PEC 1 Análisis de datos ómicos

Jorge Velázquez Gómez

2024-10-31

Contents

Resumen ejecutivo **1**

Objetivos del estudio **2**

Materiales y métodos **2**

Resultados **2**

Importación de los datos y exploración inicial 2

Limpieza de los datos 3

Creación del contenedor de tipo `SummarizedExperiment` 3

Exploración del dataset 4

Análisis estadísticos 8

Test de Kruskal-Wallis para el metabolito M48 8

Matriz de correlación entre metabolitos 9

Cluster Jerárquico Divisivo 10

Análisis de Componentes Principales 11

Discusión y conclusiones del estudio **12**

Reposición de los datos en github **13**

Resumen ejecutivo

En este estudio se ha analizado el perfil metabolómico de un conjunto de pacientes para tratar de descubrir dianas prometedoras en la detección del cáncer gástrico. En primer lugar, se han almacenado los datos y metadatos del estudio en un objeto de tipo `SummarizedExperiment` para una organización eficiente de los mismos. A continuación, se han explorado una serie de características de los metabolitos y de las muestras con el fin de obtener información relevante sobre los análisis estadísticos que se pueden llevar a cabo. Después, se ha comprobado si las diferencias en la concentración de una posible diana de interés son estadísticamente significativas a través del test de Kruskal-Wallis. Finalmente, se ha representado un cluster jerárquico para comprobar si las muestras de una misma clase se agrupan de forma natural en función de la distancia entre sus perfiles de expresión de metabolitos, y un Análisis de Componentes Principales (PCA) para comprobar si lo hacen en función de sus componentes principales.

Objetivos del estudio

Los objetivos del estudio han sido:

-Creación de un contenedor del tipo `SummarizedExperiment` que contenga los datos y metadatos del dataset de estudio.

-Exploración de los datos para encontrar variables relevantes de estudio.

-Ejecucion de un proceso simple de análisis ómico con el fin de abordar cuestiones planteadas durante la exploración de los datos.

Materiales y métodos

Se han utilizado datos pertenecientes a muestras de diferentes pacientes emparejadas mediante espectroscopia de resonancia magnética nuclear de 1H (1H - NMR), generando 77 metabolitos reproducibles. Estos datos se encuentran depositados en el **repositorio de datos de Metabolomics Workbench** (ID de proyecto PR000699).

Para la gestión y exploración de los datos, se utilizó el paquete `SummarizedExperiment` de Bioconductor. Para el análisis estadístico se emplearon algunas herramientas de R como la función `hclust()` para realizar un agrupamiento jerárquico de las muestras, la función `aov()` para realizar un test ANOVA o la función `prcomp()` para realizar un Análisis de Componentes Principales.

Resultados

Importación de los datos y exploración inicial

En primer lugar, cargamos los datos:

```
datos<-read_excel("GastricCancer_NMR.xlsx",sheet="Data")
peak<-read_excel("GastricCancer_NMR.xlsx",sheet="Peak")
head(datos)
```

```
## # A tibble: 6 x 153
##     Idx SampleID SampleType Class    M1      M2    M3    M4    M5    M6    M7
##   <dbl> <chr>    <chr>      <chr> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     1 sample_1 QC         QC     90.1   492.  203.   35   164.   19.7  41
## 2     2 sample_2 Sample     GC     43     526.  130.   NA   694.  114.   37.9
## 3     3 sample_3 Sample     BN    214.  10703.  105.   46.8 483.  152.  110.
## 4     4 sample_4 Sample     HE     31.6    59.7  86.4  14    88.6  10.3 170.
## 5     5 sample_5 Sample     GC     81.9   259.  315.    8.7 243.   18.4 349.
## 6     6 sample_6 Sample     BN    197.    128.  862.   18.7 200.    4.7  37.3
## # i 142 more variables: M8 <dbl>, M9 <dbl>, M10 <dbl>, M11 <dbl>, M12 <dbl>,
## #   M13 <dbl>, M14 <dbl>, M15 <dbl>, M16 <dbl>, M17 <dbl>, M18 <dbl>,
## #   M19 <dbl>, M20 <dbl>, M21 <dbl>, M22 <dbl>, M23 <dbl>, M24 <dbl>,
## #   M25 <dbl>, M26 <dbl>, M27 <dbl>, M28 <dbl>, M29 <dbl>, M30 <dbl>,
## #   M31 <dbl>, M32 <dbl>, M33 <dbl>, M34 <dbl>, M35 <dbl>, M36 <dbl>,
## #   M37 <dbl>, M38 <dbl>, M39 <dbl>, M40 <dbl>, M41 <dbl>, M42 <dbl>,
## #   M43 <dbl>, M44 <dbl>, M45 <dbl>, M46 <dbl>, M47 <dbl>, M48 <dbl>, ...
```

Observamos que el conjunto de datos presenta 140 filas correspondientes a las diferentes muestras tomadas y 153 columnas correspondientes a las variables de estudio (las 4 primeras contienen información descriptiva sobre las muestras y las 149 restantes las concentraciones de diferentes metabolitos).

```
str(peak)
```

```
## tibble [149 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Idx         : num [1:149] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Name        : chr [1:149] "M1" "M2" "M3" "M4" ...
##  $ Label       : chr [1:149] "1_3-Dimethylurate" "1_6-Anhydro- -D-glucose" "1_7-Dimethylxanthine" "1-
##  $ Perc_missing: num [1:149] 11.429 0.714 5 8.571 1.429 ...
##  $ QC_RSD      : num [1:149] 32.21 31.18 34.99 12.8 9.37 ...
```

La tabla peak contiene información sobre los metabolitos como el porcentaje de valores perdidos o el QC_RSD, que es una puntuación de calidad que representa la variación en las mediciones de este metabolito en todas las muestras.

Limpieza de los datos

Como se indica en el **flujo de trabajo propuesto por el CIMBC**, es conveniente evaluar la calidad de los datos y eliminar (limpiar) cualquier metabolito mal medido antes de realizar cualquier análisis estadístico. Para ello vamos a utilizar aquellos metabolitos que cumplan estos dos requisitos:

- Poseer un QC-RSD menor del 20%
- Poseer menos del 10% de valores perdidos

Dicha información se encuentra en la tabla peak, así que impondremos las restricciones en dicha tabla:

```
peak_filtrado<-peak[peak$QC_RSD < 20 & peak$Perc_missing < 10, ]
nrow(peak_filtrado)
```

```
## [1] 52
```

Observamos que solo 52 metabolitos cumplen ambos requisitos. A continuación, filtraremos en la tabla "Datos" los metabolitos seleccionados y obtendremos la matriz transpuesta, ya que en los objetos de tipo `SummarizedExperiment` las filas representan las características de interés y las columnas representan las muestras:

```
#Filtramos las columnas y calculamos la transpuesta
datos_limpios<-datos[,peak_filtrado$Name]
datos_limpios_t<-as.data.frame(t(datos_limpios))
#Indicamos como nombre de las columnas los valores de SampleID
colnames(datos_limpios_t)<-datos$SampleID
```

Creación del contenedor de tipo `SummarizedExperiment`

En la variable `datos_limpios_t` tengo los valores de concentracion de las 140 muestras (columnas) para los 52 metabolitos seleccionados (filas). Vamos a proceder a almacenar los metadatos de las columnas del `SummarizedExperiment` en una variable. Estos metadatos se encuentan en las columnas `SampleID`, `SampleType` y `Class` del dataset original.

```
colData <- data.frame(SampleID = datos$SampleID, SampleType = datos$SampleType, Class = datos$Class)
rownames(colData)<-datos$SampleID
```

Ahora crearemos una variable para los metadatos de las filas, que contendrá la información almacenada en la tabla `Peak_filtrado`:

```
rowData <- data.frame(Name = peak_filtrado$Name, Label = peak_filtrado$Label,
          Perc_missing = peak_filtrado$Perc_missing, QC_RSD = peak_filtrado$QC_RSD)
rownames(rowData) <- peak_filtrado$Name
```

Finalmente, procedemos a crear el objeto `SummarizedExperiment`:

```
se<-SummarizedExperiment(assays = list(counts = datos_limpios_t),
                         rowData = rowData, colData = colData)
```

Exploración del dataset

En primer lugar, podemos visualizar la matriz de datos y los metadatos del objeto `SummarizedExperiment`:

```
# Matriz de datos (assay)
head(assay(se)[1:10])
```

```
##     sample_1 sample_2 sample_3 sample_4 sample_5 sample_6 sample_7 sample_8
## M4      35.0       NA     46.8     14.0      8.7     18.7       NA     18.2
## M5     164.2    694.5    483.4     88.6    243.2    200.1    362.7     72.5
## M7      41.0     37.9    110.1    170.3    349.4     37.3     59.6     15.3
## M8      46.5    125.7     85.1     23.9     61.1    243.7     51.3     37.1
## M11     61.7    490.6   2441.2    140.7     48.7    103.7     58.1     54.1
## M14     35.3       NA     29.3     62.9     77.8     52.3     34.6     30.3
##     sample_9 sample_10
## M4       8.4      36.0
## M5     270.2     203.4
## M7     213.8      44.4
## M8      65.6      48.6
## M11     92.9      59.0
## M14     61.9      28.4
```

```
# Metadatos de los metabolitos (filas)
head(rowData(se))
```

```
## DataFrame with 6 rows and 4 columns
##            Name                Label Perc_missing    QC_RSD
##     <character>          <character>    <numeric> <numeric>
## M4           M4 1-Methylnicotinamide      8.57143  12.80420
## M5           M5       2-Aminoadipate      1.42857   9.37266
## M7           M7      2-Furoylglycine      2.85714   5.04916
## M8           M8 2-Hydroxyisobutyrate      0.00000   5.13234
## M11         M11   3-Aminoisobutyrate      5.00000  15.47616
## M14         M14 3-Hydroxyisobutyrate      2.14286   8.90571
```

```r
# Metadatos de las muestras (columnas)
head(colData(se))
```

```
## DataFrame with 6 rows and 3 columns
##             SampleID  SampleType       Class
##          <character> <character> <character>
## sample_1    sample_1          QC          QC
## sample_2    sample_2      Sample          GC
## sample_3    sample_3      Sample          BN
## sample_4    sample_4      Sample          HE
## sample_5    sample_5      Sample          GC
## sample_6    sample_6      Sample          BN
```

Podemos comprobar las dimensiones del objeto `SummarizedExperiment` para confirmar el número de muestras y metabolitos:

```r
dim(se)
```

```
## [1]  52 140
```

Las dimensiones se corresponden con el número de muestras y metabolitos que habíamos filtrado en los pasos previos del análisis. A continuación, elaboraremos un resumen de los metadatos de las fila para conocer cierta información sobre los metabolitos:

```r
summary(as.data.frame(rowData(se)))
```

```
##      Name               Label             Perc_missing        QC_RSD
##  Length:52          Length:52          Min.   :0.0000   Min.   : 2.432
##  Class :character   Class :character   1st Qu.:0.0000   1st Qu.: 5.582
##  Mode  :character   Mode  :character   Median :0.7143   Median : 9.436
##                                        Mean   :1.7720   Mean   :10.145
##                                        3rd Qu.:2.8571   3rd Qu.:14.490
##                                        Max.   :9.2857   Max.   :19.146
```

En este resumen sobre los metadatos de los metabolitos observamos cierta información relevante como que la mediana de `Perc_missing` es 0.7143%, lo cual sugiere que más del 50% de los metabolitos tienen menos del 1% de valores faltantes o que la media de `QC_RSD` es 10.145, lo que implica que, en promedio, los metabolitos tienen un RSD en torno al 10%, que es razonable para datos experimentales.

Sería interesante calcular la frecuencia de cada tipo de muestra y cada tipo de clase. Para ello se puede usar la funcion table() en los metadatos de las columnas:

```r
table(colData(se)$SampleType)
```

```
##
##     QC Sample
##     17    123
```

Hay 17 muestras correspondientes a control de calidad y 123 muestras de estudio.

```
table(colData(se)$Class)
```

```
##
## BN GC HE QC
## 40 43 40 17
```

Hay 40 muestras etiquetadas como tumor benigno, 43 como cáncer gástrico, 40 como control sano y 17 como control de calidad.

Podemos obtener un resumen sobre la distribución de la concentración de metabolitos en una misma muestra usando la función `summary()` en la matriz de datos:

```
summary(as.data.frame(assays(se)$count[1:10]))
```

```
##     sample_1          sample_2          sample_3          sample_4
##  Min.   :   9.90   Min.   :   16.1   Min.   :    0.2   Min.   :   4.80
##  1st Qu.:  32.42   1st Qu.:  121.7   1st Qu.:   45.6   1st Qu.:  26.25
##  Median :  86.75   Median :  272.8   Median :  136.2   Median :  60.50
##  Mean   : 449.24   Mean   : 1142.9   Mean   :  598.2   Mean   : 225.42
##  3rd Qu.: 226.32   3rd Qu.:  679.7   3rd Qu.:  461.6   3rd Qu.: 138.45
##  Max.   :8029.50   Max.   :16744.8   Max.   :12939.2   Max.   :4562.60
##                    NA's   :2         NA's   :2         NA's   :1
##     sample_5          sample_6          sample_7          sample_8
##  Min.   :    3.9   Min.   :   1.60   Min.   :    9.60   Min.   :   3.70
##  1st Qu.:   48.8   1st Qu.:  38.45   1st Qu.:   49.88   1st Qu.:  15.07
##  Median :  197.8   Median : 124.60   Median :  149.85   Median :  37.10
##  Mean   :  554.7   Mean   : 489.20   Mean   :  638.62   Mean   : 179.95
##  3rd Qu.:  415.4   3rd Qu.: 242.60   3rd Qu.:  353.65   3rd Qu.:  81.78
##  Max.   :12562.3   Max.   :8484.20   Max.   :13681.80   Max.   :2665.00
##  NA's   :1         NA's   :1         NA's   :2
##     sample_9         sample_10
##  Min.   :   0.20   Min.   :  12.60
##  1st Qu.:  39.38   1st Qu.:  34.92
##  Median :  76.20   Median :  99.55
##  Mean   : 448.43   Mean   : 436.12
##  3rd Qu.: 351.23   3rd Qu.: 236.97
##  Max.   :6864.30   Max.   :7915.50
##
```

También podemos obtener un resumen de cada metabolito:

```
summary(t(assays(se)$count[1:10, ]))
```

```
##        M4               M5               M7               M8
##  Min.   :  0.10   Min.   :   1.3   Min.   :  4.60   Min.   :  9.30
##  1st Qu.: 18.77   1st Qu.:  67.0   1st Qu.: 19.65   1st Qu.: 37.17
##  Median : 35.70   Median : 160.3   Median : 40.25   Median : 51.00
##  Mean   : 43.83   Mean   : 231.1   Mean   : 74.12   Mean   : 67.81
##  3rd Qu.: 51.33   3rd Qu.: 253.1   3rd Qu.: 65.55   3rd Qu.: 77.83
##  Max.   :242.50   Max.   :2503.0   Max.   :492.60   Max.   :525.00
##  NA's   :12       NA's   :2        NA's   :4
##       M11              M14              M15              M25
```

```
##  Min.   :   0.3   Min.   :  0.20   Min.   :  7.90   Min.   :  0.40
##  1st Qu.:  47.0   1st Qu.: 30.40   1st Qu.: 28.80   1st Qu.: 12.20
##  Median :  82.9   Median : 50.40   Median : 44.10   Median : 18.25
##  Mean   : 192.5   Mean   : 68.75   Mean   : 57.16   Mean   : 24.23
##  3rd Qu.: 164.7   3rd Qu.: 83.70   3rd Qu.: 71.08   3rd Qu.: 28.50
##  Max.   :2676.3   Max.   :437.60   Max.   :212.30   Max.   :171.80
##  NA's   :7        NA's   :3
##       M26              M31
##  Min.   :  2.00   Min.   :   0.20
##  1st Qu.: 13.80   1st Qu.:  13.60
##  Median : 20.90   Median :  24.10
##  Mean   : 28.31   Mean   :  70.59
##  3rd Qu.: 30.80   3rd Qu.:  61.55
##  Max.   :374.60   Max.   :1336.60
##  NA's   :8        NA's   :1
```

Con algo más de trabajo, podemos obtener cierta información relevante para el estudio como cuales son los metabolitos que presentan mayor concentración media para cada clase. Los resultados se encuentran expuestos en la siguiente tabla, con el nombre del metabolito seguido de su valor de expresión:

Table 1: Tabla de metabolitos con mayor concentración media en cada clase

	1	2	3	4	5	6
GC	M48: 8838.94	M45: 2386.34	M66: 2114.04	M134: 1993.97	M129: 1921.22	M138: 1521
BN	M48: 10394.27	M45: 2803.79	M129: 1578.97	M134: 1251	M66: 1242.15	M137: 1203.01
HE	M48: 11747.39	M45: 4629.9	M129: 1720.52	M66: 1271.77	M137: 1133.86	M134: 687.17
QC	M48: 7809.06	M66: 3114.12	M137: 2742.74	M129: 2412.21	M134: 696.31	M148: 654.97

Observamos que en todas las clases el metabolito que mayor concentración presenta es M48 (Creatinine), seguido de M45 (Citrate) en todos los casos excepto en el control de calidad.

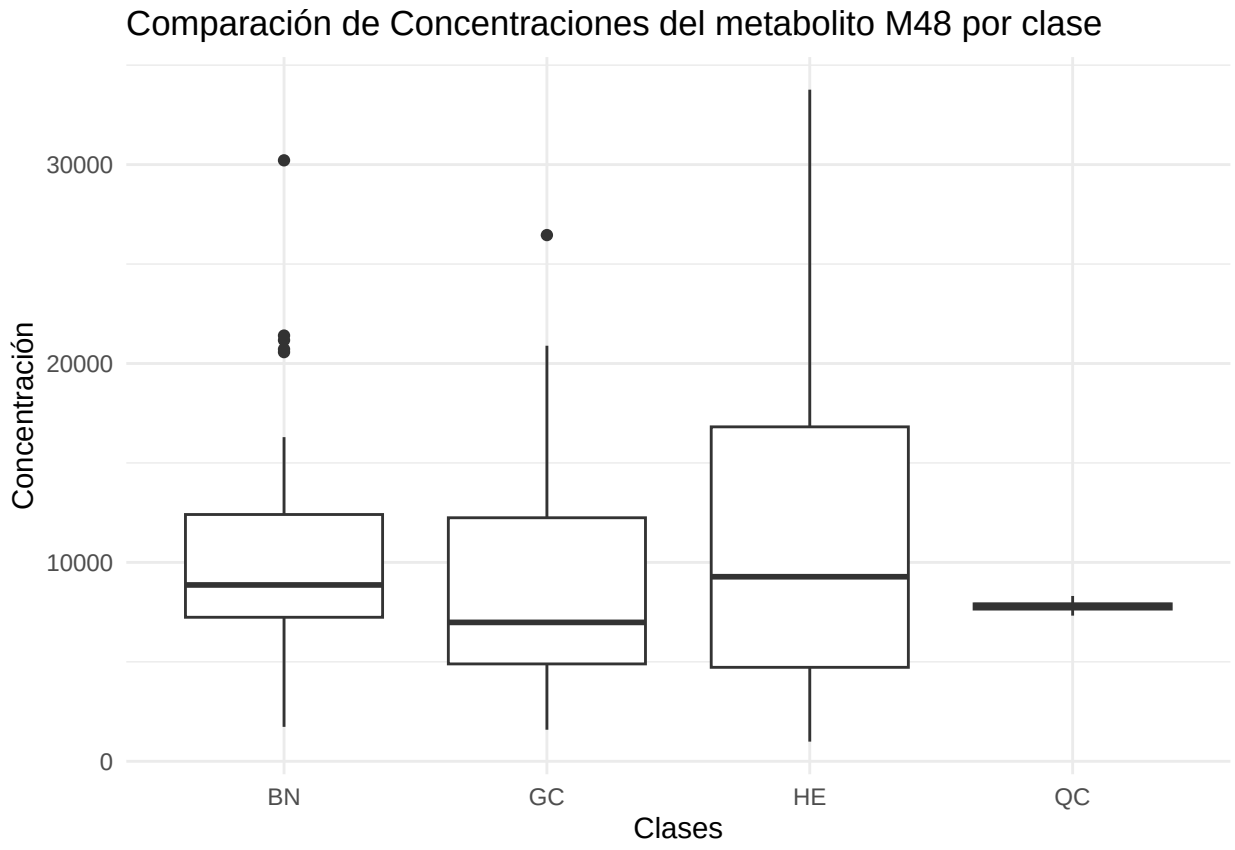
Para reforzar esta información podemos investigar cuales son los metabolitos que mayor diferencia de concentración presentan entre las muestras etiquetadas como cáncer gástrico y las muestras etiquetadas como tumor benigno y control sano, para tratar de encontrar genes implicados en el desarrollo de este tipo de tumor. No escalaremos los datos debido a que se prefiere analizar las diferencias en las concentraciones de los metabolitos en términos absolutos, para una mejor interpretación biológica. Los resultados se exponen en la siguiente tabla:

Table 2: Tabla de metabolitms con mayor diferencia de concentración media entre las clases

	1	2	3	4	5	6
Entre GC y BN	M48: 1555.33	M66: 871.89	M134: 742.98	M45: 417.45	M138: 354.06	M129: 342.25
Entre GC y HE	M48: 2908.45	M45: 2243.56	M134: 1306.81	M138: 1188.07	M66: 842.27	M89: 587.92
Entre BN y HE	M45: 1826.12	M48: 1353.12	M138: 834.01	M134: 563.83	M89: 526.9	M50: 428.91

El metabolito que presenta mayor diferencia en concentración media, entre muestras etiquetadas como cáncer gástrico y tumor benigno, y cancer gástrico y control sano , es M48 en ambos casos, por lo que la concentración de este metabolito podría estar implicada en el desarrollo de este tipo de tumor.

Podemos representar un `Boxplot` para comparar de forma más visual la distribución del metabolito M48 en las diferentes clases.



Como se esperaba la clase QC presenta la menor dispersión, ya que estas muestras actúan como control de calidad manteniendo concentraciones consistentes. Podemos ver que la concentración de M48 en la clase GC es más baja en comparación con BN y esta a su vez más baja respecto a HE, lo que sugiere que la disminución en la concentración de este metabolito podría estar asociado con el cáncer gástrico, siendo esa primera disminución en las muestras BN una respuesta temprana o una alteración metabólica que podría ocurrir en etapas previas a la malignización completa.

Análisis estadísticos

Test de Kruskal-Wallis para el metabolito M48

Para comprobar si las diferencias observadas en la concentraciones de este metabolito entre las clases son estadísticamente significativas podemos realizar un ANOVA, o en su defecto, si no hay normalidad y homogeneidad de varianza en los datos de una misma clase, un test no paramétrico como el Test de Kruskal-Wallis.

```
shapiro.test(datos_limpios_class$M48[datos_limpios_class$Class == "GC"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos_limpios_class$M48[datos_limpios_class$Class == "GC"]
## W = 0.90924, p-value = 0.002414
```

```r
shapiro.test(datos_limpios_class$M48[datos_limpios_class$Class == "HE"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos_limpios_class$M48[datos_limpios_class$Class == "HE"]
## W = 0.90244, p-value = 0.002253
```

Dado que el p-value obtenido para GC y HE es menor a 0,05, asumimos que las concetraciones de los metabolitos en cada grupo no siguen una distribución normal. Optaremos entonces por el Test de Kruskal-Wallis:

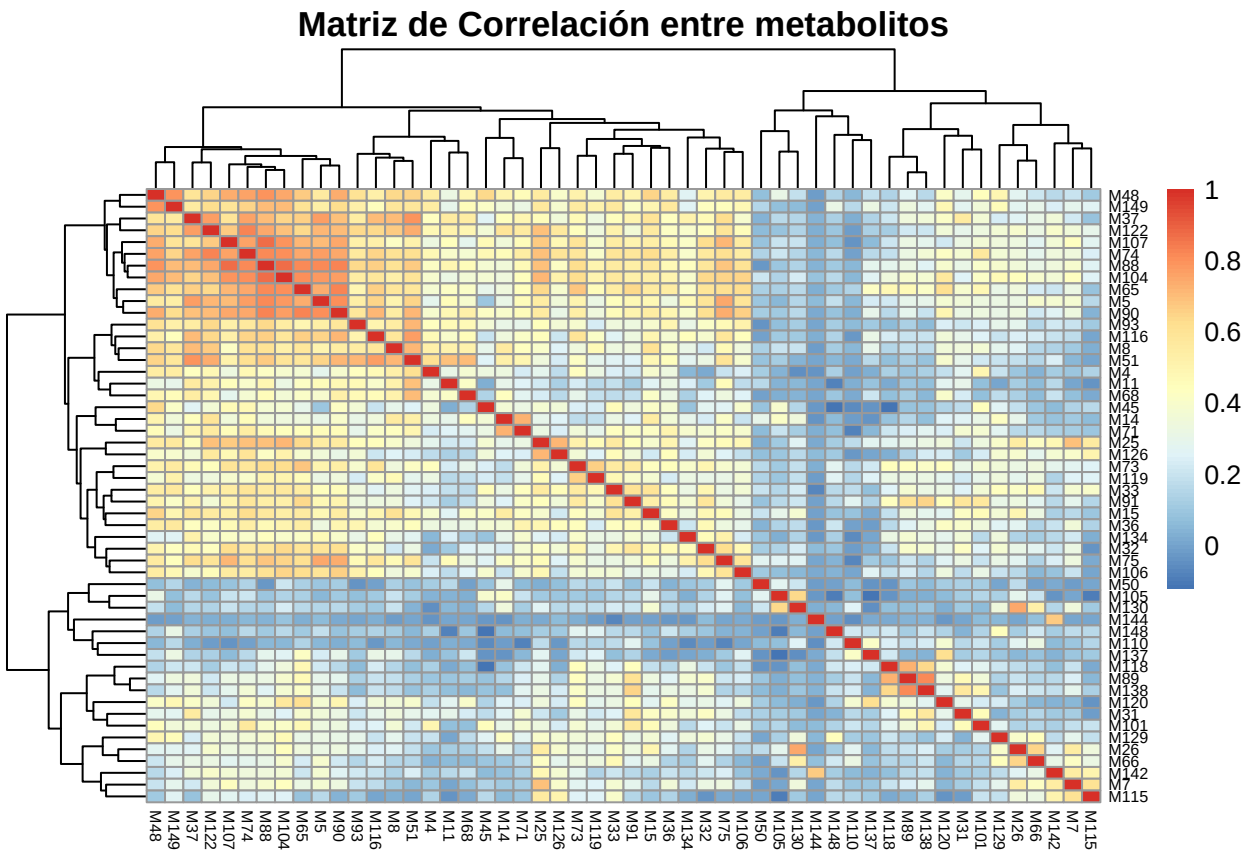
```r
kruskal.test(M48 ~ Class, data = datos_limpios_class)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  M48 by Class
## Kruskal-Wallis chi-squared = 3.7306, df = 3, p-value = 0.2921
```

El p-valor es mayor a 0,05, por lo que no hay evidencias para rechazar la hipótesis nula y asumimos que no hay diferencias estadísticamente significativas en los niveles del metabolito M48 entre las difentes clases.

Matriz de correlación entre metabolitos

Siguiendo con el análisis, vamos a indagar acerca de la correlación de los metabolitos, en especial con el metabolito M48. Para ello, podemos representar una matriz de correlación. En este caso usaremos la librería pheatmap, que nos permiten representar la matriz de una manera intuitiva en forma de mapa de calor:

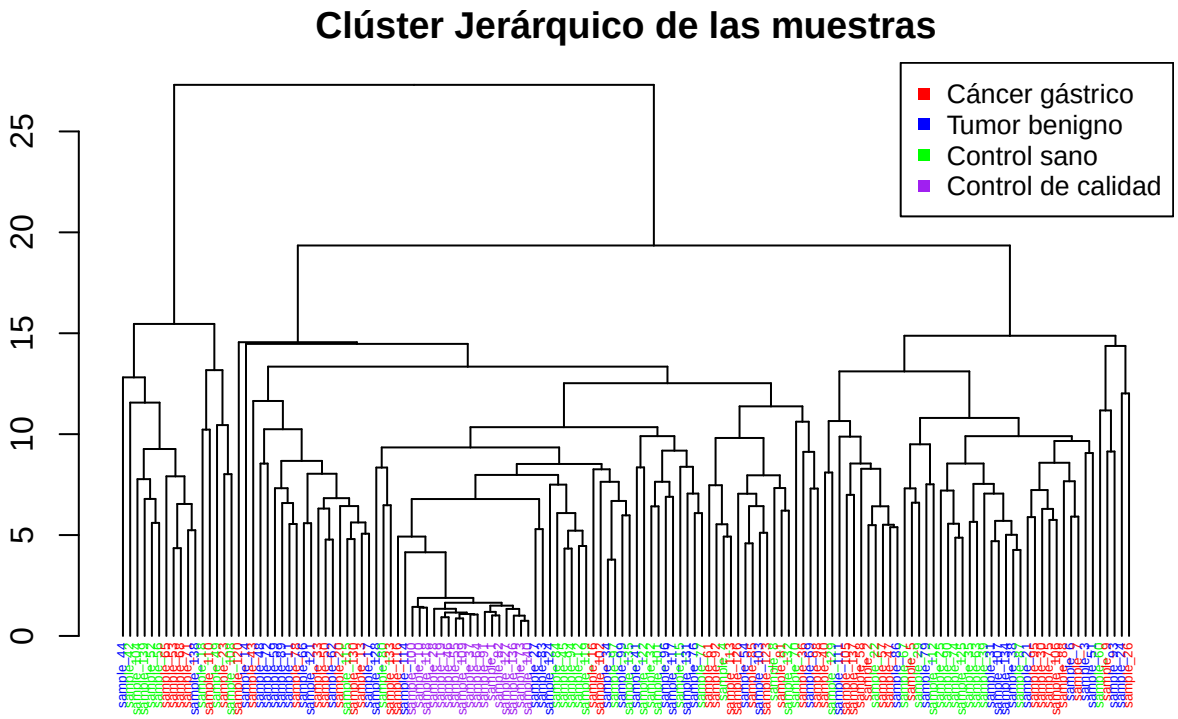


Observamos que los metabolitos que presentan mayor correlación se encuentran en la esquina superior derecha, por lo que seguramente esten involucrados en las mismas rutas metabólicas o en procesos biológicos interrelacionados.Los metabolitos con los que mayor correlación presenta M48 son M149 M107,M74,M88 y M104. Por otro lado, M48 no parece presentar demasiada correlación con los metabolitos que lo seguían en mayor diferencia de concentración entre muestras de cáncer gástrico y controles sanos (M45,M134,M138,M66 y M89).

Cluster Jerárquico Divisivo

A continuación, proceremos a investigar la forma en que las muestras se agrupan en función de la concentración de sus metabolitos y evaluaremos si estos agrupamientos se corresponden con las clases asignadas a cada muestra.

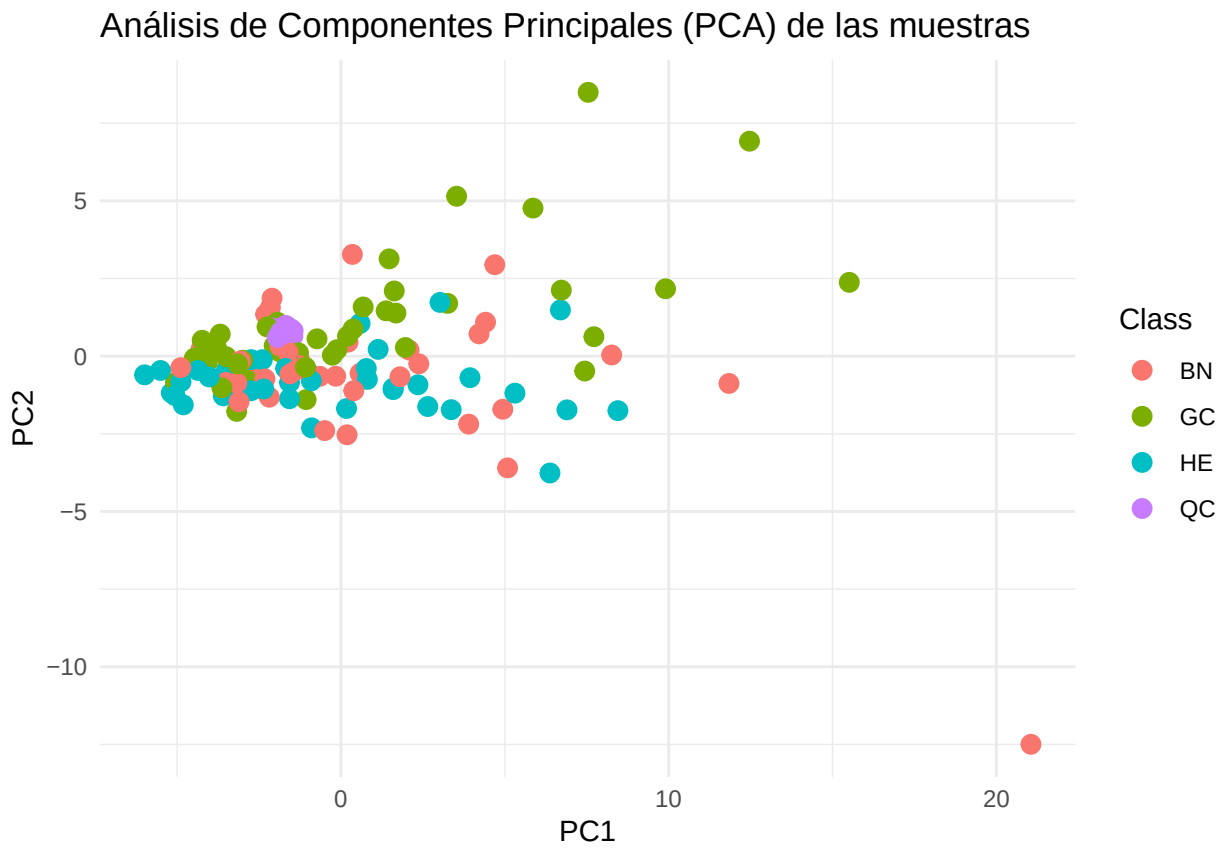
Para este cometido, en primer lugar construiremos clúster jerarquico entre las muestras para comprobar si muestras de una misma clase se agrupan de forma natural en función de sus distancias en el perfil de expresión de metabolitos. Para visualizar este agrupamiento de manera intuitiva, añadiremos color al nombre de cada muestra según la clase a la que pertenece. Antes de representar el dendograma, realizaremos un escalado estándar de los datos para que todas las variables contribuyan de manera equitativa al análisis.



Como era de esperar las muestras que mejor se agrupan son las correspondientes al control de calidad, agrupandose todos los ejemplos en un nodo relativamente alto del dendograma. Las muestras pertenecientes al resto de clases no parecen agruparse tan bien, estando repartidas en diversos subgrupos a lo largo del dendograma, aunque en el caso de los controles sanos, las muestras parecen agruparse en menos subgrupos.

Análisis de Componentes Principales

Finalmente, realizaremos un análisis de componentes principales (PCA) para ver como se distribuyen las muestras pertenecientes a las diferentes clases en función de los dos componentes principales, que son combinaciones lineales de las variables originales que buscan capturar la mayor cantidad de variabilidad posible en los datos. Antes de ello eliminaremos los valores nulos ya que la función `pcrcomp()` no puede manejar dichos valores e indicaremos la opción scale.=TRUE de esta función para hacer un escalado estándar de los datos.



En este gráfico, parece que algunas clases muestran una ligera agrupación (por ejemplo, los puntos verdes GC y azules HE), lo que sugiere que pueden tener patrones distintos. Sin embargo, también hay solapamiento entre las clases, lo que indica que las diferencias no son completamente distinguibles en estas dos dimensiones. En cuanto a los valores de los componentres principales, las muestras de cáncer gástrico parece que tienden a tener un mayor valor de PC2 si las comparamos con los controles sanos principalmente.

Discusión y conclusiones del estudio

En las búsqueda de posibles metabolitos implicados en el desarrollo del cáncer gástrico se encontró que el metabolito que mayor diferencia en concentración media presentaba entre muestras con cáncer gástrico y el resto de clases, era el denominado M48 (Creatinine). Sin embargo, se realizó un test de Kruskal-Walli para comprobar si las diferencias observadas eran estadísticamente significativas y se obtuvo un p-valor superior a 0.05, por lo que no se pueden asumir dichar diferencias. Una solución para obtener mejores resultados podría ser escalar los datos de concentración antes de realizar estos análisis, aunque como ya se mencionó previamente, para este análisis se prefirió trabajar con las concentraciones originales para una mayor veracidad en la interpretación biológica.

En cuanto a los resultados del Clusting Jerárquico y el Análisis de Componentes Principales (PCA), no se encontró un patrón de agrupamiento muy marcado en muestras pertenecientes a la misma clase, excepto para las de control de calidad. Esto puede deberse a que las variables estudiadas no capturan las diferencias biológicas clave entre las clases o que haya mucha variabilidad dentro de las propias clases, por lo que sería interesante tratar de identificar y analizar subtipos dentro de las propias clases.

Reposición de los datos en github

En primer lugar, a traves de las siguientes líneas he creado todo los archivos que se piden:

```
#Exportación del contenedor en formato .rda
save(se, file = "Summarized_Experiment.Rda")
#Exportación de la matriz en formato txt
write.table(assay(se), file = "datos_limpios.txt",
            sep = "\t", row.names = TRUE, col.names = TRUE, quote = FALSE)
#Exportación de los datos originales
write.table(datos, file = "datos_originales.txt",
            sep = "\t", row.names = TRUE, col.names = TRUE, quote = FALSE)
#Exportación de los metadatos de las muestras en formato md
library(knitr)
markdown_table_col <- kable(colData, format = "markdown")
writeLines(markdown_table_col, "metadata_muestras.md")
#Exportación de los metadatos de los metabolitos en formato md
library(knitr)
markdown_table_row <- kable(rowData, format = "markdown")
writeLines(markdown_table_row, "metadata_metabolitos.md")
```

Los pasos que he seguido para vinculor los datos han sido:

1. En primer lugar he creado un repositorio en github: https://github.com/Jorgevg0/Velazquez-Gomez-Jorge-PEC1
2. En Rstudio he abierto las opciones del proyecto y en las opciones de Git/SVN he marcado Git como versión de control de sistemas.
3. He abierto un nuevo terminal e introducido el comando: -git remote add origin https://github.com/Jorgevg0/Velazquez-Gomez-Jorge-PEC1
4. He usado la opcion Commit del menu Git en Rstudio y he marcado los archivos correspondientes.
5. Finalmente he introducido en el terminal el comando: -git push origin master.