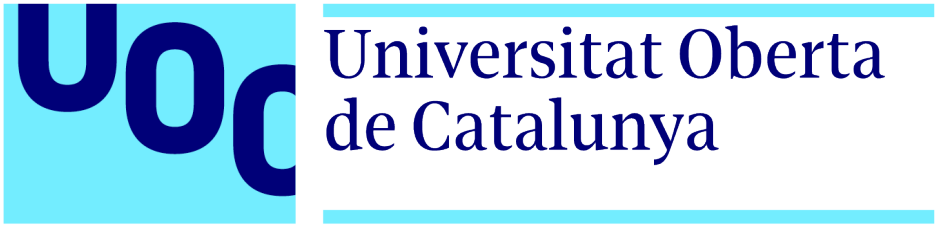


Análisis de datos ómicos (M0-157) PEC 1 - 6/11/2024

Álvaro Rodríguez Sans

Contents

1 Abstract

Esta actividad se corresponde con la prueba de evaluación continua (PEC1) de la asignatura de análisis de datos ómicos.

Esta práctica permitirá la consolidación de los conocimientos adquiridos hasta el momento acerca de las ómicas así como su manejo bioinformático con Bioconductor haciendo uso del objeto **SummarizedExperiment**.

Esta PEC desarrolla la planificación y ejecución del proceso de análisis básico de datos ómicos, mediante el uso de las herramientas y métodos vistos hasta el momento donde se demuestra un conocimiento teórico básico de las tecnologías ómicas y las posible / diversas herramientas informáticas disponibles (Bioconductor, R, etc.). Para ello, se resuelve un ejercicio que consiste en analizar un dataset elegido la azahar de entre los proporcionados, el cual ha sido descargado a RStudio de manera automática con los códigos R generados al efecto.

El dataset seleccionado 'GastricCancer_NMR.xlsx'viene del estudio publicado por Broadhurst (2018), donde se investiga si el cáncer gástrico (GC) tiene un perfil metabolómico urinario distintivo en comparación con enfermedades gástricas benignas (BN) y personas sanas (HE). Con los datos proporcionados se construye el objeto **SummarizedExperiment**. Según el estudio se identificaron 77 metabolitos reproducibles con aplicaron de análisis estadísticos univariantes y multivariantes junto con la generación de un modelo de regresión logística LASSO para la selección de los metabolitos clave (en este caso el "urinario"), Broadhurst (2018).

En nuestro caso hemos desarrollado los puntos referentes a la selección, descarga y análisis exploratorio del dataset.

2 Objetivos

El objetivo de esta PEC es la planificación y ejecución del proceso de análisis básico de datos ómicos, mediante el uso de las herramientas y métodos vistos hasta el momento.

Se busca obtener un conocimiento teórico básico de las tecnologías ómicas y las posibles / diversas herramientas disponibles para trabajar con ellas como la librería Bioconductor y sus contenedores ómicos al efecto (en nuestro caso **SummarizedExperiment**), los gestores de control de versiones como github y las herramientas de exploración de datos dentro del paraguas del lenguaje R, con el que hemos realizado todo el proceso.

3 Materiales y métodos

Los datos con los que hemos tratado en esta PEC son de carácter metabolómico y en nuestro caso son accesibles desde el repositorio proporcionado github, Nutrimetabolomics (n.d.). También pueden ser accedidos desde el estudio publicado por Broadhurst (2018).

Se ha seleccionado un dataset y se ha creado una estructura de datos del tipo **expressionSet** pero en este caso utilizado la clase **SummarizedExperiment**, Morgan (n.d.), la cual contiene una matriz de datos, una tabla con información sobre las covariables y otros aspectos del experimento. Todo ello usando el lenguaje de programación R y git como gestor de control de versiones.

La construcción y exploración del objeto **SummarizedExperiment** se ha llevado a cabo principalmente siguiendo las plantillas de los casos de estudio compartidos en el aula así como la ayuda oficial de la clase **SummarizedExperiment**:

- Analisis_de_datos_omicos-Ejemplo_0-Microarrays, Sanchez-Pla (n.d.a).
- Omics_Data_Analysis-Case_Study_0-Introduction_to_BioC, Sanchez-Pla (n.d.b).
- SummarizedExperiment for Coordinating Experimental Assays, Samples, and Regions of Interest, Morgan (n.d.).

La exploración del objeto generado ha consistido en la obtención de una visión general del mismo en la línea de lo que hemos visto en las actividades vía:

- Análisis univariante vía boxplots y/o histogramas para ver la forma general de los mismos.

- Análisis multivariante vía análisis de componentes Principales (PCA) y agrupamiento jerárquico, y a así ver si los grupos se relacionan entre ellos.

El código R empleado en esta PEC es accesible desde el repositorio github creado al efecto (ver anexo para su acceso). Se han ocultado los códigos de todos los puntos de este informe salvo resultados relevantes mediante la opción knitr::purl(), de esta manera se ha extraído el código R generado del informe ("knitr::purl("ADO-PEC1-Res.Rmd")") y se ha creado el archivo"ADO-PEC1-Res.R" accesible en el repositorio github.

4 Reasultados

4.1 Selección y descarga de los datos para el estudio

Los datos han sido seleccionados de entre los provistos en el repositorio de github, Nutrimetabolomics (n.d.), antes mencionado.

Para la selección se ha accedido al repositorio en remoto mediante un código escrito en R que selecciona de manera aleatoria el dataset a analizar de entre los disponibles haciendo uso de una semilla para asegurar repetibilidad.

Estos son los datasets disponibles para seleccionar:

Dataset	Samples	Features	Description
2018-MetabotypingPaper	39	690	Data used in the paper "Metabotypes of response to bariatric surgery independent of the magnitude of weight loss"
2018-Phosphoproteomics	12	1320	The acompanying dataset has been obtained from a phosphoproteomics experiment that was performed to analyze (3 + 3) PDX models of two different subtypes using Phosphopeptide enriched samples.
2023-CIMCBTutorial	140	149	NMR data from a gastric cancer study used in a metabolomics data analysis tutorial ("Basic Metabolomics Data Analysis Workflow" (https://cimcb.github.io/MetabWorkflowTutorial/Tutorial1.html)
2023-UGrX-4MetaboAnalystTutorial	24	145	Data from MetabolomicsWorkbench (ID ST000002)
2024-fobitools-UseCase_1	45	1541	This dataset is used in the fobitools Bioconductor package, in one its vignettes, [Use Case ST000291] analyzing the data from Metabolomics Workbench Dataset
2024-Cachexia	77	63	Cachexia is a complex metabolic syndrome associated with an underlying illness (such as cancer) and characterized by loss of muscle with or without loss of fat mass (Evans et al., 2008). A total of 77 urine samples were collected being 47 of them patients with cachexia, and 30 control patients (from the "specmine.datasets" R package)

Este es el dataset seleccionado:

```
# A tibble: 1 x 1
  Dataset
  <chr>
1 2023-CIMCBTutorial
```

Localizamos el archivo en repositorio, el cual se encuentra en la ruta:

- Datasets/2023-CIMCBTutorial/GastricCancer_NMR.xlsx

Lo descargamos, con R, a nuestra área de trabajo del proyecto en RStudio para su posterior lectura y análisis.

Verificamos que hemos descargado correctamente el fichero "GastricCancer_NMR.xlsx" mostrando las primeras filas.

```
[1] "Archivo descargado correctamente."
```

```
# 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>, ...
```

4.1.1 Daset descargado - Estudio de referencia

Antes de proseguir con el análisis se ha indagado de donde provienen los datos y cual es el objeto original del estudio.

Estos datos que hemos seleccionado viene del estudio publicado por Broadhurst (2018), donde se investiga si el cáncer gástrico (GC) tiene un perfil metabolómico urinario distintivo en comparación con enfermedades gástricas benignas (BN) y personas sanas (HE). Para ello, analizaron muestras de orina de 43 pacientes con GC, 40 con BN y 40 personas sanas HE utilizando espectroscopía de resonancia magnética nuclear de protón (1H-NMR).

Identificaron 77 metabolitos reproducibles y aplicaron análisis estadísticos univariantes y multivariantes. Mediante un modelo de regresión logística LASSO, seleccionaron tres metabolitos clave para diferenciar GC de HE: 2-hidroxibutirato, 3-indoxilsulfato y alanina. Este modelo mostró un alto poder predictivo, con un área bajo la curva ROC de 0,95, Broadhurst (2018).

Los resultados sugieren que el perfil metabolómico urinario podría ser útil para el diagnóstico temprano del cáncer gástrico, dado su potencial para diferenciar entre GC y controles sanos, Broadhurst (2018).

4.2 Generación del contenedor

El tipo del contenedor es **SummarizedExperiment**, Morgan (n.d.), que contendrá los datos y metadatos (información acerca del dataset, las filas y las columnas) del dataset proporcionado.

Tal y como se explica en rdrr.iO y SummarizedExperiment for Coordinating Experimental Assays, Samples, and Regions of Interest, Martin (2020), está clase esta diseñada convenientemente para almacenar datos numéricos y otros tipos de datos derivados de un experimento de secuenciación. Con una estructura rectangular, tipo matriz, ofrece anotaciones adicionales en las filas y columnas, y con posibilidad de gestionar varios ensayos simultáneamente. Los datos del ensayo han de estar en una matriz.

Los aspectos mas importantes a la hora de entender y trabajar con esta clase son:

- Los datos se acceden mediante la función **assays**, la cual devuelve un objeto **SimpleList**. Cada elemento de la lista debe ser a su vez una matriz con las mismas que las dimensiones del SummarizedExperiment en el que están almacenados. Los nombres de filas y columnas de cada matriz deben ser

NULL o coincidir con los del SummarizedExperiment durante la construcción. Es conveniente que los elementos de la **SimpleList** de ensayos tengan nombre.

- Las filas de un objeto SummarizedExperiment representan las características de interés y la información sobre estas características se almacena en un objeto DataFrame, que es accesible mediante la función **rowData**. Este DataFrame debe tener tantas filas como filas tenga el objeto SummarizedExperiment, y cada fila proporciona información sobre la característica en la fila correspondiente del objeto SummarizedExperiment, donde las columnas del DataFrame representan diferentes atributos de las características de interés.
- Cada columna de un objeto SummarizedExperiment representa una muestra y la información sobre las misma ea almacenada en un DataFrame, que es accesible mediante la función **colData**. Este DataFrame debe tener tantas filas como columnas tenga el objeto SummarizedExperiment, y cada fila proporciona información sobre la muestra en la columna correspondiente del objeto, donde las columnas del DataFrame representan diferentes atributos de las muestras.
- En un objeto SummarizedExperiment también podemos incluir información sobre el experimento en general. Esta información se almacena como un objeto list, y es accesible mediante la función **metadata**.

Estos son los datos que obtenemos de la lectura del fichero excel (hojas Data y Peak):

Data:

```
# 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>, ...
```

```
[1] "tbl_df"     "tbl"        "data.frame"
```

Tal y como se explica en cimcb, CIMCB (n.d.):

- Las columnas M1…M149 describen las concentraciones de metabolitos.
- La columna "SampleType" indica si la muestra era un control de calidad combinado o una muestra de estudio.
- La clase de columna "Class" indica el resultado clínico observado para ese individuo: GC = cáncer gástrico, BN = tumor benigno, HE = control sano.

Peak:

```
# A tibble: 6 x 5
    Idx Name  Label                    Perc_missing QC_RSD
  <dbl> <chr> <chr>                           <dbl>  <dbl>
1     1 M1    1_3-Dimethylurate              11.4    32.2
2     2 M2    1_6-Anhydro- -D-glucose         0.714  31.2
3     3 M3    1_7-Dimethylxanthine            5      35.0
```

```
4     4 M4    1-Methylnicotinamide           8.57   12.8
5     5 M5    2-Aminoadipate                 1.43    9.37
6     6 M6    2-Aminobutyrate                5      47.0
```

```
[1] "tbl_df"     "tbl"        "data.frame"
```

- "Name" es el nombre de la columna correspondiente a este metabolito.
- "Label" proporciona un nombre único para el metabolito (o un identificador uNNN).
- "Perc_missing" indica qué porcentaje de muestras no contienen una medición para este metabolito (datos faltantes).
- "QC_RSD" es una puntuación de calidad que representa la variación en las mediciones de este metabolito en todas las muestras.

4.2.1 Preparación de assay (matriz de datos de expresión)

Para crear el objeto, es necesario que los datos estén en formato de matriz y DataFrame con lo que tenemos que convertir "data" matriz. Por el contrario "peak" ya esta en el formato adecuado.

Extraemos los valores de las mediciones (M1, M2, etc.) de "data" y utilizamos "SampleID" como nombres de filas.

Mostramos la matriz resultante (solo las 6 primeras filas y columnas):

```
            M1       M2    M3   M4    M5    M6
sample_1  90.1   491.6 202.9 35.0 164.2  19.7
sample_2  43.0   525.7 130.2   NA 694.5 114.5
sample_3 214.3 10703.2 104.7 46.8 483.4 152.3
sample_4  31.6    59.7  86.4 14.0  88.6  10.3
sample_5  81.9   258.7 315.1  8.7 243.2  18.4
sample_6 196.9   128.2 862.5 18.7 200.1   4.7
```

4.2.2 Preparacion de rowData y colData

Obtenemos los siguientes datos:

- colData: Usa de peak »» "Name", "Label", "Perc_missing" y "QC_RSD".
- rowData: Usa de data »» "SampleID", "SampleType" y "Class".

Mostramos los primeros elementos de cada elemento:

```
# A tibble: 6 x 3
  SampleID SampleType Class
  <chr>    <chr>      <chr>
1 sample_1 QC         QC
2 sample_2 Sample     GC
3 sample_3 Sample     BN
4 sample_4 Sample     HE
5 sample_5 Sample     GC
6 sample_6 Sample     BN
```

```
# A tibble: 6 x 4
  Name  Label                   Perc_missing QC_RSD
  <chr> <chr>                          <dbl>  <dbl>
1 M1    1_3-Dimethylurate             11.4    32.2
2 M2    1_6-Anhydro- -D-glucose        0.714  31.2
3 M3    1_7-Dimethylxanthine           5      35.0
4 M4    1-Methylnicotinamide           8.57   12.8
5 M5    2-Aminoadipate                 1.43    9.37
6 M6    2-Aminobutyrate                5      47.0
```

4.2.3 Creacion de SummarizedExperiment

Ahora generamos el objeto siguiendo las pautas indicada en SummarizedExperiment for Coordinating Experimental Assays, Samples, and Regions of Interest, Morgan (n.d.).

Objeto generado:

```
class: SummarizedExperiment
dim: 140 149
metadata(0):
assays(1): counts
rownames(140): sample_1 sample_2 ... sample_139 sample_140
rowData names(3): SampleID SampleType Class
colnames(149): M1 M2 ... M148 M149
colData names(4): Name Label Perc_missing QC_RSD
```

4.2.4 Inserción de metadatos a SummarizedExperiment

Los metadatos a nivel de característica (metabolito) y sample ya los hemos añadido en "rowData" y "colData". Estos describen propiedades específicas de cada metabolito, como nombre, etiqueta, porcentaje de datos faltantes, etc.

Los mostramos:

```
DataFrame with 6 rows and 4 columns
          Name                  Label Perc_missing    QC_RSD
   <character>            <character>    <numeric> <numeric>
M1          M1      1_3-Dimethylurate    11.428571  32.20800
M2          M2 1_6-Anhydro- -D-gluc..     0.714286  31.17803
M3          M3   1_7-Dimethylxanthine     5.000000  34.99060
M4          M4   1-Methylnicotinamide     8.571429  12.80420
M5          M5         2-Aminoadipate     1.428571   9.37266
M6          M6        2-Aminobutyrate     5.000000  46.97715
```

```
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
```

Los metadatos a nivel de experimento describen información general sobre el experimento, como el experimentador, laboratorio, título, etc. Estos datos se añadirán a metadata su origen puede encontrarse en, CIMCB (n.d.), secciones 2.1 y 2.2, y ademas en, Broadhurst (2018).

Estos son, a nivel de experimento, los metadatos del objeto generado (ver código para mas detalles):

Campo	Descripción
Project_ID	PR000699
Project_DOI	doi: 10.21228/M8B10B
Project_Title	1H-NMR urinary metabolomic profiling for diagnosis of gastric cancer
Institute	University of Alberta
Autor	Broadhurst, David
Address	270 Joondalup Drive, Joondalup, WA 6027, AUSTRALIA
Email	d.broadhurst@ecu.edu.au
Project_Summary	Background: Metabolomics has shown promise in gastric cancer (GC) detection. This research sought to identify whether GC has a unique urinary metabolomic profile compared with benign gastric disease (BN) and healthy (HE) patients. Methods: Urine from 43 GC, 40 BN, and 40 matched HE patients was analysed using 1H nuclear magnetic resonance (1H-NMR) spectroscopy, generating 77 reproducible metabolites (QC-RSD <25%). Univariate and multivariate (MVA) statistics were employed. A parsimonious biomarker profile of GC vs HE was investigated using LASSO regularised logistic regression (LASSO-LR). Model performance was assessed using Receiver Operating Characteristic (ROC) curves. Results: GC displayed a clear discriminatory biomarker profile; the BN profile overlapped with GC and HE. LASSO-LR identified three discriminatory metabolites: 2-hydroxyisobutyrate, 3-indoxylsulfate, and alanine, which produced a discriminatory model with an area under the ROC of 0.95. Conclusions: GC patients have a distinct urinary metabolite profile. This study shows clinical potential for metabolic profiling for early GC diagnosis.
Data_Sample	Las columnas M1...M149 describen las concentraciones de metabolitos.
Data_SampleType	La columna SampleType indica si la muestra era un control de calidad combinado o una muestra de estudio.
Data_Class	La clase de columna Class indica el resultado clínico observado para ese individuo: GC = cáncer gástrico, BN = tumor benigno, HE = control sano.
Peak_Name	Name es el nombre de la columna correspondiente a este metabolito.
Peak_Label	Label proporciona un nombre único para el metabolito (o un identificador uNNN).
Peak_Perc_missing	Perc_missing indica qué porcentaje de muestras no contienen una medición para este metabolito (datos faltantes).
Peak_QC_RSD	QC_RSD es una puntuación de calidad que representa la variación en las mediciones de este metabolito en todas las muestras.

4.3 Creacion de del fichero .Rda

También procedemos a guardar el objeto generado como un fichero binario .Rda que se llama "Gastric-Cancer_NMR.Rda" y procedemos a su carga otra vez para realizar el siguiente punto.

```
[1] "El archivo se ha creado correctamente."
```

Procedemos a cárgalo y mostrarlo:

```
class: SummarizedExperiment
dim: 6 149
metadata(15): Project_ID Project_DOI ... Peak_Perc_missing Peak_QC_RSD
assays(1): counts
rownames(6): sample_1 sample_2 ... sample_5 sample_6
rowData names(3): SampleID SampleType Class
colnames(149): M1 M2 ... M148 M149
colData names(4): Name Label Perc_missing QC_RSD
```

```
Formal class 'SummarizedExperiment' [package "SummarizedExperiment"] with 5 slots
```

```
..@ colData        :Formal class 'DFrame' [package "S4Vectors"] with 6 slots
.. .. ..@ rownames       : chr [1:149] "M1" "M2" "M3" "M4" ...
.. .. ..@ nrows          : int 149
.. .. ..@ elementType    : chr "ANY"
.. .. ..@ elementMetadata: NULL
.. .. ..@ metadata       : list()
.. .. ..@ listData       :List of 4
.. .. .. ..$ Name        : chr [1:149] "M1" "M2" "M3" "M4" ...
.. .. .. ..$ Label       : chr [1:149] "1_3-Dimethylurate" "1_6-Anhydro- -D-glucose" "1_7-Dimethylxant
.. .. .. ..$ 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 ...
..@ assays         :Formal class 'SimpleAssays' [package "SummarizedExperiment"] with 1 slot
.. .. ..@ data:Formal class 'SimpleList' [package "S4Vectors"] with 4 slots
.. .. .. .. ..@ listData       :List of 1
.. .. .. .. .. ..$ counts: num [1:140, 1:149] 90.1 43 214.3 31.6 81.9 ...
.. .. .. .. .. .. ..- attr(*, "dimnames")=List of 2
.. .. .. .. .. .. .. ..$ : chr [1:140] "sample_1" "sample_2" "sample_3" "sample_4" ...
.. .. .. .. .. .. .. ..$ : chr [1:149] "M1" "M2" "M3" "M4" ...
.. .. .. .. ..@ elementType    : chr "ANY"
.. .. .. .. ..@ elementMetadata: NULL
.. .. .. .. ..@ metadata       : list()
..@ NAMES          : chr [1:140] "sample_1" "sample_2" "sample_3" "sample_4" ...
..@ elementMetadata:Formal class 'DFrame' [package "S4Vectors"] with 6 slots
.. .. ..@ rownames       : NULL
.. .. ..@ nrows          : int 140
.. .. ..@ elementType    : chr "ANY"
.. .. ..@ elementMetadata: NULL
.. .. ..@ metadata       : list()
.. .. ..@ listData       :List of 3
.. .. .. ..$ SampleID  : chr [1:140] "sample_1" "sample_2" "sample_3" "sample_4" ...
.. .. .. ..$ SampleType: chr [1:140] "QC" "Sample" "Sample" "Sample" ...
.. .. .. ..$ Class     : chr [1:140] "QC" "GC" "BN" "HE" ...
..@ metadata       :List of 15
.. ..$ Project_ID       : chr "PR000699"
.. ..$ Project_DOI      : chr "doi: 10.21228/M8B10B"
.. ..$ Project_Title    : chr "1H-NMR urinary metabolomic profiling for diagnosis of gastric cancer"
.. ..$ Institute        : chr "University of Alberta"
.. ..$ Autor            : chr "Broadhurst,\tDavid"
.. ..$ Address          : chr "270 Joondalup Drive, Joondalup, WA 6027, AUSTRALIA"
.. ..$ Email            : chr "d.broadhurst@ecu.edu.au"
.. ..$ Project_Summary  : chr "Background: Metabolomics has shown promise in gastric cancer (GC) dete
.. ..$ Data_Sample      : chr "Las columnas M1...M149 describen las concentraciones de metabolitos."
.. ..$ Data_SampleType  : chr "La columna SampleType indica si la muestra era un control de calidad c
.. ..$ Data_Class       : chr "La clase de columna Class indica el resultado clínico observado para e
.. ..$ Peak_Name        : chr "Name es el nombre de la columna correspondiente a este metabolito."
.. ..$ Peak_Label       : chr "Label proporciona un nombre único para el metabolito (o un identificad
.. ..$ Peak_Perc_missing: chr "Perc_missing indica qué porcentaje de muestras no contienen una medici
.. ..$ Peak_QC_RSD      : chr "QC_RSD es una puntuación de calidad que representa la variación en las
```

4.4 Exploración del dataset

Realizamos un análisis exploratorio de los datos. Primero vamos a ver cual es el tamaño de nuestro set de datos:

```
[1] 140 149
```

```
[1] 20860
```

De este total de 20860 datos, observamos que tenemos datos vacíos (NA):

```
[1] 1069
```

Y datos disponibles:

```
[1] 19791
```

Ahora bien, tenemos una dimensión de 140 x 149 y atendiendo a lo expuesto en, CIMCB (n.d.), sección 3, parare que podemos reducir la dimensionalidad de assay con datos cuyo QC-RSD sea menor del 20% y el percMiss sea de menos del 10%.

De esta manera vemos que nos quedamos en lugar de 149 samples con:

```
[1] "Numero de 'peaks' de interes: 52"
```

```
# A tibble: 52 x 1
   Name
   <chr>
 1 M4
 2 M5
 3 M7
 4 M8
 5 M11
 6 M14
 7 M15
 8 M25
 9 M26
10 M31
# i 42 more rows
```

Este paso nos sirve para ver una de las propiedades de este tipo de objetos, que es que cuando cambia algo (numero de filas o columnas) este cambio se replica a las demás partes del objeto sin necesidad de hacer el cambio manualmente.

Vemos como las dimensiones han cambiado en assay, colData y rowData manteniendo su sincronía.

```
[1] 140  52
```

```
           M4    M5    M7    M8    M11  M14
sample_1 35.0 164.2  41.0  46.5   61.7 35.3
sample_2   NA 694.5  37.9 125.7  490.6   NA
sample_3 46.8 483.4 110.1  85.1 2441.2 29.3
sample_4 14.0  88.6 170.3  23.9  140.7 62.9
sample_5  8.7 243.2 349.4  61.1   48.7 77.8
sample_6 18.7 200.1  37.3 243.7  103.7 52.3
```

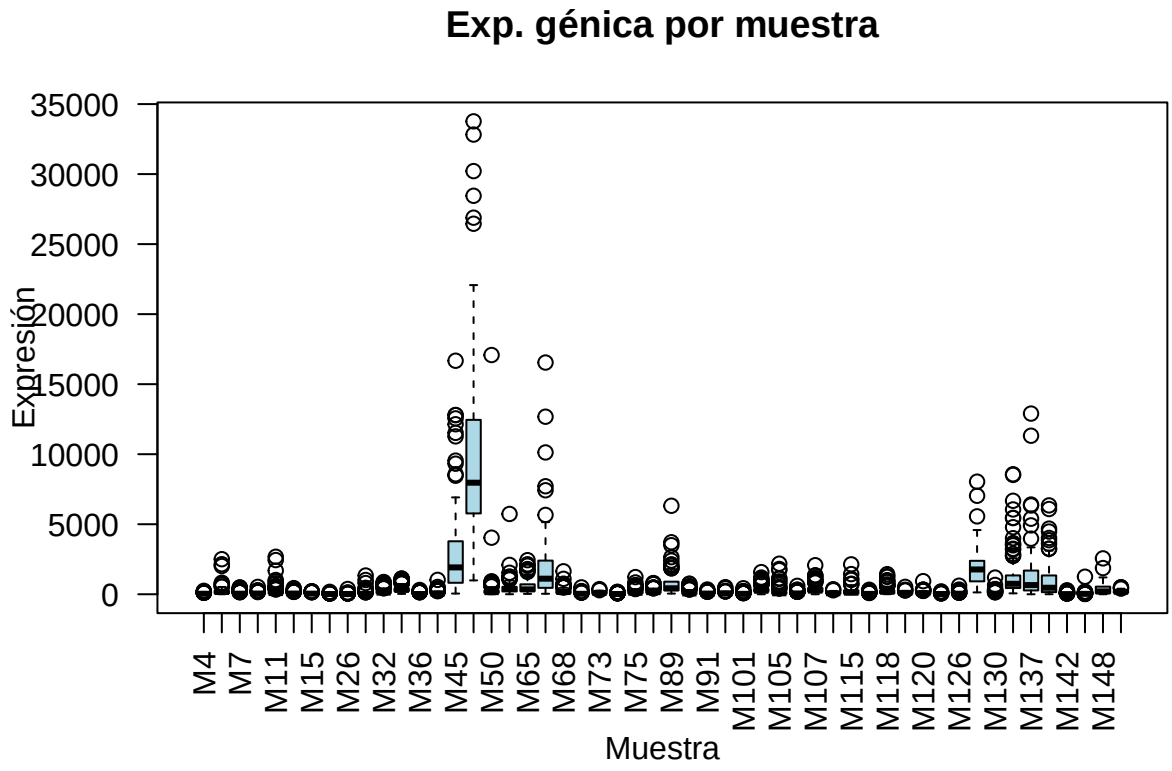
```
[1] 140   3
```

```
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
```

```
[1] 52  4
```

```
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
```

Ahora procedemos a realizar las visualizaciones oportunas.



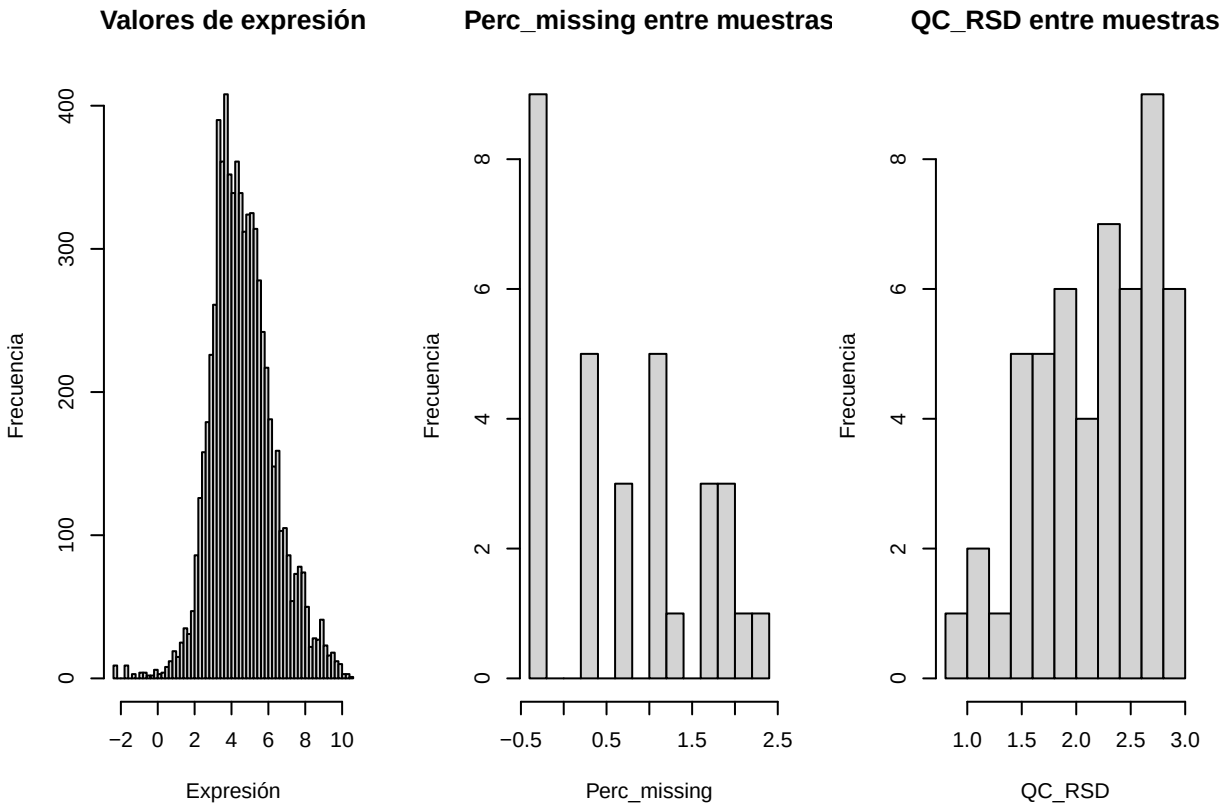
Importante

Los datos al ser asimétricos en varios de los "samples", sugiere que puede tener sentido trabajar con los mismos datos en escala logarítmica.

4.4.1 Histograma

Observamos la distribución de frecuencias "assay" y "colData" para ver cómo se distribuyen los valores de las variables de colData (Perc_missing y QC_RSD) y la matriz (assay).

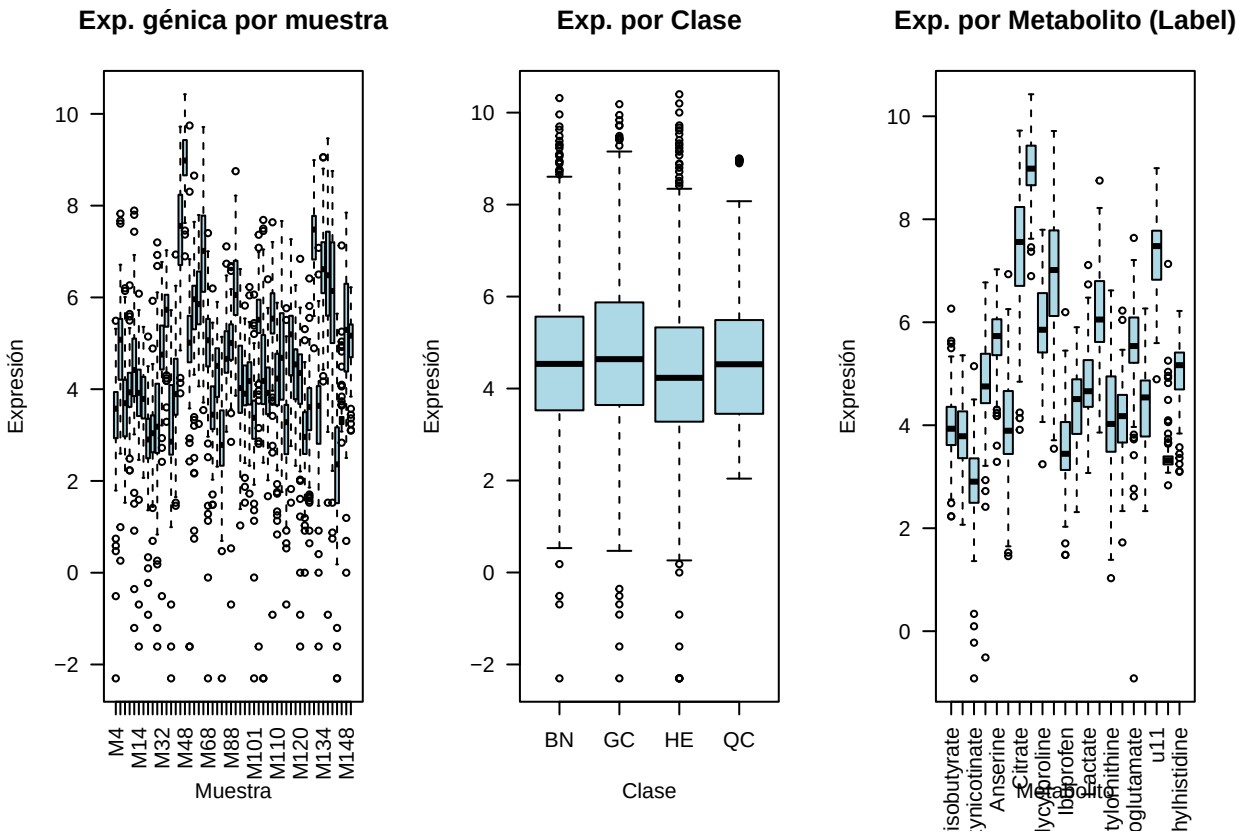
En en le caso del assay vemos como su distribución es simétrica, en colData_Perc_missing parace que tenemos un sesgo negativo (a la izquierda) donde la cola larga está en el lado izquierdo indicandonos que la mayoría de los valores son altos, pero hay algunos más bajos y finalmente colData_QC_RSD que tiene un sesgo positivo (a la derecha) donde la cola larga (lado derecho) indica que la mayoría de los valores bajos son pocos y los altos son más abundantes.



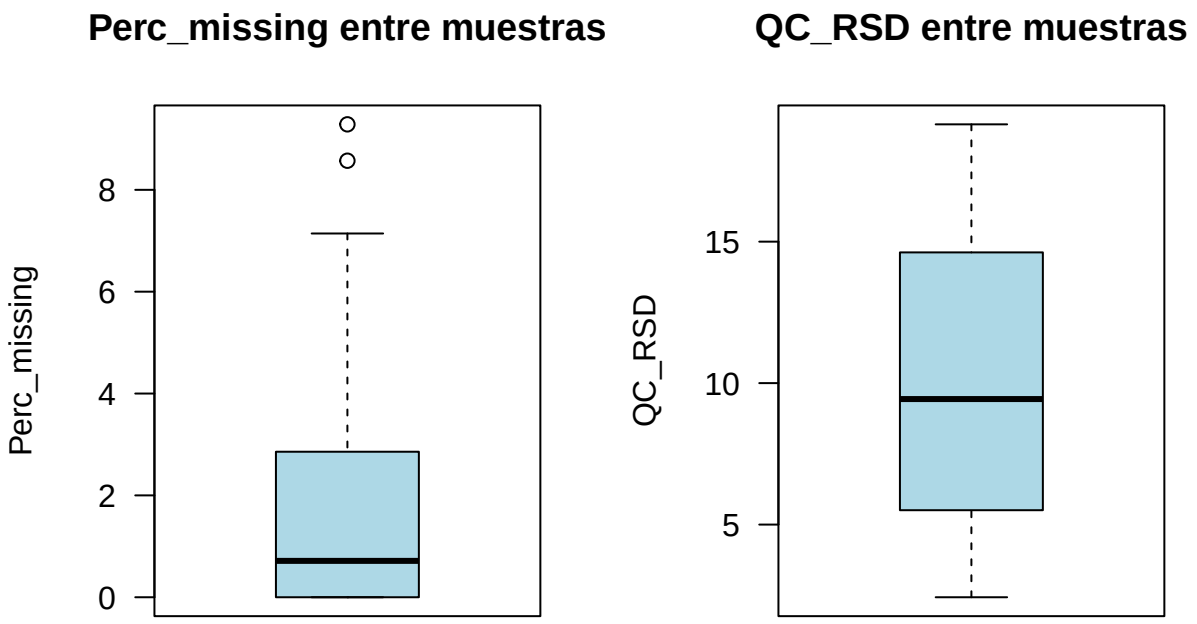
4.4.2 Boxplot

Un boxplot (diagrama de caja y bigotes) es una herramienta gráfica que permite visualizar la distribución, simetría, dispersión y posibles valores atípicos en un conjunto de datos permitiendo obtener una visión rápida y detallada de la distribución de los datos.

En nuestro caso si las cajas son muy amplias, la expresión génica indica que tiene mucha variabilidad y vemos que este no es el caso. También vemos que los bigotes superiores no son excesivamente largos, el caso contrario nos indicaría que algunos genes tienen niveles de expresión más altos de lo habitual. Como punto final si que vemos como en alguna expresiones los puntos fuera de los bigotes (outliers) podrían ser genes con expresión particularmente alta o baja y merecen su revisión para ver si estos son errores o resultados reales importantes.



En el caso de QC_RSD y Perc_missing vemos como sus representación entra dentro de los valores esperados después de la limpieza realizada.

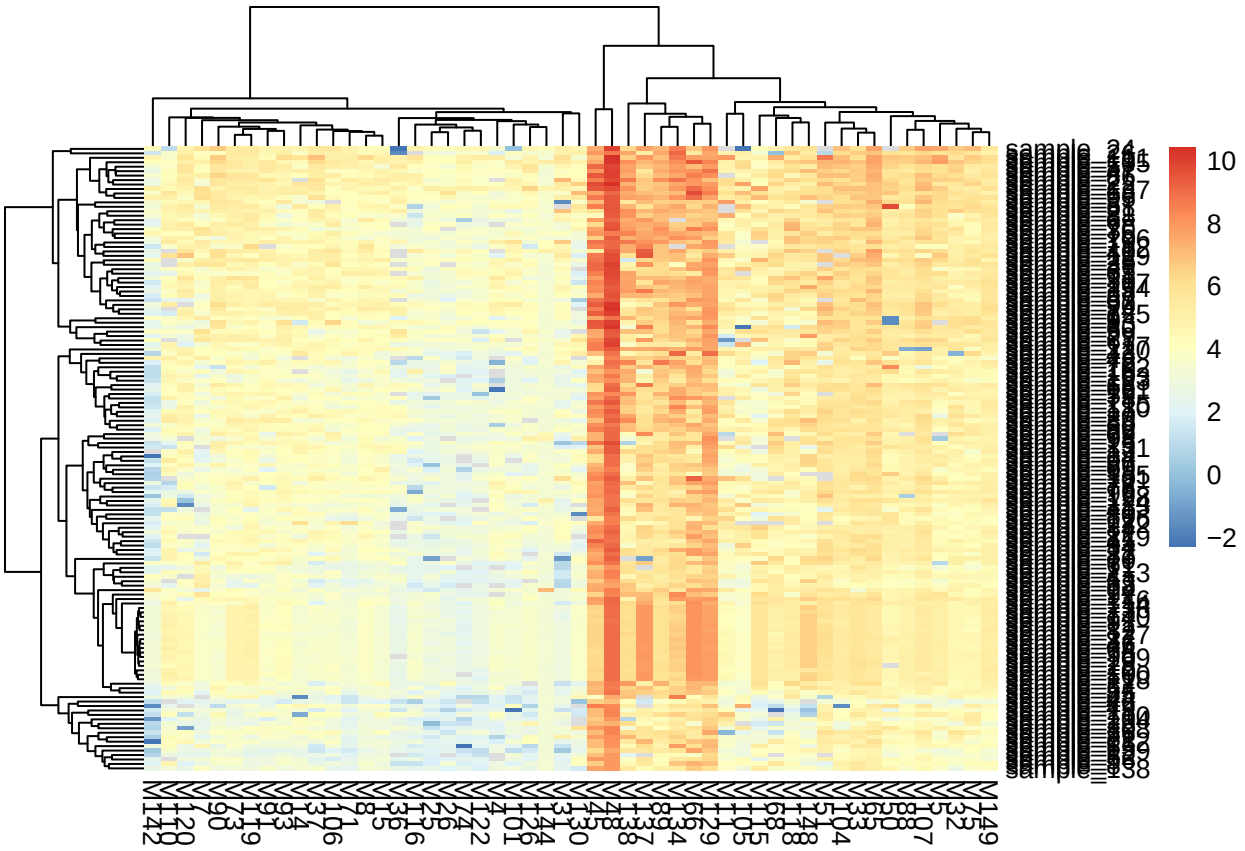


4.4.3 Heatmap

Nos ofrece una visualización intuitiva para identificar rápidamente patrones, agrupaciones y posibles valores atípicos en los datos de expresión o metabolitos, facilitando la comparación y el análisis exploratorio de las relaciones en el conjunto de datos.

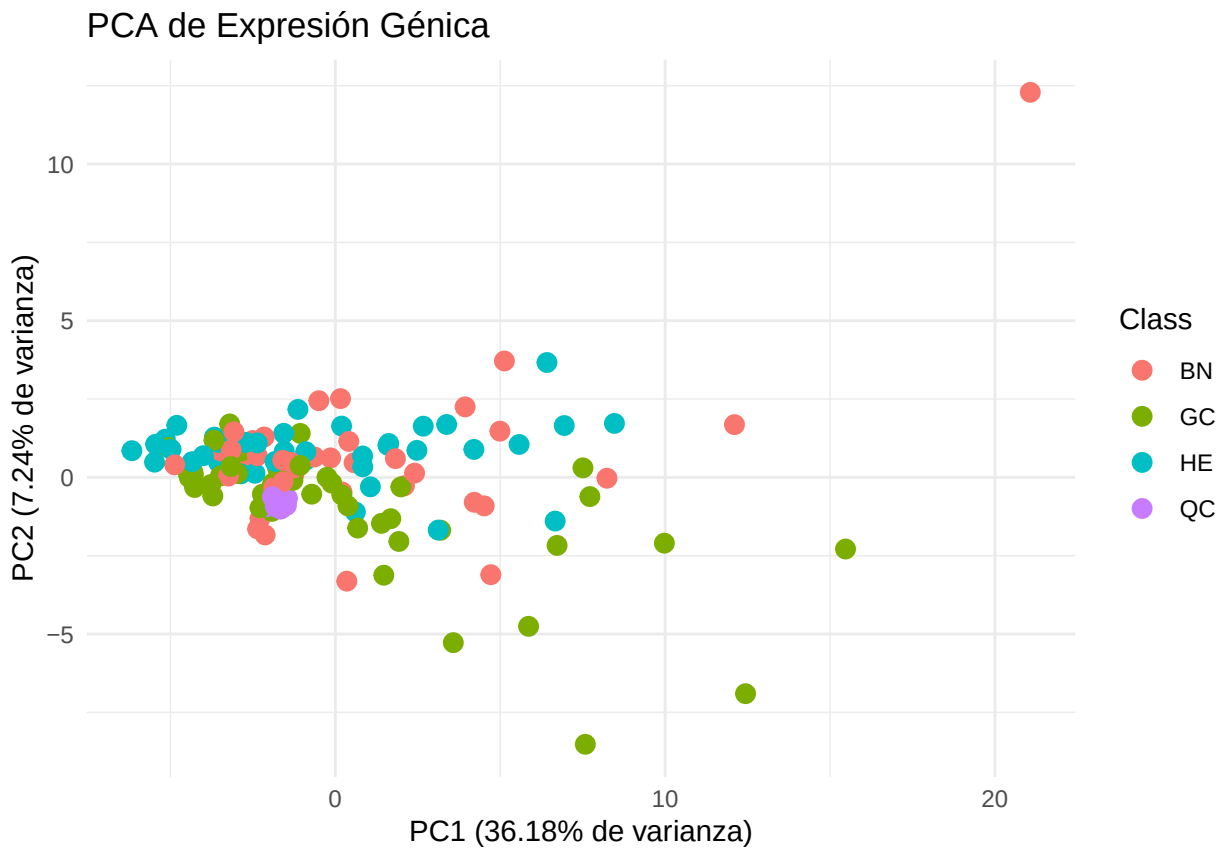
- Las filas (genes o metabolitos) cercanas nos indican que suelen compartir un patrón de expresión similar (e.j. participación en funciones biológicas relacionadas).
- Las columnas (muestras) agrupadas tienen perfiles de expresión parecidos, y nos podría indicar que pertenecen al mismo grupo de condición o estado.

También nos vale para la posible identificación de grupos y biomarcadores donde los genes o muestras que se destacan por niveles de expresión fuera del patrón general pueden ser candidatos interesantes, ya sea como biomarcadores (si representan patrones específicos de una condición) o como genes/metabolitos de interés para análisis posteriores. En nuestro caso vemos como **M48 - Creatinine** es un elemento a tener muy presente.



4.4.4 PCA

El Análisis de Componentes Principales (PCA) es una técnica que permite reducir la dimensionalidad de un conjunto de datos, visualizando la variabilidad de los datos en un espacio de menor dimensión donde los componentes principales (PC1 y PC2) representan la mayor parte de la variabilidad en los datos.




```
[1] "PC1 (36.18% de la varianza) - PC2 (7.24% de la varianza)"
```

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     4.3373 1.94056 1.76174 1.59331 1.46954 1.40850 1.33331
                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     1.19708 1.15586 1.09986 1.07587 1.04140 0.96581 0.93136
                          PC15   PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     0.88148 0.8654 0.83567 0.82192 0.79194 0.76286 0.75150
                         PC22    PC23    PC24    PC25    PC26    PC27    PC28
Standard deviation     0.7101 0.66567 0.61167 0.59192 0.58668 0.57373 0.54250
                          PC29    PC30   PC31    PC32    PC33    PC34    PC35
Standard deviation     0.51799 0.49164 0.4675 0.43915 0.42087 0.39553 0.39311
                         PC36    PC37    PC38    PC39    PC40    PC41    PC42
Standard deviation     0.3602 0.34701 0.33350 0.32850 0.31370 0.28008 0.26616
                          PC43    PC44    PC45    PC46    PC47    PC48    PC49
Standard deviation     0.25433 0.23414 0.23180 0.22107 0.20111 0.18857 0.17778
                          PC50    PC51    PC52
Standard deviation     0.16868 0.14594 0.13508
 [ reached getOption("max.print") -- omitted 2 rows ]
```

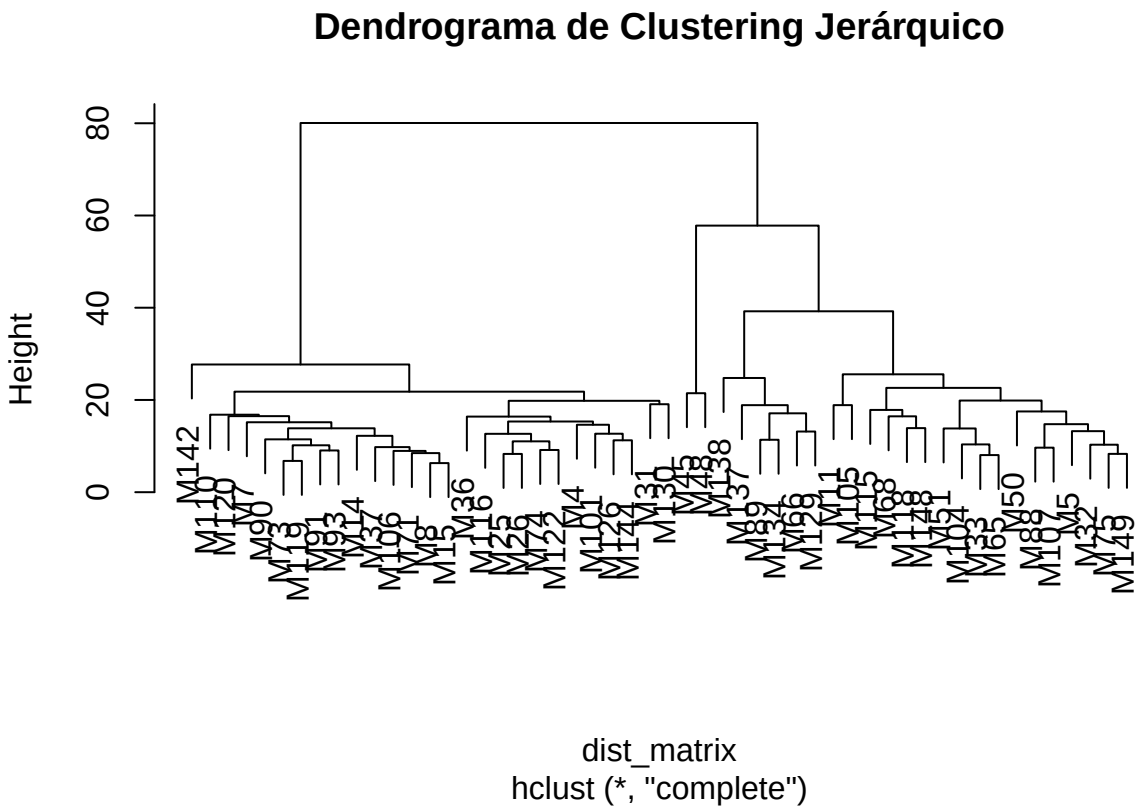
4.4.5 Agrupaciones - Cluster

El dendrograma nos ofrece un análisis de clúster jerárquico con una representación visual de la relación entre las muestras basadas en sus perfiles de expresión génica.

Cosas que podemos revisar en el denograma son:

- Fusión de clústeres: Las ramas del dendrograma indican cómo se agrupan las muestras y estas se fusionan en una "altura menor" si son más similares entre sí en términos de expresión génica (cuanto más baja sea la altura en el eje vertical donde se fusionan, mayor es la similitud).
- Distancia: La longitud de las ramas nos indica la distancia entre los clústeres. Ramas más cortas significan que las muestras son más similares, mientras que ramas más largas indican que las muestras son más diferentes.

En nuestro caso observamos la formación de grupos o clústeres en el dendrograma. Cada grupo puede representar un conjunto de muestras que comparten características similares en sus perfiles de expresión.



5 Discusión

El análisis exploratorio realizado, tras imputar, limpiar y tomar logaritmos sobre los datos finales, ha puesto de manifiesto que existen mas dos fuentes de variación distintas ya que las PC1 y PC2 explican no mas del 44% de la variabilidad.

Ademas en las visualizaciones no se muestran grupos diferenciadores que pudieran a ayudar a identificar uno o unos metabolitos mas relacionados con el cáncer gástrico (GC).

En lo que tiene que ver con la agrupación en el PCA, la clase "QC" esta perfectamente agrupada en un pequeño espacio y la clase "GC" es la que se puede agrupar en un área separada y amplia pero que también se mezcla con otras clases en otras áreas.

En cuanto a la relación entre muestras las distancias entre los puntos en el gráfico de PCA nos indica la similitud o disimilitud entre estas, donde muestras cercanas tienen perfiles de expresión génica similares y las que están más separadas son diferentes. Como vemos "QC" es la única similar.

Vemos outliers si que están relacionados con las clase "GC" y "BN".

Como conclusión entendemos que habría que seguir con el estudio y aplicar mas tecnicas, como las descritas en Broadhurst (2018), para poder llegar a las conclusiones descritas en ese estudio.

6 Apendice

6.1 URL repositorio github

El repositorio de esta PEC con todos los entregables solicitados se en:

https://github.com/arodriguezsans/Rodriguez-Sans-Alvaro-PEC1

En este repositorio se encuentran los siguientes ficheros de interes:

- ADO-PEC1-Res.pdf (informe pdf de la pec)
- ADO-PEC1-Res.R (código R)
- ADO-PEC1-Res.Rmd (informe rmarkdown de la pec)
- ADOreferences-Res.bib (referencia bibliográficas usadas)
- GastricCancer_NMR.Rda (fichero .Rda solicitado)
- GastricCancer_NMR.xlsx (fichero excel descargado del repositorio de referencia)

Bibliografia

Broadhurst, David. 2018. "1H-NMR Urinary Metabolomic Profiling for Diagnosis of Gastric Cancer." *Metabolomics Workbench: NIH Data Repository*. University of Alberta. https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Project&ProjectID=PR000699.

CIMCB. n.d. "Tutorial 1: Basic Metabolomics Data Analysis Workflow." *Cimcb*. cimcb. https://cimcb.github.io/MetabWorkflowTutorial/Tutorial1.html.

Martin, Obenchain, Morgan; Valerie. 2020. "Summarizedexperiment-Class: Summarizedexperiment Objects in Summarizedexperiment: Summarizedexperiment Container." *R Package Documentation*. https://rdrr.io/bioc/SummarizedExperiment/man/SummarizedExperiment-class.html.

Morgan, Valerie, Martin; Obenchain. n.d. "SummarizedExperiment for Coordinating Experimental Assays, Samples, and Regions of Interest." *Bioconductor*. Bioconductor. https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html.

Nutrimetabolomics. n.d. "Nutrimetabolomics/Metabodata: A Repository with a Few Public Metabolomics Datasets Borrowed from Different Public Open Sources." *GitHub*. https://github.com/nutrimetabolomics/metaboData/.

Sanchez-Pla, Alex. n.d.a. "Aspteaching/Analisis_de_datos_omicos-Ejemplo_0-Microarrays: Exploración de Datos de Micrtoarrays USANDO Funciones Básicas de r." *GitHub*. UB-UOC. https://github.com/ASPteaching/Analisis_de_datos_omicos-Ejemplo_0-Microarrays.

———. n.d.b. "Aspteaching/OMICS_DATA_ANALYSIS-Case_study_0-Introduction_to_BioC." *GitHub*. UB-UOC. https://github.com/ASPteaching/Omics_Data_Analysis-Case_Study_0-Introduction_to_BioC.