Análisis de Datos Ómicos - PEC 1

Beatriz Jiménez Guijarro

3 de noviembre, 2024

Contents

1. *Abstract*

Este estudio analiza datos de expresión metabolómica en muestras intestinales pre y post trasplante, con el objetivo de identificar patrones de variabilidad entre las muestras. A partir de datos de expresión metabolómica provenientes de un repositorio público, se han aplicado herramientas bioinformáticas en R para estructurar y analizar los datos. La integración de los datos en un objeto `SummarizedExperiment` ha permitido una manipulación estructurada y el análisis exploratorio de la expresión de los metabolitos.

Los resultados del análisis multivariante, incluyendo análisis de componentes principales (PCA) y agrupación jerárquica, sugieren que existen diferencias sutiles entre las muestras, aunque no se observa una clara separación biológica entre los grupos pre y post trasplante en la mayoría de los componentes principales (PC), lo cual indica que los perfiles metabólicos entre ambos grupos son relativamente similares. Sin embargo, la agrupación jerárquica sugiere que algunos metabolitos pueden contribuir de forma específica a la diferenciación entre grupos. Las implicaciones de estos hallazgos en la identificación de posibles biomarcadores se exploran en el estudio.

2. Objetivos del estudio

El objetivo principal de este estudio es investigar los cambios en el perfil metabolómico intestinal en individuos sometidos a un trasplante intestinal, comparando muestras pre y post trasplante. Se busca identificar patrones de expresión y variabilidad entre los metabolitos que puedan proporcionar información sobre los cambios metabólicos asociados al trasplante, y así mejorar la comprensión de las diferencias metabólicas y su relación con la adaptación del injerto.

Además, se pretende evaluar la utilidad de herramientas bioinformáticas para el análisis metabolómico en este contexto, especialmente mediante el uso de análisis exploratorio de datos y técnicas multivariantes como el análisis de componentes principales (PCA) y la agrupación jerárquica.

3. Materiales y Métodos

En este apartado se describe el conjunto de datos utilizado para el análisis, las herramientas bioinformáticas empleadas y el procedimiento general seguido para procesar y analizar los datos de expresión metabolómica asociados a los intestinos humanos. Este análisis tiene como objetivo identificar patrones de agrupamiento y variabilidad que puedan aportar información sobre las diferencias metabólicas entre muestras de individuos antes y después de un trasplante intestinal.

3.1. Origen y selección de los datos

El conjunto de datos utilizado en este estudio se obtuvo del repositorio público https://github.com/nutrimetabolomics/metaboData/, que proporciona acceso a diversos conjuntos de datos de metabolómica. En particular, el dataset utilizado está relacionado con los intestinos humanos (https://github.com/nutrimetabolomics/metaboData/tree/main/Datasets/2023-UGrX-4MetaboAnalystTutorial), más concretamente con individuos que se han sometido a un trasplante intestinal, y de los que se han recogido muestras intestinales, antes y después del trasplante, para tomar medidas de expresión de diferentes metabolitos asociados.

Cargaremos el dataset con los datos de metabolitos, sus metadatos y la información general del dataset desde archivos de texto.

El dataset contiene valores de expresión de diferentes metabolitos en muestras intestinales de individuos antes y después de un trasplante intestinal. Cada fila representa un metabolito, mientras que las columnas corresponden a muestras individuales. También se incluye una fila que describe el grupo de cada muestra (antes, *Before*, o después, *After*, del trasplante) y las muestras están etiquetadas inicialmente con las letras A y B, según al grupo al que pertenezcan. Los datos de las muestras están almacenados en formato .csv (archivo "ST000002_AN000002_clean.csv")

La información general del dataset la obtenemos de un documento markdown (archivo "ST000002_AN000002_dataset_info.m y nos proporciona una breve documentación explicando el origen de los datos, su organización, la autoría, etc.

Por último, los metadatos de los metabolitos analizados la obtenemos de otro documento de texto (archivo "ST000002_AN000002_metadata.txt") donde cada fila representa un metabolito diferente y las columnas son diferentes carácterísticas propias de estos metabolitos.

3.2. Herramientas y paquetes utilizados

El análisis se ha llevado a cabo utilizando el lenguaje de programación R, y se ha apoyado en herramientas de Bioconductor para facilitar el procesamiento y análisis de los datos ómicos. A continuación se enumeran los paquetes específicos empleados:

- **SummarizedExperiment**: Este paquete permite estructurar los datos y metadatos en un contenedor unificado, optimizado para análisis ómicos. Facilita el manejo de los datos de expresión y permite conservar las relaciones entre las variables y muestras.

- **hist, plot, density, boxplot**: Estas funciones de R se emplean para generar gráficos descriptivos de los datos, como histogramas, gráficos de densidad y diagramas de cajas (boxplots). Estos gráficos permiten observar la distribución de los datos y detectar patrones o valores atípicos en las muestras de forma visual.
- **prcom**: Esta función de R se utiliza para realizar el análisis de componentes principales (PCA), una técnica de reducción de dimensionalidad que facilita la identificación de patrones y agrupaciones entre las muestras al transformar las variables originales en nuevas componentes principales.
- **pairs**: Esta función de R se emplea para visualizar las relaciones entre las primeras componentes principales generadas por el PCA, permitiendo explorar posibles agrupaciones o correlaciones entre las muestras a través de gráficos de pares de componentes.
- **hclust y dist**: Estas funciones de R se utilizan para el análisis de agrupación jerárquica (clustering). `dist` calcula la matriz de distancias entre muestras, y `hclust` construye el dendrograma, mostrando relaciones de similitud y posibles agrupaciones entre las muestras basadas en las medidas de expresión de metabolitos.

3.3. Procedimiento general de análisis

El análisis general se desarrollará en varias etapas. Una vez cargados los datos y los metadatos del dataset relacionado con el intestino humano pre/post trasplante, se deben organizar en un objeto `SummarizedExperiment`, una estructura específica de Bioconductor y una extensión de `ExpressionSet`, que permite almacenar de manera conjunta las mediciones ómicas y sus metadatos. Esto facilitará el acceso y la manipulación de los datos durante el análisis.

Posteriormente, se realizará un análisis exploratorio inicial utilizando estadísticas univariantes para examinar la variabilidad y distribución de los datos.

A continuación, se aplicará un análisis de componentes principales (PCA) para reducir la dimensionalidad y facilitar la identificación de patrones y posibles agrupaciones entre las muestras.

Finalmente, se empleará una agrupación jerárquica para confirmar y visualizar cualquier posible agrupación natural en los datos, ofreciendo una perspectiva adicional sobre la relación entre las muestras en función de su perfil metabolómico.

4. Resultados

En esta sección vamos a presentar los resultados obtenidos del análisis exploratorio del dataset de metabolómica seleccionado (muestras intestinales pre/post transplante), enfocado en la identificación de patrones y características clave de los datos de expresión de metabolitos. Estudiaremos, en primer lugar, la estructura de los datos del dataset y, a continuación, realizaremos el análisis exploratorio de los datos, mediante el cuál, a través de una serie de técnicas estadísticas y visualizaciones, se busca una comprensión general de la variabilidad, distribución y relaciones entre las muestras y los metabolitos en el dataset.

4.1. Estructura de los datos

En primer lugar, se ha creado un objeto `SummarizedExperiment`, que facilita la integración de los datos de expresión y los metadatos. Este objeto es un contenedor de tipo matriz donde las filas representan características de interés (por ejemplo, genes, transcripciones, exones, etc.) y las columnas representan muestras. También contiene la matriz de expresión de los datos, los metadatos de las filas y las columnas, y la información general del estudio.

Cargamos primero el paquete necesario y después, convertimos los datos de estudio en una matriz numérica y creamos un `data.frame` para los metadatos de las muestras. Con esta información, creamos un objeto `SummarizedExperiment` que contiene los datos de los metabolitos, los metadatos de las muestras y los

metabolitos, y la información del experimento en general, facilitando su análisis y manipulación en un solo objeto estructurado. Veamos el resultado del objeto `SummarizedExperiment` creado:

```
## class: SummarizedExperiment
## dim: 142 12
## metadata(1): dataset_info
## assays(1): counts
## rownames(142): 1-monoolein 1-monostearin ... xanthine xylose
## rowData names(8): moverz_quant ri ... other_id other_id_type
## colnames(12): A_684508 A_684512 ... B_684499 B_684503
## colData names(1): Groups
```

Podemos ver un resumen de las dimensiones, filas, columnas y metadatos del objeto `SummarizedExperiment` creado. Este objeto contiene datos de 142 metabolitos en 12 muestras, junto con metadatos específicos sobre cada metabolito y muestra.

Comprobemos los datos de la clase.

```
## [1] "SummarizedExperiment"
## attr(,"package")
## [1] "SummarizedExperiment"
```

Veamos las dimensiones del objeto, es decir, el número de filas (metabolitos) y columnas (muestras) en el conjunto de datos.

```
## [1] 142  12
```

Examinemos la matriz que contiene los valores de los metabolitos para cada muestra. Dado que la cantidad de datos es muy elevada, vamos a mostrar sólo una cabecera (`head`).

```
##                         A_684508 A_684512 A_684516 A_684520 A_684524 A_684528
## 1-monoolein                 6047     2902     1452     3428     2985    16334
## 1-monostearin               9771     6521     1302     2781     5789     4338
## 2-hydroxybutanoic acid     13238    29774     4134     4419    13334     2115
## 2-hydroxyglutaric acid      7160    11501     3202    17238    20376     1109
## 2-ketoisocaproic acid        812     2011      738     2550      871      628
## 2-monopalmitin              1511      622      883      796      623     5716
##                         B_684483 B_684487 B_684491 B_684495 B_684499 B_684503
## 1-monoolein               244142     6968     1928    19228     3029    23277
## 1-monostearin              16848    10206     9398     1013     4190    11114
## 2-hydroxybutanoic acid     11587    65635    32433     1823     4429    30427
## 2-hydroxyglutaric acid      8276    12402    20964    25913     2709    70972
## 2-ketoisocaproic acid       2096     3472    10669      432     1055     1005
## 2-monopalmitin              3405     3196     1457     1416     1275    14445
```

Vamos a comprobar ahora los metadatos de los metabolitos (filas) del dataset. Se mostrará la información sobre cada metabolito, como sus nombres, categorías y demás características propias.

```
## DataFrame with 142 rows and 8 columns
##                         moverz_quant        ri     ri_type pubchem_id inchi_key
##                            <integer> <integer> <character>  <integer> <logical>
## 1-monoolein                      129    952993       Fiehn    5283468        NA
## 1-monostearin                    399    959625       Fiehn     107036        NA
## 2-hydroxybutanoic acid           131    258175       Fiehn      11266        NA
## 2-hydroxyglutaric acid           129    506359       Fiehn         43        NA
## 2-ketoisocaproic acid            200    310629       Fiehn         70        NA
## ...                              ...       ...         ...        ...       ...
## uric acid                        441    731185       Fiehn       1175        NA
## uridine                          258    856953       Fiehn       6029        NA
```

```
## valine                          144    313224       Fiehn       6287        NA
## xanthine                        353    702391       Fiehn       1188        NA
## xylose                          103    542808       Fiehn     135191        NA
##                           kegg_id  other_id other_id_type
##                       <character> <integer>   <character>
## 1-monoolein                           213963       BinBase
## 1-monostearin              D01947    202835       BinBase
## 2-hydroxybutanoic acid     C05984    199800       BinBase
## 2-hydroxyglutaric acid     C02630    214409       BinBase
## 2-ketoisocaproic acid      C00233    213388       BinBase
## ...                           ...       ...           ...
## uric acid                  C00366    221495       BinBase
## uridine                    C00299    213127       BinBase
## valine                     C00183    199605       BinBase
## xanthine                   C00385    203224       BinBase
## xylose                     C00181    200500       BinBase
```

A continuación, también comprobaremos los metadatos pero esta vez de las muestras (columnas) del dataset. Se mostrarán los nombres y la información sobre los grupos o categorías de cada muestra.

```
## DataFrame with 12 rows and 1 column
##               Groups
##          <character>
## A_684508       After
## A_684512       After
## A_684516       After
## A_684520       After
## A_684524       After
## ...              ...
## B_684487      Before
## B_684491      Before
## B_684495      Before
## B_684499      Before
## B_684503      Before
```

En este caso, podemos ver que las muestras se dividen en dos grupos, *After* y *Before*, que indican si las muestras se obtivieron antes (*Before*) o después (*After*) de un transplante intestinal. También podemos ver como los nombres de las muestras también contienen esta información, pues aquellas que pertenecen al grupo *After* comienzan con la letra A y aquellas que pertenecen al grupo *Before* comienzan con la letra B.

Con la función `metadata()` podemos ver los metadatos generales del dataset, es decir, la información general del conjunto de datos, que se cargó previamente, sin embargo, como comprobamos a continuación, al provenir esta información de un documento de texto (.md, en este caso) su lectura es complicada.

```
## $dataset_info
## [1] "#METABOLOMICS WORKBENCH ofiehn_20130123_9589761_mwtab.txt DATATRACK_ID:34 STUDY_ID:ST000002 ANA
```

Por ello, vamos a generar los metadatos del dataset en formato tabla para poder leer más adecuadamente la información.

Clave	Valor
#METABOLOMICS WORKBENCH	
ofiehn_20130123_9589761_mwtab.txt	
DATATRACK_ID:34 STUDY_ID:ST000002	
ANALYSIS_ID:AN000002 PROJECT_ID:PR000002	
PR:PROJECT_TITLE	Intestinal Samples II pre/post transplantation

Clave	Valor
PR:PROJECT_TYPE	Human intestinal samples
PR:PROJECT_SUMMARY	Intestinal Samples II pre/post transplantation
PR:INSTITUTE	University of California, Davis
PR:DEPARTMENT	Davis Genome Center
PR:LABORATORY	Fiehn
PR:LAST_NAME	Fiehn
PR:FIRST_NAME	Oliver
PR:ADDRESS	451 E. Health Sci. Drive, Davis, California 95616, USA
PR:EMAIL	ofiehn@ucdavis.edu
PR:PHONE	-
ST:STUDY_TITLE	Intestinal Samples II pre/post transplantation
ST:STUDY_TYPE	MS analysis
ST:STUDY_SUMMARY	Intestinal Samples II pre/post transplantation
ST:INSTITUTE	University of California, Davis
ST:DEPARTMENT	Davis Genome Center
ST:LABORATORY	Fiehn
ST:LAST_NAME	Hartman
ST:FIRST_NAME	Amber
ST:ADDRESS	451 E. Health Sci. Drive, Davis, California 95616, USA
ST:EMAIL	-
ST:PHONE	-
ST:NUM_GROUPS	2
ST:TOTAL_SUBJECTS	12
SU:SUBJECT_TYPE	Human
SU:SUBJECT_SPECIES	Homo sapiens
SU:TAXONOMY_ID	9606
SU:SPECIES_GROUP	Human
#SUBJECT_SAMPLE_FACTORS	SUBJECT(optional)[tab]SAMPLE[tab]FACTORS(NAME:VAI pairs separated by |)[tab]Additional sample data
SUBJECT_SAMPLE_FACTORS	- LabF_684508 Transplantation:After transplantation
SUBJECT_SAMPLE_FACTORS	- LabF_684512 Transplantation:After transplantation
SUBJECT_SAMPLE_FACTORS	- LabF_684516 Transplantation:After transplantation
SUBJECT_SAMPLE_FACTORS	- LabF_684520 Transplantation:After transplantation
SUBJECT_SAMPLE_FACTORS	- LabF_684524 Transplantation:After transplantation
SUBJECT_SAMPLE_FACTORS	- LabF_684528 Transplantation:After transplantation
SUBJECT_SAMPLE_FACTORS	- LabF_684483 Transplantation:Before transplantation
SUBJECT_SAMPLE_FACTORS	- LabF_684487 Transplantation:Before transplantation
SUBJECT_SAMPLE_FACTORS	- LabF_684491 Transplantation:Before transplantation
SUBJECT_SAMPLE_FACTORS	- LabF_684495 Transplantation:Before transplantation

Clave	Valor
SUBJECT_SAMPLE_FACTORS	- LabF_684499 Transplantation:Before transplantation
SUBJECT_SAMPLE_FACTORS	- LabF_684503 Transplantation:Before transplantation
CO:COLLECTION_SUMMARY	-
CO:SAMPLE_TYPE	Tissue
TR:TREATMENT_SUMMARY	-
TR:TREATMENT_PROTOCOL_COMMENTS	Before transplantation | After transplanation
SP:SAMPLEPREP_SUMMARY	-
SP:EXTRACTION_METHOD	Extraction Proteomics 2004, 4, 78-83; Splitratio splitless 25 purge
CH:CHROMATOGRAPHY_TYPE	GC
CH:INSTRUMENT_NAME	Agilent 6890N
CH:COLUMN_NAME	-
AN:ANALYSIS_TYPE	MS
MS:MS_COMMENTS	-
MS:INSTRUMENT_NAME	Leco Pegasus III GC TOF
MS:INSTRUMENT_TYPE	GC-TOF
MS:MS_TYPE	EI
MS:ION_MODE	POSITIVE
MS_METABOLITE_DATA:UNITS	Peak height

4.1. Análisis exploratorio de los datos

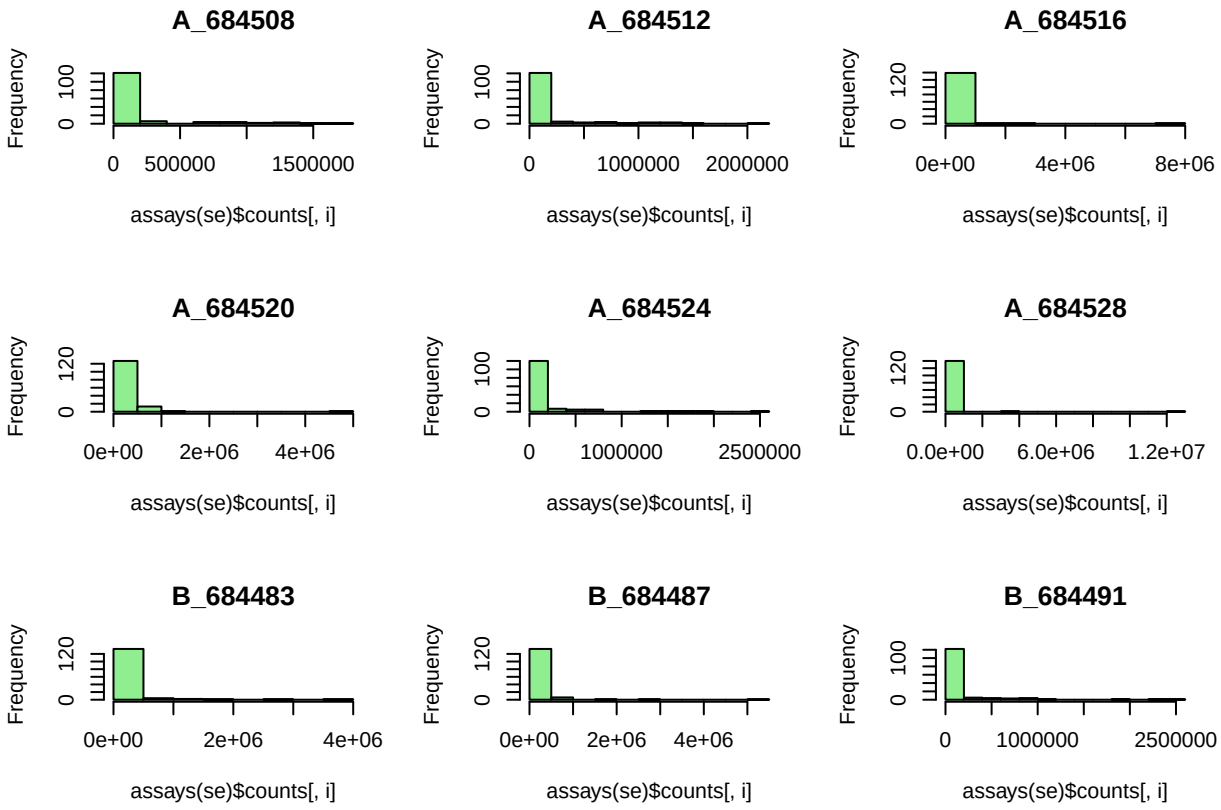
Una vez integrados los datos en un objeto `SummarizedExperiment` y analizado la estructura y obtenida la información de los mismos, vamos a proceder a realizar una exploración general que nos proporcione más información sobre el estado de las muestras. Calcularemos los estadísticos descriptivos habituales, empezando por medidas univariantes y progresando a estadísticos multivariantes.

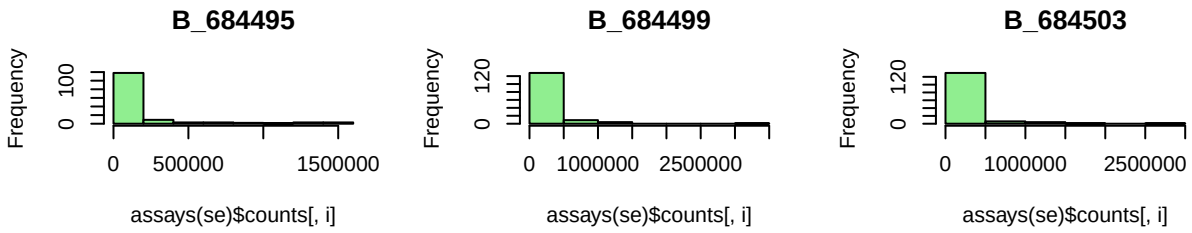
4.1.1. Análisis exploratorio univariante

Comenzaremos con un **análisis estadístico básico** que incluye medidas como la media, el mínimo, el máximo y la desviación estándar de la expresión de cada muestra, para caracterizar la variabilidad de los datos.

```
##         A_684508 A_684512 A_684516 A_684520 A_684524 A_684528 B_684483 B_684487
## Min.          95      336       98      186      114       48      309      192
## 1st Qu.     1261     2815      911     2214     1527      592     2449     2051
## Median      4728    10370     4877     5989     7428     3164    10900    12006
## Mean      140978   141017   141063   140922   140911   140966   141038   141185
## 3rd Qu.    52750    60511    36756    33838    67985    17146    41716    63356
## Max.     1665633  2165933  7204190  4694846  2498885 12543992  3937010  5370106
##         B_684491 B_684495 B_684499 B_684503
## Min.         464       88      164       67
## 1st Qu.     3004     2449     1592     3474
## Median      9611    10563     5836    11010
## Mean      141187   140878   140910   141294
## 3rd Qu.    81266    59358    67631    69077
## Max.     2458026  1515847  3434602  2754573
```

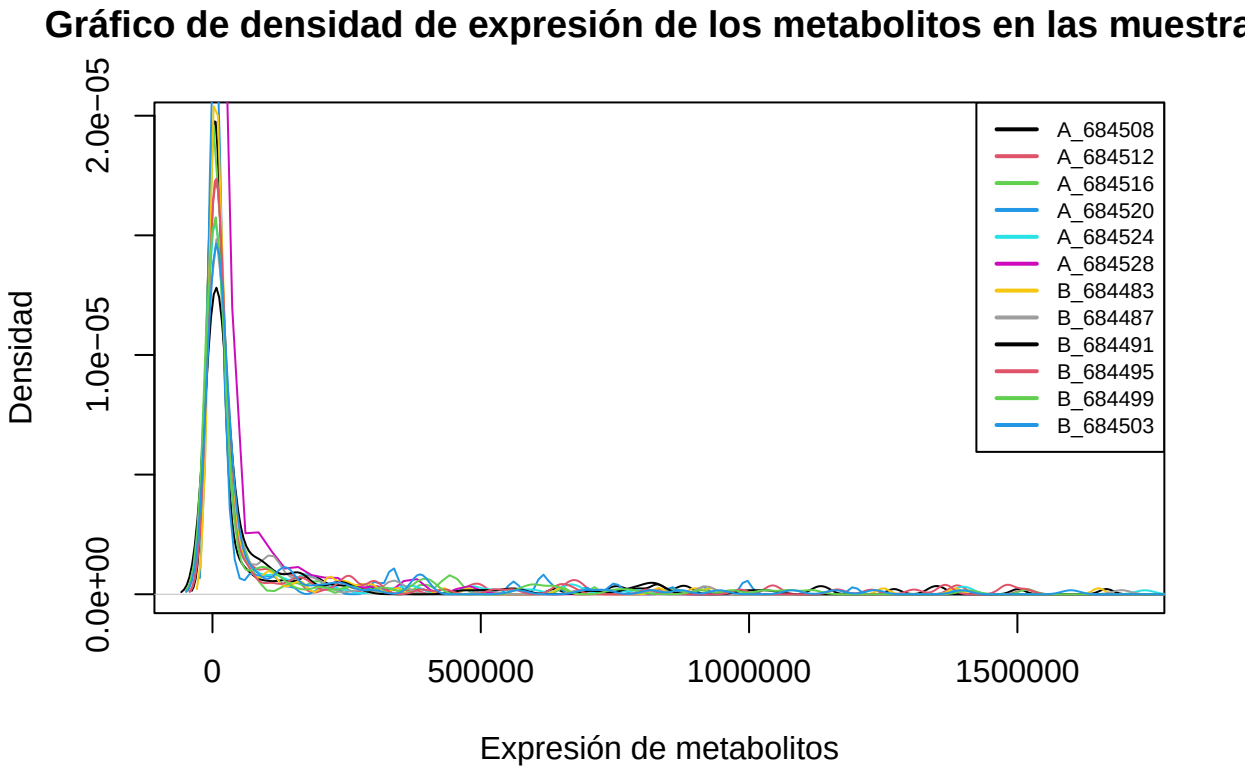
A continuación, vamos a obtener un **histograma** de expresión por cada muestra, que permite observar la frecuencia de niveles de expresión de los metabolitos dentro de cada muestra y detectar tendencias o anomalías específicas.


A_684508
Frequency
0
100
0
500000
1500000
assays(se)$counts[, i]
A_684512
Frequency
0
100
0
1000000
2000000
assays(se)$counts[, i]
A_684516
Frequency
0
120
0e+00
4e+06
8e+06
assays(se)$counts[, i]
A_684520
Frequency
0
120
0e+00
2e+06
4e+06
assays(se)$counts[, i]
A_684524
Frequency
0
100
0
1000000
2500000
assays(se)$counts[, i]
A_684528
Frequency
0
120
0.0e+00
6.0e+06
1.2e+07
assays(se)$counts[, i]
B_684483
Frequency
0
120
0e+00
2e+06
4e+06
assays(se)$counts[, i]
B_684487
Frequency
0
120
0e+00
2e+06
4e+06
assays(se)$counts[, i]
B_684491
Frequency
0
100
0
1000000
2500000
assays(se)$counts[, i]



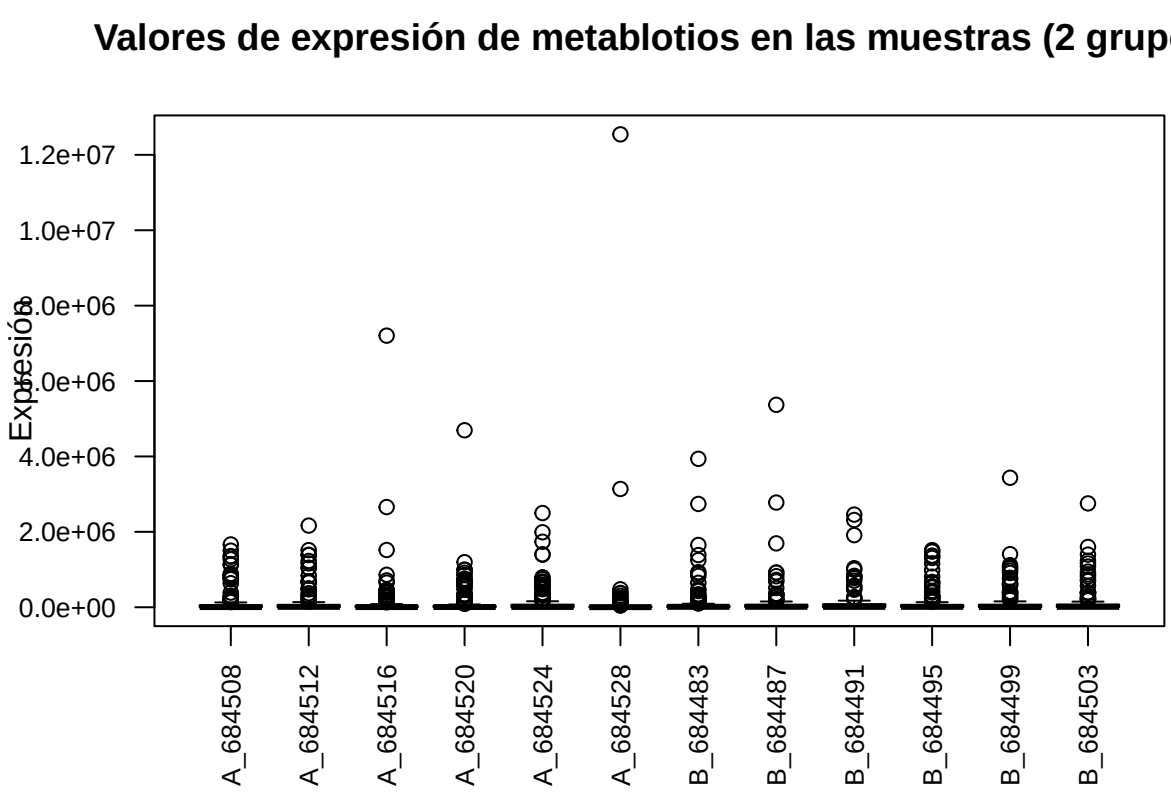
Con estos histogramas podemos hacernos una idea inicial de que todas las muestras tienen valores metabolíticos muy similares, sin mostrar grandes diferencias. Vamos a realizar más gráficos para poder obtener más información al respecto.

Mostramos ahora un **gráfico de densidad** de todas las muestras. Mediante gráfico se observa una visualización de la distribución general de la expresión de los metabolitos en todas las muestras, permitiendo observar cómo se distribuyen los niveles de expresión y detectar posibles diferencias entre muestras. Cada muestra tiene un color diferente, y una leyenda ayuda a identificar cada línea.

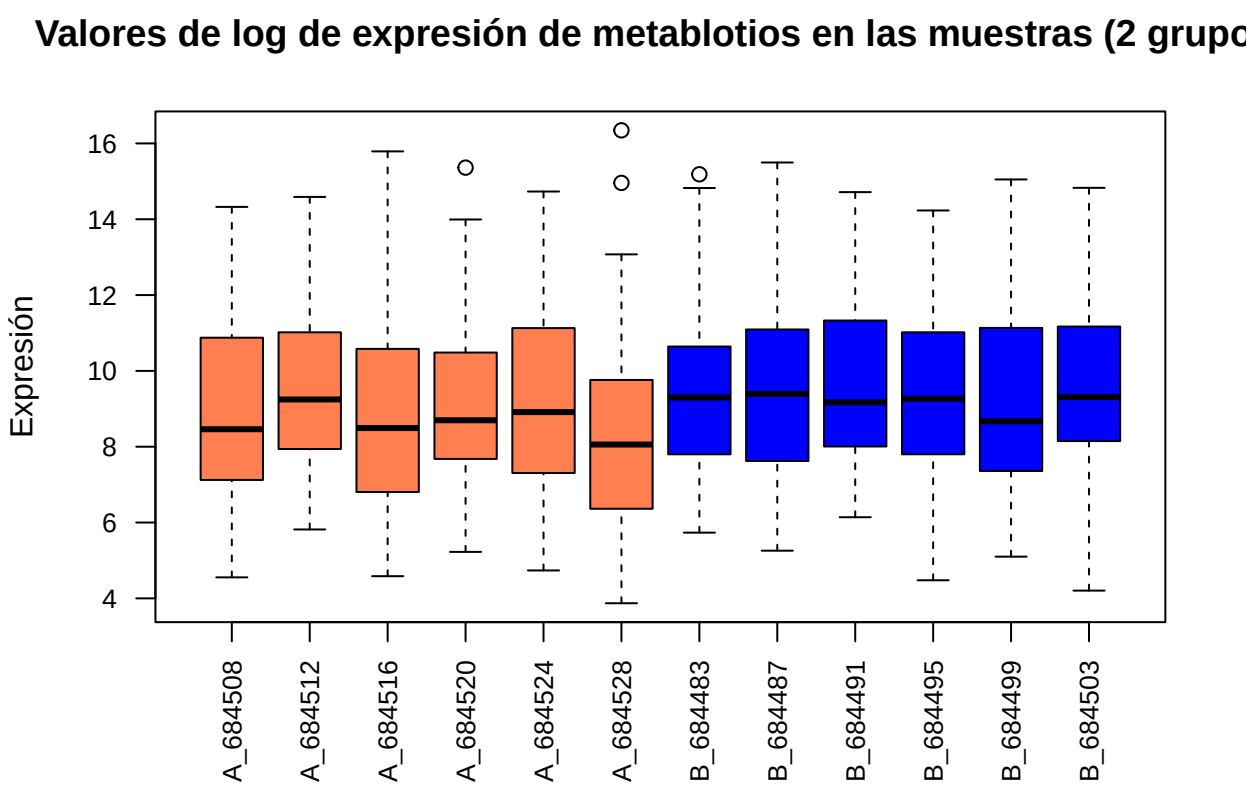


Al igual que con los gráficos anteriores, con el gráfico de densidad observamos que las curvas de densidad son, también, similares en todas las muestras, con algunos picos diferentes pero, en general, sin mostrar grandes diferencias.

Con los **diagramas de cajas (*Boxplot*)** quizá podramos observar de manera más clara las posibles diferencias entre las muestras, puesto que estos diagramas proporcionan una comparación gráfica de la distribución de los niveles de expresión de los metabolitos a través de todas las muestras, identificando posibles valores atípicos y diferencias en la variabilidad, y nos pueden proporcionar alguna pista sobre la conveniencia de realizar algún tipo de procesamiento de los datos.



Podemos comprobar mediantes los diagramas de cajas que los datos presentan asimetría. Vamos a comprobar si esta asimetría podría corregirse escalando los datos mediante **logaritmos**.



En vista de los resultados obtenidos con los diagramas de cajas, mediante los datos escalados logarítmicamente, parece **más razonable trabajar con estos datos transformados**.

4.1.2. Análisis exploratorio multivariante

En esta sección nos vamos a centrar en analizar **las componentes principales (PCA)** del estudio. Así realizaremos un análisis de reducción de dimensionalidad para identificar patrones y posibles agrupaciones entre las muestras, facilitando su visualización, lo que ayudará a detectar relaciones complejas y correlaciones entre metabolitos.

El análisis de componentes principales (PCA) transforma las variables originales en nuevas componentes. Estas nuevas componentes son independientes entre sí (ortogonales), puesto que cada una explica diferentes aspectos de los datos, y además, explican la variabilidad observada, con capacidad decreciente (la primera componente explica la mayor variablidad y la última, la menor).

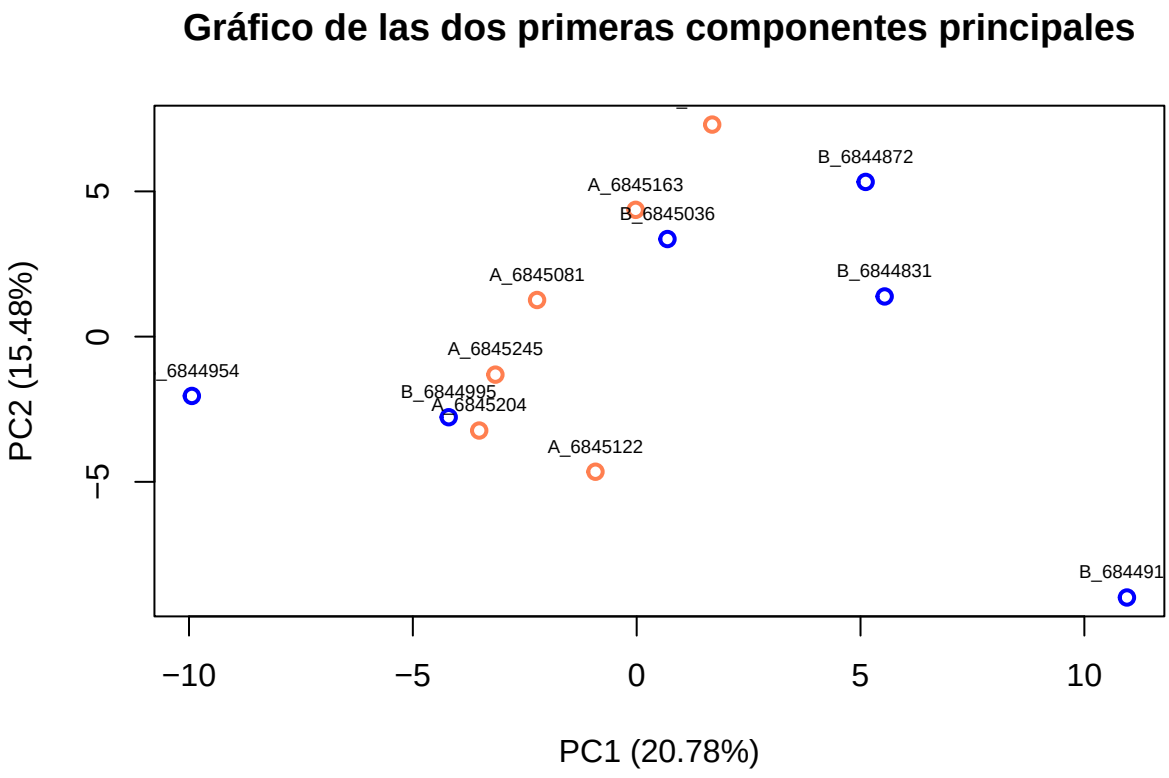
Calculemos, en primer lugar, las **componentes principales (PC)**.

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5     PC6     PC7
## Standard deviation     5.4321 4.6887 4.3912 4.1223 3.8224 3.28776 2.92520
## Proportion of Variance 0.2078 0.1548 0.1358 0.1197 0.1029 0.07612 0.06026
## Cumulative Proportion  0.2078 0.3626 0.4984 0.6181 0.7210 0.79709 0.85735
##                           PC8     PC9    PC10    PC11      PC12
## Standard deviation     2.58586 2.57008 2.27736 1.33322 1.548e-15
## Proportion of Variance 0.04709 0.04652 0.03652 0.01252 0.000e+00
## Cumulative Proportion  0.90444 0.95096 0.98748 1.00000 1.000e+00
```

Podemos comprobar que no se explica más de un 70% de la variabilidad de los datos hasta la componente 5,

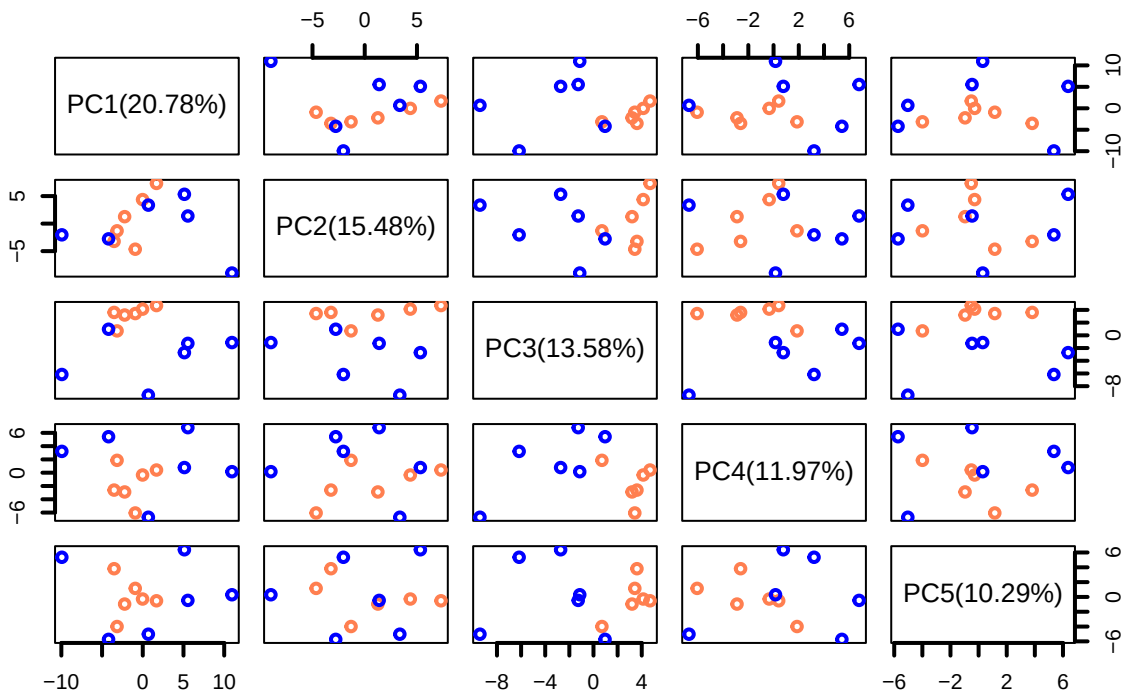
por lo que precisaríamos de las PC desde la 1 a la 5. También vemos que 11 componentes explican el 100% de la variabilidad de los datos, lo que significa que la última componente (PC12) no aporta variabilidad adicional porque tiene una desviación estándar cercana a cero.

Vamos a visualizar, en primer lugar, los resultados de las **dos primeras componentes principales (PC1 y PC2)** mediante la generación de un gráfico.



A continuación, visualizaremos los resultados de las **cinco primeras componentes principales (PC1 a PC5)** mediante la generación de un gráfico, puesto que hemos concluido que precisaríamos de las cinco primeras componentes principales para garantizar la identificación de patrones.

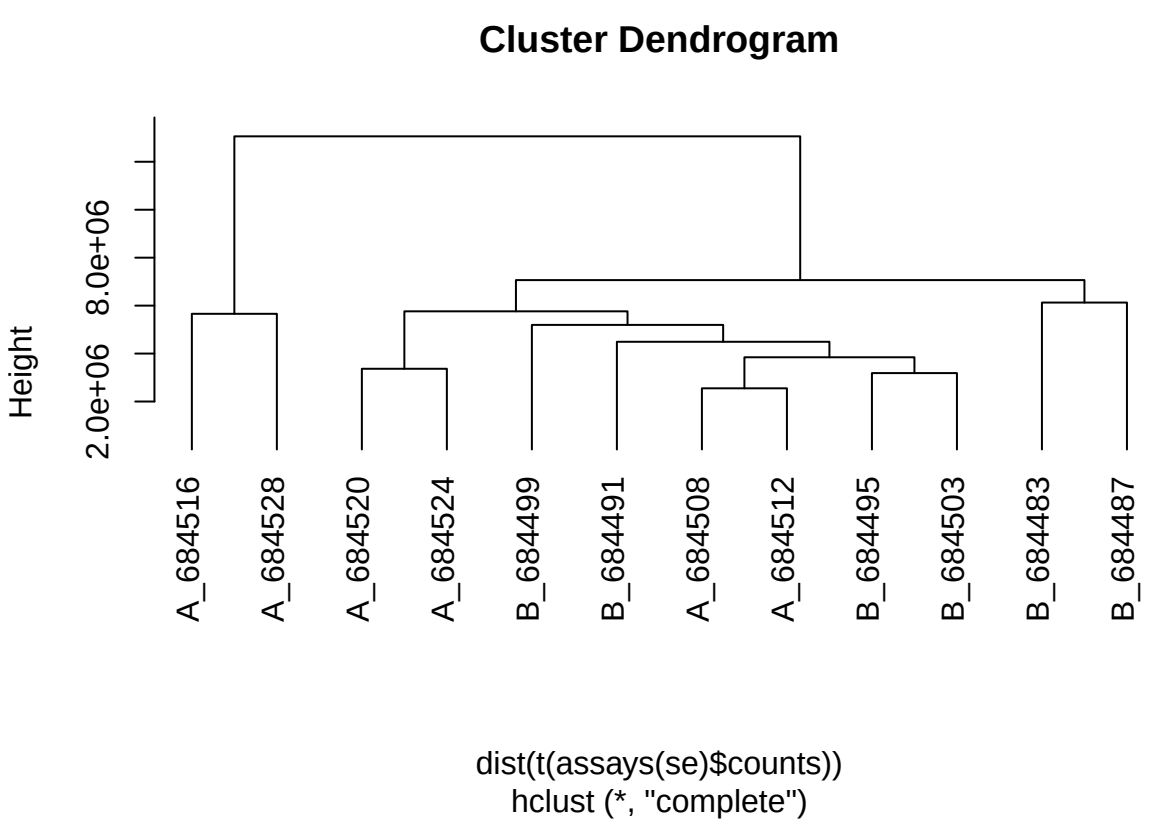
Gráfico de pares de las primeras cinco componentes principales



En ambos gráficos podemos observar como **no se distingue, en general, una agrupación** clara asociada con el grupo (*After* o *Before*), existiendo algunas excepciones, como entre PC1 y PC3, donde se ve una ligera separación entre los grupos. Tampoco se observa una clara separación entre las muestras a partir de los metabolitos ni agrupaciones, lo que nos indica que no hay similitudes en los perfiles de metabolitos y no se detectan grupos biológicos.

Todo esto nos indica que las cargas que cada metabolito tiene asociadas en cada componente principal son bajas y que ningún metabolito contrubuye de manera significativa a cada componente. Esto último lo podemos deducir del bajo porcentaje de varianza que tenía cada componente.

Por último, vamos a utilizar una **agrupación jerárquica (*cluster*)** para visualizar, mediante un dendograma, cualquier posible agrupación de las muestras que no se haya podido detectar con el análisis de componentes principales.



Parece ser que, con la agrupación jerárquica, hemos descubierto dos grupos diferenciados entre las muestras. Esto nos indica que sí debe de haber algún metabolito que contrubuya de manera más significativa a algún componente principal, y que las dos primeras muestras del dendograma se asociarían más a ciertos metabolitos que el resto de las muestras.

5. Discusión, limitaciones y conclusiones del estudio

A través del análisis de componentes principales (PCA) y la agrupación jerárquica, se observó que, aunque las muestras no se agrupan de forma clara en función del estado pre o post trasplante, algunos patrones sugieren que ciertos metabolitos podrían estar asociados a las diferencias metabólicas después del trasplante. La agrupación jerárquica, en particular, indica que puede haber una diferenciación sutil entre las muestras.

Entre las limitaciones del estudio, cabe destacar la limitada cantidad de muestras y la posible necesidad de aplicar técnicas de normalización adicionales para mejorar la calidad de los datos y eliminar un posible efecto batch antes del análisis. Además, los resultados no permiten identificar metabolitos específicos como indicadores claros de los cambios post trasplante, lo cual podría deberse a la variabilidad biológica inherente o a factores de confusión no considerados en este análisis. A pesar de estas limitaciones, los datos utilizados no presentaban problemas, pues no había valores faltantes (NA) y las muestras estaban bien distribuidas.

En conclusión, aunque no se encontraron patrones de agrupación claros, los hallazgos preliminares sugieren que un análisis más detallado, posiblemente con un tamaño de muestra mayor y técnicas de procesamiento de datos adicionales, podría identificar metabolitos clave en la monitorización post trasplante.

Apéndice 1: Repositorio GitHub

El informe final, el documento Rmarkdown original, el objeto contenedor con los datos y los metadatos en formato binario (.Rda), el documento con el código R para la exploración de los datos y los datos y metadatos acerca del dataset se pueden encontrar en el siguiente repositorio de GitHub: https://github.com/BeatrizJimenezGuijarro/Jimenez-Guijarro-Beatriz-PEC1

Apéndice 2: código R

Todo el código R de este informe se puede encontrar tanto en el informe original en formato Rmarkdown ("Jimenez_Guijarro_Beatriz_PEC1.Rmd") como en el documento .R ("Jimenez_Guijarro_Beatriz_PEC1.Rmd") que engloba únicamente las celdas de código utilizado a lo largo del informe y que se ha generado mediante la instrucción `knitr::purl("Jimenez_Guijarro_Beatriz_PEC1.Rmd", output = "Jimenez_Guijarro_Beatriz_PEC1.R")`. Este documento .R se ha incluido en este apéndice dentro de una última celda de código que no se ejecuta y que no se muestra en el informe final, pero sí en el documento Rmarkdown.

Referencias

Bioconductor. 2024. *SummarizedExperiment*. https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html#constructing-a-summarizedexperiment.

Sanchez-Pla, Alex. 2024. "Exploración Multivariante de Datos Ómicos: Descriptivo, PCA y Clustering." Informe. Universitat de Barcelona, Departamento de Genética, Microbiología y Estadística. https://aula.uoc.edu/courses/47009/assignments/527835?module_item_id=1781171.

Teaching, ASP. 2024. "Exploración de Microarrays - Análisis de Datos Ómicos." https://aspteaching.github.io/Analisis_de_datos_omicos-Ejemplo_0-Microarrays/ExploreArrays.html.