: Control de calidad (verde), GC: Cancer gástrico (azul), BN: Tumor benigno (rojo), HE: Control saludable (amarillo).

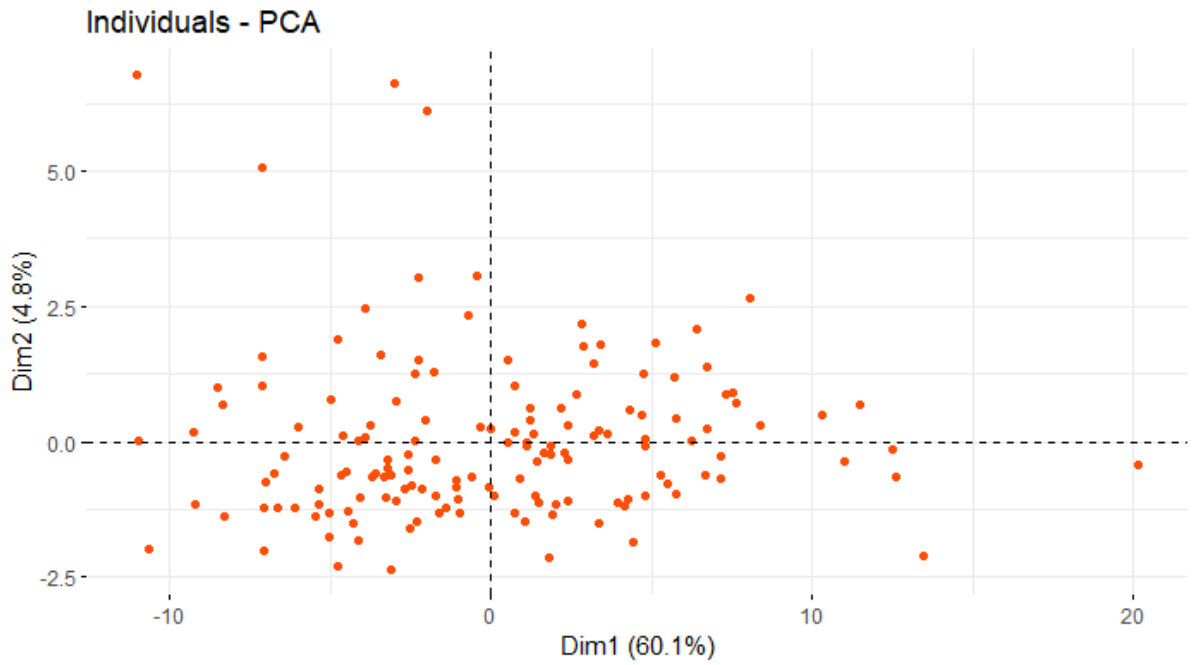


Se observa que la concentración de metabolitos es muy variable entre los grupos evaluados pero que los valores más altos en la concentración global de metabolitos son para los pacientes con cáncer gástrico. Hace falta realizar análisis por cada metabolito parra establecer cuales son los que presentan mayor concentración en estos pacientes.

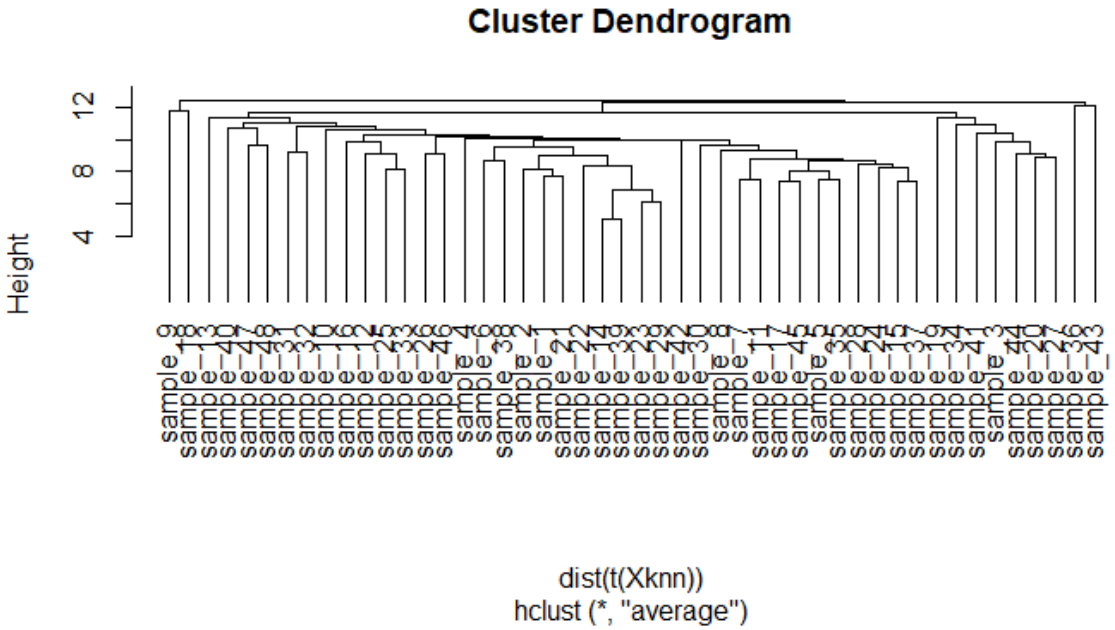
Para poder llevar a cabo el PCA era necesario que los valores NA desaparecieran, por lo que los valores logarítmicos base 10 de la tabla fueron escalados (cada valor menos la media de la columna dividido entre la desviación estándar de la columna), y cada NA fue sustituido por el valor del vecino más cercano en un rango de tres:

```
# cambiar los NA por 0
library(tidyverse)
library(VIM)
library(scales)

x <- assays(eset)$exprs
x_ <- log10(x)
x_1 <- assays(eset)$exprs #backup
Xscale <- scale(x_) # escalar

Xknn <- kNN(Xscale, k = 3, imp_var = FALSE) # imputar valores NA
head(Xknn)

apply(Xknn, MARGIN = -1, function(x) sum(is.na(x))) # evaluar na
```

El gráfico de los dos primeros componentes principales no muestra grupos distintos entre los individuos evaluados a simple vista:



Y el dendograma si muestra agrupaciones, pero no los tres o cuatro grandes grupos evaluados. Sin embargo, si es posible observar tres ramas que podrían simbolizar grupos formados por la condición evaluada en las concentraciones de metabolitos:



En el caso del proyecto en Hithub se pudo construir y subir todo lo requerido en el guión de la PEC1:

Enlace al repositorio de Hithub:

https://github.com/jdjuliosoto/Soto-Lopez-JulioDavid-PEC1

Discusión y limitaciones del estudio:

En este trabajo se puede observar que el enfoque fue puramente metodológico. Y en las metodologías utilizadas no se pudo ahondar en los análisis estadísticos predictivos, técnicas de machine learning u otros análisis numéricos de los datos para poder llegar a una conclusión acerca de los datos utilizados. Esto se debe principalmente a que el objetivo de la PEC1 era demostrar el manejo de los datos, así como la construcción de objetos que puedan ser leídos por otras plataformas durante el análisis de los datos. Es evidente que estos datos pueden ser aprovechados para ejemplificar análisis de discriminación mucho más potentes o algoritmos de asociación de metabolitos con los grupos estudiados. Sin embargo, fueron útiles para mostrar lo aprendido durante la unidad del curso que nos compete.