# PEC 1

Judit Maria Sebares Huerta

## Índice



## Resumen

Este estudio explora una base de datos de cáncer gástrico mediante análisis metabolómico. Se realizó una limpieza exhaustiva de datos, seguida de análisis estadísticos descriptivos y de componentes principales utilizando la función SummarizedExperiment de Bioconductor. Los resultados revelan patrones distintivos en tres metabolitos clave: M7 (sobreexpresado en cáncer gástrico), M138 (elevado en procesos tumorales sin distinción entre benignos y malignos) y M70 (reducido en cáncer gástrico comparado con tejido sano).

El análisis de componentes principales muestra cierto solapamiento entre grupos, aunque con algunos valores atípicos en pacientes con cáncer que podrían indicar diferentes estadios de la enfermedad. Aunque se identificaron posibles biomarcadores, se recomienda realizar pruebas estadísticas adicionales para evaluar la significancia de las diferencias observadas. La metodología empleada facilitó tanto el análisis como la visualización dinámica de datos mediante la librería iSEE.

## Objetivos

Los objetivos del estudio son :

- Exploración de una base de datos de cáncer gástrico, entendiendo los datos que se están estudiando, y la función de las diferentes columnas.
- Hacer una limpieza de la base de datos que nos permita comprender lo que estamos estudiando y además, facilitar un futuro estudio cuantitativo.
- Exploración estadística mediante estadística descriptiva y análisis de componentes.
- Visualización, mediante gráficos de violín, de la abundancia de diferentes metabolitos.
- Ver cómo afecta el tipo de cáncer a la matriz de conteos obtenida.
- Emplear la función de *SummarizedExperiment* del Bioconductor
- Detectar qué metabolitos se pueden emplear como biomarcadores de cáncer gástrico.

## Métodos

Se ha empleado para este estudio una base de datos de cáncer gástrico que ha sido previamente publicada en el siguiente artículo ***"H-NMR urinary metabolomic profiling for diagnosis of gastric cancer"*** donde se ha analizado el perfil metabolómico de pacientes con cáncer gástrico, con tumores gástricos benignos y con pacientes sanos.

Se ha realizado un análisis descriptivo de los datos y luego un análisis de los componentes principales.

### Estructura de la base de datos

Podemos ver los datos que acabamos de cargar para entenderlos mejor. La base de datos cuenta con dos hojas , la primera llamada **Data** y la segunda llamada **Peak.**

La tabla **Data** contiene 153 columnas (en nuestro código la hemos renombrado **Raw Data**) :

- Índice
- Id de cada muestra (SampleID)
- Class : es el tipo de paciente. Los valores que puede tomar son QC: Control de calidad, BN : tumor benigno, HE : paciente sano y GC: cáncer gástrico. Emplearemos estas clases para hacer el análisis de los componentes.
- M1 a M149 son los diferentes metabolitos que se han estudiado.

Por otro lado, tenemos la tabla llamada **peak** con algunos metadatos.

La tabla consta de las siguientes columnas (En nuestro código la hemos renombrado **metadata**) :

- Índice
- El nombre del metabolito
- La etiqueta del metabolito
- El Perc_missing indica el porcentaje de muestras que no incluyen medidas para el metabolito
- Columna QC_RSD indica la variación en las medidas del metabolito a lo largo de todas las muestras.

## Resultados

Antes de procesar los datos tendremos que limpiar la base de datos. La columna Perc_missing nos indica el % de metabolitos para los que no hay datos.

Podemos escoger únicamente las muestras que tengan menos del 10% de datos faltantes.

### Creación de la clase Summarized Experiment

La clase *SummarizedExperiment* de la librería Bioconductor se emplea para almacenar matrices de resultados experimentales.

Cada objeto almacena observaciones de una o más muestras además de metadatos.

El primer aspecto clave es que Summarized Experiment almancena los conteos en matrices donde las muestras van en las columnas mientras que los metabolitos (en nuestro caso particular) van en las filas.

Empezaremos creando la clase

```
#Seleccionar solo las columnas relevantes (SampleID y mediciones)
samples <- raw_data %>%
  dplyr::select(SampleID, starts_with("M"))

#Transponer los datos para que SampleID sean columnas
df_transposed <- samples %>%
  pivot_longer(cols = -SampleID, names_to = "Measurement", values_to = "Value") %>%
  pivot_wider(names_from = SampleID, values_from = Value)
```

```r
#Convertir en matriz de conteos
counts_matrix <- df_transposed %>%
  column_to_rownames(var = "Measurement") %>%
  as.matrix()

#Crear `rowData` con metadatos de mediciones
rowData_df <- metadata %>%
  dplyr::select(-Idx) %>%
  column_to_rownames(var = "Name")

#Crear `colData` con metadatos de muestras
colData_df <- raw_data %>%
   dplyr::select(SampleID, SampleType, Class) %>%
  column_to_rownames(var = "SampleID")

#Crear `SummarizedExperiment`
se <- SummarizedExperiment(
  assays = list(counts = counts_matrix),
  colData = colData_df,
  rowData = rowData_df
)

#Verificar que el objeto se creó correctamente
se
```

```
class: SummarizedExperiment
dim: 125 140
metadata(0):
assays(1): counts
rownames(125): M2 M3 ... M148 M149
rowData names(3): Label Perc_missing QC_RSD
colnames(140): sample_1 sample_2 ... sample_139 sample_140
colData names(2): SampleType Class
```

Column (sample) data

```r
colData(se)[1:4,] #Restringimos las columnas a mostrar a 4, por espacio.
```

```
DataFrame with 4 rows and 2 columns
          SampleType       Class
         <character> <character>
```

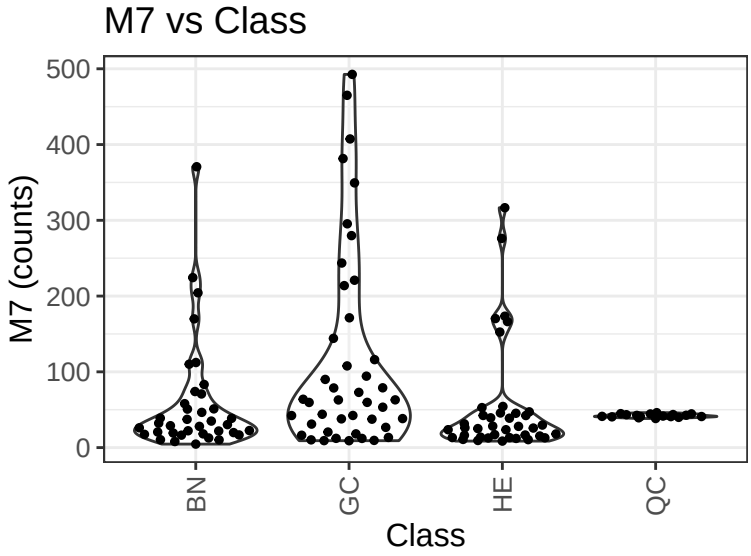
```
sample_1          QC          QC
sample_2      Sample          GC
sample_3      Sample          BN
sample_4      Sample          HE
```

En este objeto estamos accediendo a los metadatos que describen cada muestra.

Hemos visualizado la abundancia de expresión de los diferentes metabolitos en los diferentes tipos de pacientes con la librería **iSEE** que crea gráficos de violín. A continuación he seleccionado los que he considerado que tienen mayor relevancia.

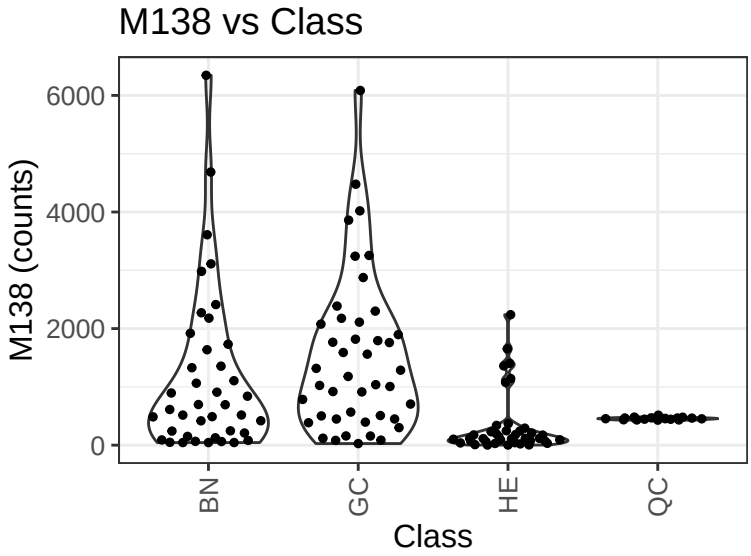
**Metabolito 7**

```
se <- iSEE::cleanDataset(se)
colormap <- synchronizeAssays(ExperimentColorMap(), se)

set.seed(100)
plot.data <- data.frame(Y = assay(se, "counts")["M7", ], X = factor(colData(se)[, "Class"]))
  subset(!is.na(Y)) |>
  transform(GroupBy = X, jitteredX = iSEE::jitterViolinPoints(X, Y, width = 0.4, method = 'q

set.seed(124)
plot.data <- plot.data[sample(nrow(plot.data)), , drop = FALSE]

dot.plot <- ggplot(plot.data, aes(x = X, y = Y)) +
  geom_violin(aes(group = GroupBy), alpha = 0.2, scale = 'width', width = 0.8) +
  geom_point(aes(x = jitteredX), alpha = 1, color = '#000000', size = 1) +
  labs(x = "Class", y = "M7 (counts)", title = "M7 vs Class") +
  coord_cartesian(ylim = range(plot.data$Y, na.rm = TRUE), expand = TRUE) +
  theme_bw() +
  theme(legend.position = 'bottom',
        axis.text.x = element_text(angle = 90, size = 10, hjust = 1, vjust = 0.5),
        axis.text.y = element_text(size = 10),
        axis.title = element_text(size = 12), title = element_text(size = 12))

dot.plot
```




En el gráfico de violín para el metabolito M7 podemos ver una sobredispersión del numero de conteos y además un alto número de conteos de la clase GC , lo que indica una variabilidad metabólica en las muestras de cáncer. Vemos que las muestraGs de pacientes anos tienen menos conteos, lo que nos indican que es menos a abundante en tejido sano.

Metabolito M138[1]

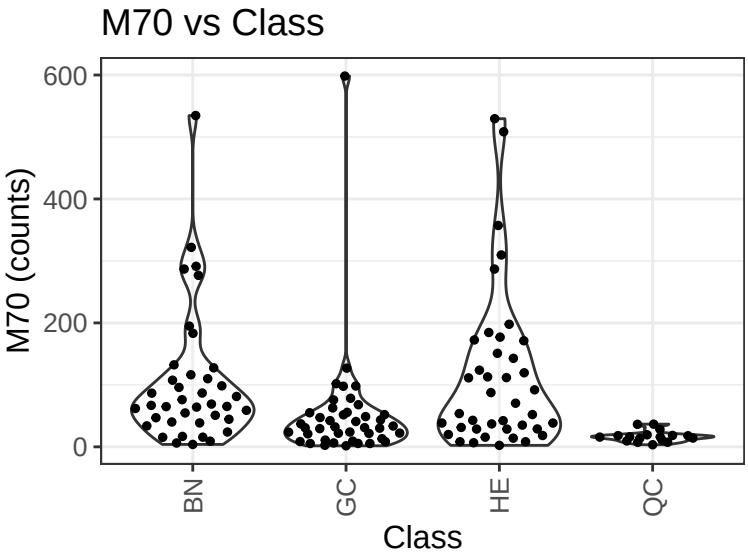



En este caso, también observamos que es un metabolito sobreexpresado en procesos tumorales aunque al contrario que M1, parece no distinguir entre tumores benignos y malignos ya que

[1]El código de R empleado para hacer el gráfico no se muestra dado que es el mismo que elempleado para el metabolito 1

parece expresado de manera similar. Podría indicarnos otro biomarcador útil para tumores aunque no para distinguirlos entre ellos.

**Metabolito 70**[2]




A diferencia de los metabolitos anteriores (M7 y M138), el metabolito M70 muestra un patrón muy distinto: los valores más bajos se encuentran en el grupo GC (cáncer gástrico), con la mayoría de las muestras concentradas por debajo de 100 conteos.

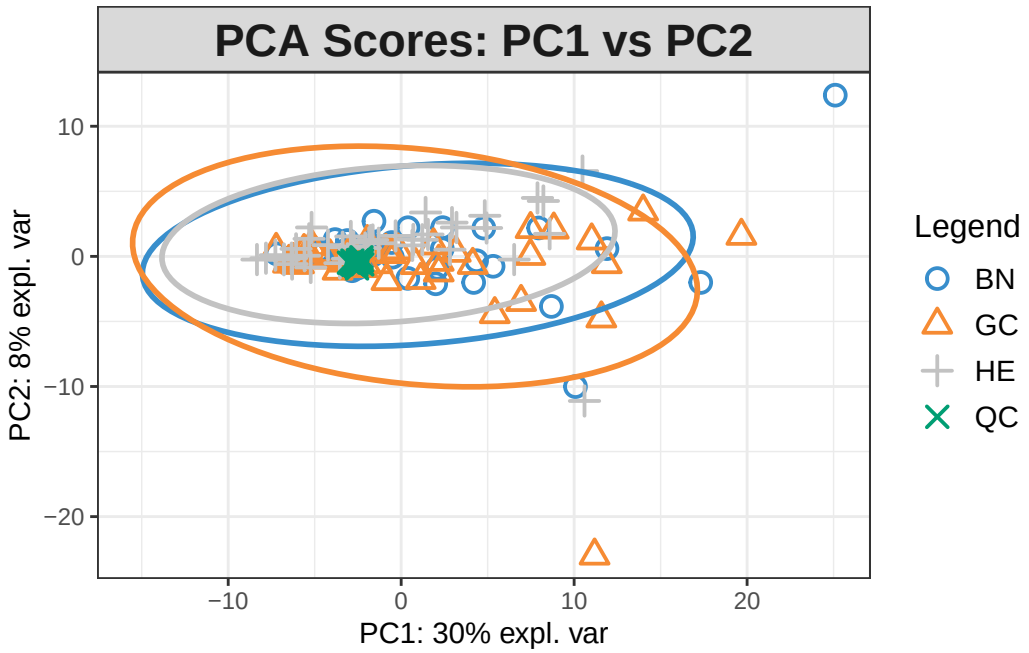
- El hecho de que los tumores benignos muestren un nivel intermedio sugiere que la reducción de M70 podría correlacionarse con la progresión de la malignidad.

### Análisis de los componentes principales empleando MixOmics

Se ha realizado un análisis de los componentes principales empleando el paquete **MixOmics**.

```
X <- raw_data[, -c(1:4)]
Y <- raw_data$Class

# Realizar PCA usando mixOmics
pca_modelo <- pca(X, ncomp = 5, scale = TRUE)

# Gráfico de scores
plotIndiv(pca_modelo, group = Y,ind.names = FALSE, ellipse = TRUE, legend = TRUE,
          title = "PCA Scores: PC1 vs PC2")
```

[2]El código de R empleado para hacer el gráfico no se muestra dado que es el mismo que elempleado para el metabolito 1




En este gráfico podemos ver losdos primeros componentes principales (PC1 y PC2).

- El PC1 explica el 30% de la varianza total de los datos mientras que el PC2 explica el 8% de la varianza total.

Los círculos azules responden a muestras con tumor benigno, los triángulos naranjas corresponden a las muestras con cáncer gástrico, las cruces grises a los sanos y las cruces verdes al control de calidad

Se observan algunos outlier especialmente en los grupos GC (cáncer gástrico) y HE (pacientes sanos), pero en general, podemos ver que no hay una clara separación entre los diferentes grupos.

## Discusión

El **análisis del conteo de metabolitos** nos ha permitido determinar los metabolitos que están más expresados en según qué tipos de pacientes. Dado que nuestra base de datos contaba con más de 149 metabolitos, solamente hemos seleccionado 3 para realizar el análisis.

El **metabolito 7** (M7) parece estar sobreexpresada en cáncer gástrico en comparación con las muestras del tejido sano. Por otro lado, se puede apreciar que los tumores benignos presentan un nivel intermedio de esta característica.

Al observar una separación en los conteos entre los pacientes de cáncer y los sanos podemos deducir que se puede emplear como biomarcador.

El **metabolito 138** (M138) al igual que el metabolito 1, parece estar sobreexpresado en procesos tumorales, aunque en este caso no distingue tanto si se trata de tumores malginos o benignos (sí que se aprecia que los pacientes con tumor gástrico tienen más valores intermedios). Aunque los pacientes sanos en general tienen menos conteos, sí que hay algunas muestras que alcanzan hasta los 2000 conteos, por lo que más estudios estadísticos son necesarios para determinar la significación de los datos.

El **metabolito 70** (M70) al contario de los dos metabolitos estudiados anteriormente, parece estar dismunido en el cáncer gástrico en comparación los tejidos sanos y los tumores benignos. Se podría pensar que este metabolito está presente en vías metabólicas que están suprimidas en el cáncer gástrico.

En el **análisis de los componentes principales** no se ve una clara separación entre los diferentes grupos, lo que nos puede indicar un cierto solapamiento entre los perfiles metabólicos.

Sí que se observan algunos pacientes con cáncer gástrico (GC) con valores atípicos que podrían ser significativos, esto podría indicar diferentes estadios dentro de la enfermedad o diferentes tipos de metabolitos. Los pacientes sanos y con tumores benignos, muestras patrones de distribución más compactos.

## Conclusiones

Mediante los estudios que hemos realizado podemos detectar algunos metabolitos que se podrían emplear como biomarcadores, aunque en este estudio solo hemos reseñado tres (M1, M70 y M138) se recomienda emplear el código proporcionado para testar mediante el paquete iSEE los diferentes gráficos de violín. Para sacar conclusiones contundentes sobre la utilidad de los metabolitos mencionados como biomarcadores se deberian realizar más test estadísticos que evalúen la diferencia de conteos en los diferentes grupos.

Esto es importante ya que en el estudio de análisis principales realizado posteriormente se ve cierto solapamiento entre las diferentes clases, lo que nos hace poder intuir que las diferencias observadas en el conteo de los metabolitos pueden no ser significativas.

Por otro lado, cabe destacar el empleo de la clase Summarized Experiment nos ha permitido por un lado conocer mejor el data set y por otro lado, ha facilitado la visualización de los datos con la librería iSee que se puede ejecturar en el código de R, lanzando una app para la visualizción dinámica de los datos teniendo en cuanta la clase de paciente estudiado.

## Referencias

Repositorio de Github : https://github.com/juditseb/AnalisisDatosOmicos.git

Artículo con la base de datos : Chan, A., Mercier, P., Schiller, D.*et al.*1H-NMR urinary metabolomic profiling for diagnosis of gastric cancer.*Br J Cancer***114**, 59–62 (2016). https://doi.org/10.1038/bjc.2015.414 \end{document}