Informe de análisis de datos metabolómicos

Paschal Chukwudum Ogbogu Emeghalu

2024-11-06

Table of Contents

Abstract

Este informe tiene como objetivo explorar y analizar los datos metabolómicos de un estudio obtenido de Metabolomics Workbench. Los datos se han procesado y explorado para obtener una visión general de las características de las muestras y sus metabolitos, lo cual puede proporcionar una base para estudios futuros en metabolómica clínica utilizando los mismos.

Objetivos del estudio

Los objetivos de este análisis son los siguientes:

1. Descargar y organizar los datos metabolómicos del estudio, familiarizándose, de esta manera, con las herramientas de R y los flujos de Git pertinentes.

2. Explorar la estructura y contenido del dataset, incluyendo variables y metadatos.
3. Visualizar y analizar los datos para observar patrones generales y diferencias entre los grupos de tratamiento.
4. Identificar limitaciones y proponer posibles direcciones para análisis futuros.

Materiales y métodos

Herramientas utilizadas

Este informe fue realizado en R, utilizando las siguientes bibliotecas:

- `metabolomicsWorkbenchR` para descargar y procesar los datos desde Metabolomics Workbench.
- `SummarizedExperiment` para organizar los datos en un contenedor adecuado.
- `ggplot2` para visualización de datos.

Adicionalmente, se ha utilizado Git y GitHub para el control de versiones y compartir el código con otros colaboradores.

Preparación del entorno de trabajo

Se creó un nuevo proyecto en RStudio y se vinculó con un repositorio en GitHub para preparar el entorno de trabajo. Para ello, utilizamos la opción de crear un proyecto a partir de un control de versiones, seleccionando Git como sistema de control. A continuación, creemos un repositorio en GitHub con el nombre `Ogbogu-Emeghalu-Paschal-PEC1`, y establecimos la conexión entre el proyecto en RStudio y este repositorio en GitHub. Esta configuración permite gestionar el proyecto directamente desde RStudio, llevando un registro detallado de los cambios y facilitando el trabajo colaborativo.

Preparación de los datos

En cuanto a los datos, se optó por utilizar un conjunto de datos de metabolómica disponible en el repositorio de GitHub `metaboData` (https://github.com/nutrimetabolomics/metaboData/). Se accedió a dicho repositorio y se exploró su estructura. En la carpeta '2024-fobitools-UseCase_1' se encuentra varios archivos, entre los cuales 'description.md' que proporciona información referente al conjunto de datos del repositorio. Específicamente, este conjunto de datos se encuentra en el repositorio `metabolomics Workbench` con el ID de ST000291. Cargamos el dataset en el proyecto de R leyéndolo directamente de la base de datos de Metabolomics Workbench utilizando el paquete Bioconductor `metabolomicsWorkbenchR` y lazando una consulta (query) sobre el mismo con los parámetros del estudio como se ve a continuación:

```
library(metabolomicsWorkbenchR)
library(SummarizedExperiment)

## Loading required package: MatrixGenerics
```

```
## Loading required package: matrixStats

##
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars

## Loading required package: GenomicRanges

## Loading required package: stats4

## Loading required package: BiocGenerics

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, table,
##     tapply, union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'
```

```
## The following object is masked from 'package:utils':
##
##     findMatches

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

## Loading required package: GenomeInfoDb

## Loading required package: Biobase

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

##
## Attaching package: 'Biobase'

## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians

## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
```

```r
# Query para obtener los datos del estudio ST000291
query_result <- do_query(
  context = "study", # Contexto de la consulta
  input_item = "study_id", # Item de entrada
  input_value = "ST000291", # Valor de entrada
  output_item = "SummarizedExperiment") # Item de salida
query_result
```

```
## $AN000464
## class: SummarizedExperiment
## dim: 1786 45
## metadata(8): data_source study_id ... description subject_type
## assays(1): ''
## rownames(1786): ME104202 ME104203 ... ME105898 ME105899
## rowData names(3): metabolite_name metabolite_id refmet_name
## colnames(45): a1 a10 ... c8 c9
## colData names(6): local_sample_id study_id ... raw_data Treatment_
##
## $AN000465
## class: SummarizedExperiment
## dim: 747 45
```

```
## metadata(8): data_source study_id ... description subject_type
## assays(1): ''
## rownames(747): ME105925 ME105926 ... ME106646 ME106647
## rowData names(3): metabolite_name metabolite_id refmet_name
## colnames(45): a1 a10 ... c8 c9
## colData names(6): local_sample_id study_id ... raw_data Treatment_
```

Vemos, en primera instancia, la estructura y contenido básicos del estdio, con dos conjuntos de datos de tipo `SummarizedExperiment`. Guadaremos este objeto contenedor con los datos y los metadatos en formato binario (.Rda), los datos en formato texto y los metadatos acerca del dataset en un archivo markdown. Luego procederemos a un análisis exploratorio más detallado de los datos.

```
# Guardamos el objeto contenedor con los datos y los metadatos en formato
binario
save(query_result, file = "metabolomics_data_container.Rda")
```

Exploración y análisis de los datos

El estudio ST000291, disponible en Metabolomics Workbench[1], fue diseñado para investigar los efectos del consumo de jugo de arándano en el metabolismo humano, comparándolo con el jugo de manzana como referencia[2]. Este estudio se realizó mediante enfoques de metabolómica y utilizó espectrometría de masas para analizar cambios en el perfil metabólico de la orina en mujeres jóvenes. Los objetivos principales fueron identificar diferencias significativas en metabolitos después de la ingesta de estos jugos y explorar cómo ciertos compuestos de los arándanos, como los polifenoles, pueden influir en la salud, especialmente en cuano a ofrecer efectos protectores para el sistema cardiovascular.

Estructura de los datos

En el estudio, los objetos etiquetados como $AN000464 y $AN000465 corresponden a diferentes conjuntos de datos metabolómicos derivados de las muestras de orina de los participantes, recolectados tras el consumo de jugo de arándano o de manzana y representados en forma de objetos `SummarizedExperiment`. Concretamente, $AN000464 contiene un conjunto de metabolitos obtenidos utilizando la técnica de la espectrometría de masas (MS) en “modo positivo” (es decir, los metabolitos se ionizan de manera que se generan iones cargados positivamente). Por contra, $AN000465 contiene metabolitos obtenidos en “modo negativo” de la espectrometría de masas (es decir, los metabolitos se ionizan de manera que se generan iones cargados negativamente).

---

[1] Metabolomics Workbench. (2023). Dataset ST000291. https://www.metabolomicsworkbench.org/

[2] https://www.omicsdi.org/dataset/metabolomics_workbench/ST000291

Cabe mencionar que el uso del objeto `SummarizedExperiment` permite organizar los datos en tres componentes principales:

- Assays: Contiene la matriz de intensidades de los metabolitos en cada muestra.
- colData: Incluye las características y anotaciones de las muestras, como el tratamiento y el identificador local.
- rowData: Contiene las anotaciones de los metabolitos, incluyendo nombres y referencias en bases de datos como PubChem o KEGG.

Así, podemos, en primera instancia, verificar la estructura y el contenido general de cada objeto `SummarizedExperiment` de la siguiente manera:

```
# Vemos estructura general del dataset. Utilizamos max.level para simplificar la salida
str(query_result$AN000464, max.level = 3)
```

```
## Formal class 'SummarizedExperiment' [package "SummarizedExperiment"] with 5 slots
##   ..@ colData        :Formal class 'DFrame' [package "S4Vectors"] with 6 slots
##   ..@ assays         :Formal class 'SimpleAssays' [package "SummarizedExperiment"] with 1 slot
##   ..@ NAMES          : chr [1:1786] "ME104202" "ME104203" "ME104204" "ME104205" ...
##   ..@ elementMetadata:Formal class 'DFrame' [package "S4Vectors"] with 6 slots
##   ..@ metadata       :List of 8
##   .. ..$ data_source     : chr "Metabolomics Workbench"
##   .. ..$ study_id        : chr "ST000291"
##   .. ..$ analysis_id     : chr "AN000464"
##   .. ..$ analysis_summary: chr "ESI Positive mode"
##   .. ..$ units           : chr "Peak area"
##   .. ..$ name            : chr "ST000291:AN000464"
##   .. ..$ description     : chr "LC-MS Based Approaches to Investigate Metabolomic Differences in the Urine of Young Women after Drinking Cranbe"|
__truncated__
##   .. ..$ subject_type    : logi NA
```

El análisis de los metabolitos en modo positivo revela un conjunto de datos robusto que contiene 1786 observaciones distribuidas en 45 muestras. Cada muestra, identificada por su local_sample_id, está asociada a un único study_id, lo que indica que provienen del mismo experimento. La cantidad de metabolitos observados es notablemente alta, lo que proporciona un contexto adecuado para detectar variaciones significativas y potencialmente identificar biomarcadores relevantes asociados con la intervención dietética. Sin embargo, también se observan valores faltantes en algunas mediciones, un fenómeno común en estudios metabolómicos, que puede influir en el análisis estadístico y la interpretación de los resultados.

Podemos extraer y guardar la matriz de intensidades de metabolitos, los metadatos de las muestras y los nombres y datos de los metabolitos de la siguiente manera:

```
# Extraemos y guardamos la matriz de intensidades de metabolitos
metabolite_data_positive <- assay(query_result$AN000464)
write.csv(metabolite_data_positive, "metabolomics_data_positive.csv",
row.names = TRUE)

# Extraemos y guardamos los metadatos de las muestras
sample_metadata_positive <- colData(query_result$AN000464)
write.csv(sample_metadata_positive, "metabolomics_metadata_positive.csv",
row.names = TRUE)

# Extraemos y guardamos los nombres y datos de los metabolitos
metabolite_metadata_positive <- rowData(query_result$AN000464)
write.csv(metabolite_metadata_positive, "metabolites_metadata_positive.csv",
row.names = TRUE)
```

Y hacemos lo propio para el subconjunto de datos en modo negativo:

```
# Extraemos y guardamos la matriz de intensidades de metabolitos
metabolite_data_negative <- assay(query_result$AN000465)
write.csv(metabolite_data_negative, "metabolomics_data_negative.csv",
row.names = TRUE)

# Extraemos y guardamos los metadatos de las muestras
sample_metadata_negative <- colData(query_result$AN000465)
write.csv(sample_metadata_negative, "metabolomics_metadata_negative.csv",
row.names = TRUE)

# Extraemos y guardamos los nombres y datos de los metabolitos
metabolite_metadata_negative <- rowData(query_result$AN000465)
write.csv(metabolite_metadata_negative, "metabolites_metadata_negative.csv",
row.names = TRUE)
```

También verificamos la estructura y el contenido general del objeto `SummarizedExperiment`:

```
# Vemos la estructura general del dataset
str(query_result$AN000465, max.level = 3) # Para el dataset de modo negativo

## Formal class 'SummarizedExperiment' [package "SummarizedExperiment"] with
5 slots
##   ..@ colData        :Formal class 'DFrame' [package "S4Vectors"] with 6
slots
##   ..@ assays         :Formal class 'SimpleAssays' [package
"SummarizedExperiment"] with 1 slot
##   ..@ NAMES          : chr [1:747] "ME105925" "ME105926" "ME105922"
"ME105923" ...
##   ..@ elementMetadata:Formal class 'DFrame' [package "S4Vectors"] with 6
slots
```

```
##   ..@ metadata       :List of 8
##   .. ..$ data_source     : chr "Metabolomics Workbench"
##   .. ..$ study_id        : chr "ST000291"
##   .. ..$ analysis_id     : chr "AN000465"
##   .. ..$ analysis_summary: chr "ESI Negative mode"
##   .. ..$ units           : chr "Peak area"
##   .. ..$ name            : chr "ST000291:AN000465"
##   .. ..$ description     : chr "LC-MS Based Approaches to Investigate
Metabolomic Differences in the Urine of Young Women after Drinking Cranbe"|
__truncated__
##   .. ..$ subject_type    : logi NA
```

Vemos que el análisis en modo negativo incluye 747 observaciones de metabolitos a partir de 45 muestras. Al igual que en el modo positivo, cada muestra se identifica por un local_sample_id y se agrupa bajo el mismo study_id. Aunque el número de metabolitos es menor que en el modo positivo, la variedad de muestras sugiere una cobertura adecuada para explorar las diferencias metabólicas. Este conjunto de datos también presenta valores faltantes, y la cantidad reducida de metabolitos podría limitar la capacidad para detectar cambios significativos en respuesta al tratamiento con jugo de arándano. A pesar de esta limitación, el análisis en modo negativo sigue siendo esencial para complementar la comprensión general de cómo esta intervención dietética afecta el metabolismo.

Comparando los dos conjuntos de datos, se observa que el análisis en modo positivo ofrece una mayor diversidad y número de metabolitos, lo cual podría facilitar un análisis más detallado y, por lo tanto, fiable. La presencia de valores faltantes es un aspecto que ambos conjuntos comparten, aunque su impacto puede ser más pronunciado en el análisis en modo negativo debido a la menor cantidad de metabolitos disponibles.

Por lo tanto, aunque ambos conjuntos de datos son complementarios y proporcionan una visión integral sobre el impacto del jugo de arándano en el perfil metabólico, el análisis en modo positivo podría permitir una exploración más exhaustiva de las diferencias metabólicas y la identificación de patrones que podrían no ser tan evidentes en el modo negativo.

Datos de las muestras

Para comprender la distribución de las muestras, analizamos las variables disponibles en `colData`, siendo los resultados idénticos para ambos conjuntos de datos (modo positivo y modo negativo) ya que, como hemos visto anteriormente, ambos comparten el mismo número y estructura de muestras:

```
# Vemos los datos de muestra
col_data <- colData(query_result$AN000464)
col_data
```

```
## DataFrame with 45 rows and 6 columns
##     local_sample_id    study_id sample_source mb_sample_id    raw_data
##         <character> <character>   <character>  <character> <character>
## a1               a1    ST000291         Urine     SA013237
```

```
## a10             a10    ST000291         Urine     SA013236
## a11             a11    ST000291         Urine     SA013246
## a12             a12    ST000291         Urine     SA013244
## a13             a13    ST000291         Urine     SA013245
## ...             ...         ...           ...          ...          ...
## c4               c4    ST000291         Urine     SA013251
## c6               c6    ST000291         Urine     SA013250
## c7               c7    ST000291         Urine     SA013248
## c8               c8    ST000291         Urine     SA013249
## c9               c9    ST000291         Urine     SA013254
##                                Treatment_
##                                  <factor>
## a1      Urine after drinking apple juice
## a10     Urine after drinking apple juice
## a11     Urine after drinking apple juice
## a12     Urine after drinking apple juice
## a13     Urine after drinking apple juice
## ...                                   ...
## c4  Urine after drinking cranberry juice
## c6  Urine after drinking cranberry juice
## c7  Urine after drinking cranberry juice
## c8  Urine after drinking cranberry juice
## c9  Urine after drinking cranberry juice
```

Vemos que cada muestra está etiquetada con un identificador único (local_sample_id) y se clasifica según su fuente, que en todos los casos es orina. También se proporciona un identificador de muestra biológica (mb_sample_id), que representa el identificador único para cada muestra en el contexto del laboratorio o del análisis experimental.

La columna Treatment_ clasifica las muestras en dos grupos según el tratamiento recibido antes de la recolección de la orina: las muestras fueron recolectadas después de consumir zumo de manzana o zumo de arándano. Esta categorización es esencial para el análisis comparativo, ya que permite identificar posibles variaciones en el perfil metabolómico de la orina dependiendo del tipo de jugo ingerido, lo que puede reflejar la influencia de los diferentes compuestos bioactivos presentes en estos alimentos.

Datos de los metabolitos

Los datos de los metabolitos se encuentran en `rowData`. Exploramos el contenido para verificar los nombres de los metabolitos y sus identificadores, como `metabolite_id`, que permite enlazar con otras bases de datos de compuestos:

```
rowData(query_result$AN000464)
```

```
## DataFrame with 1786 rows and 3 columns
##                 metabolite_name metabolite_id            refmet_name
##                     <character>   <character>            <character>
## ME104202 10-Desacetyltaxuyunn..      ME104202
## ME104203 10-Hydroxydecanoic a..      ME104203 10-Hydroxydecanoic a..
## ME104204      10-Oxodecanoate_1      ME104204    10-Oxodecanoic acid
```

```
## ME104205      10-Oxodecanoate_2      ME104205    10-Oxodecanoic acid
## ME104206 11beta,21-Dihydroxy-..      ME104206 11beta,21-Dihydroxy-..
## ...                          ...          ...                    ...
## ME105895            Zingerone_3      ME105895              Zingerone
## ME105896          Zinnimidine_1      ME105896            Zinnimidine
## ME105897          Zinnimidine_2      ME105897            Zinnimidine
## ME105898          Zinnimidine_3      ME105898            Zinnimidine
## ME105899               Zoxamide      ME105899
```

```
cat("\n")
```

```
rowData(query_result$AN000465)
```

```
## DataFrame with 747 rows and 3 columns
##                  metabolite_name metabolite_id            refmet_name
##                      <character>   <character>            <character>
## ME105925 10-Deacetyl-2-debenz..      ME105925 10-Deacetyl-2-debenz..
## ME105926        10-Oxodecanoate      ME105926    10-Oxodecanoic acid
## ME105922 1,1-Diethyl-2-hydrox..      ME105922
## ME105923 1,1-Diethyl-2-hydrox..      ME105923
## ME105924 1,2-Dihydroxynaphtha..      ME105924
## ...                          ...          ...                    ...
## ME106643           Xanthosine_2      ME106643             Xanthosine
## ME106644            Zalcitabine      ME106644            Zalcitabine
## ME106645            Zinnimidine      ME106645            Zinnimidine
## ME106646          Zizybeoside I      ME106646
## ME106647           Zoxazolamine      ME106647           Zoxazolamine
```

En el conjunto AN000464, se listan 1786 metabolitos con tres columnas: el nombre del metabolito (metabolite_name), el identificador único (metabolite_id) y el nombre de referencia (refmet_name). Algunos metabolitos, como 10-Hydroxydecanoic acid y 10-Oxodecanoic acid, tienen sus nombres de referencia claramente asignados, lo que permite una identificación más precisa en el análisis. Otros, sin embargo, carecen de nombre de referencia, lo que podría indicar una menor caracterización previa o menor abundancia en la literatura científica.

Por otro lado, en el conjunto AN000465, se incluyen 747 metabolitos, también con tres columnas: metabolite_name, metabolite_id, y refmet_name. Al comparar ambos conjuntos, AN000465 tiene menos del 50% de la cantidad de metabolitos encontrados en AN000464. Sin embargo, ambos conjuntos incluyen metabolitos que se identifican con sus nombres de referencia, como 10-Oxodecanoic acid, lo cual sugiere una coincidencia en ciertos metabolitos entre los dos conjuntos, lo que podría tener implicaciones significativas en términos de consistencia de los datos experimentales.

Matriz de intensidades de los metabolitos

La matriz de intensidades contiene los valores medidos para cada metabolito en cada muestra. Realizamos un análisis descriptivo básico de los valores de intensidad de los metabolitos. Podemos centrarnos únicamente en el conjunto AN000464, ya que son resultados más detallados y extensos:

```r
# Obtenemos la matriz de intensidades y visualizamos algunos datos
metabolite_data <- assay(query_result$AN000464)
summary(metabolite_data)
```

```
##        a1                  a10                  a11
##  Min.   :9.460e+02   Min.   :9.480e+02   Min.   :1.120e+03
##  1st Qu.:3.058e+04   1st Qu.:2.680e+04   1st Qu.:1.280e+05
##  Median :2.525e+05   Median :2.700e+05   Median :8.955e+05
##  Mean   :1.079e+07   Mean   :1.702e+07   Mean   :2.828e+07
##  3rd Qu.:1.600e+06   3rd Qu.:1.950e+06   3rd Qu.:4.338e+06
##  Max.   :6.050e+09   Max.   :1.210e+10   Max.   :2.190e+10
##  NA's   :74          NA's   :81          NA's   :40
##       a12                 a13                  a14
##  Min.   :1.270e+03   Min.   :8.770e+02   Min.   :8.380e+02
##  1st Qu.:1.218e+05   1st Qu.:5.200e+04   1st Qu.:1.360e+04
##  Median :9.095e+05   Median :4.340e+05   Median :1.540e+05
##  Mean   :3.441e+07   Mean   :1.865e+07   Mean   :1.065e+07
##  3rd Qu.:5.355e+06   3rd Qu.:2.452e+06   3rd Qu.:1.220e+06
##  Max.   :2.780e+10   Max.   :1.230e+10   Max.   :6.570e+09
##  NA's   :50          NA's   :58          NA's   :123
##       a15                 a16                  a17
##  Min.   :7.910e+02   Min.   :9.300e+02   Min.   :6.980e+02
##  1st Qu.:5.970e+04   1st Qu.:2.488e+04   1st Qu.:1.505e+04
##  Median :4.330e+05   Median :2.795e+05   Median :1.430e+05
##  Mean   :1.780e+07   Mean   :1.442e+07   Mean   :9.344e+06
##  3rd Qu.:2.505e+06   3rd Qu.:1.898e+06   3rd Qu.:1.280e+06
##  Max.   :1.430e+10   Max.   :9.770e+09   Max.   :5.330e+09
##  NA's   :51          NA's   :90          NA's   :119
##        a2                  a4                   a6
##  Min.   :1.190e+03   Min.   :8.350e+02   Min.   :6.870e+02
##  1st Qu.:1.570e+05   1st Qu.:8.130e+04   1st Qu.:2.192e+04
##  Median :9.220e+05   Median :5.760e+05   Median :2.325e+05
##  Mean   :2.957e+07   Mean   :2.224e+07   Mean   :1.331e+07
##  3rd Qu.:4.780e+06   3rd Qu.:2.910e+06   3rd Qu.:1.828e+06
##  Max.   :2.090e+10   Max.   :1.530e+10   Max.   :8.930e+09
##  NA's   :39          NA's   :45          NA's   :104
##        a7                  a8                   a9
##  Min.   :9.420e+02   Min.   :1.070e+03   Min.   :9.040e+02
##  1st Qu.:7.915e+04   1st Qu.:9.218e+04   1st Qu.:1.360e+04
##  Median :5.540e+05   Median :6.625e+05   Median :1.650e+05
##  Mean   :1.605e+07   Mean   :2.480e+07   Mean   :1.040e+07
##  3rd Qu.:2.768e+06   3rd Qu.:3.665e+06   3rd Qu.:1.208e+06
##  Max.   :9.710e+09   Max.   :1.640e+10   Max.   :5.900e+09
##  NA's   :68          NA's   :46          NA's   :120
##        b1                  b10                  b11
##  Min.   :1.170e+03   Min.   :6.270e+02   Min.   :1.140e+03
##  1st Qu.:1.080e+05   1st Qu.:8.320e+04   1st Qu.:2.100e+05
##  Median :7.230e+05   Median :5.940e+05   Median :1.180e+06
##  Mean   :2.655e+07   Mean   :2.233e+07   Mean   :3.362e+07
##  3rd Qu.:3.570e+06   3rd Qu.:3.490e+06   3rd Qu.:6.012e+06
```

```
##  Max.   :1.920e+10   Max.   :1.550e+10   Max.   :2.070e+10
##  NA's   :40          NA's   :59          NA's   :30
##       b12                 b13                 b14
##  Min.   :1.150e+03   Min.   :6.830e+02   Min.   :9.310e+02
##  1st Qu.:1.290e+05   1st Qu.:4.848e+04   1st Qu.:4.315e+04
##  Median :7.890e+05   Median :4.215e+05   Median :3.960e+05
##  Mean   :2.807e+07   Mean   :1.960e+07   Mean   :1.740e+07
##  3rd Qu.:4.250e+06   3rd Qu.:2.632e+06   3rd Qu.:2.275e+06
##  Max.   :2.300e+10   Max.   :1.130e+10   Max.   :1.240e+10
##  NA's   :49          NA's   :56          NA's   :67
##       b15                 b16                 b17
##  Min.   :8.990e+02   Min.   :8.490e+02   Min.   :9.440e+02
##  1st Qu.:1.012e+05   1st Qu.:2.232e+04   1st Qu.:2.468e+04
##  Median :7.035e+05   Median :2.300e+05   Median :2.600e+05
##  Mean   :2.379e+07   Mean   :1.214e+07   Mean   :9.849e+06
##  3rd Qu.:3.520e+06   3rd Qu.:1.588e+06   3rd Qu.:1.450e+06
##  Max.   :1.650e+10   Max.   :7.320e+09   Max.   :6.810e+09
##  NA's   :44          NA's   :80          NA's   :90
##        b2                  b4                  b6
##  Min.   :9.760e+02   Min.   :9.950e+02   Min.   :7.460e+02
##  1st Qu.:1.422e+05   1st Qu.:1.570e+05   1st Qu.:3.620e+04
##  Median :7.880e+05   Median :9.840e+05   Median :3.380e+05
##  Mean   :2.600e+07   Mean   :3.701e+07   Mean   :1.698e+07
##  3rd Qu.:3.890e+06   3rd Qu.:5.042e+06   3rd Qu.:2.320e+06
##  Max.   :1.700e+10   Max.   :2.610e+10   Max.   :1.270e+10
##  NA's   :32          NA's   :38          NA's   :77
##        b7                  b8                  b9
##  Min.   :9.590e+02   Min.   :1.120e+03   Min.   :8.250e+02
##  1st Qu.:4.260e+04   1st Qu.:1.302e+05   1st Qu.:1.560e+04
##  Median :3.760e+05   Median :8.090e+05   Median :1.880e+05
##  Mean   :1.322e+07   Mean   :2.311e+07   Mean   :1.005e+07
##  3rd Qu.:2.050e+06   3rd Qu.:4.145e+06   3rd Qu.:1.330e+06
##  Max.   :7.870e+09   Max.   :1.430e+10   Max.   :6.970e+09
##  NA's   :67          NA's   :44          NA's   :99
##        c1                  c10                 c11
##  Min.   :7.440e+02   Min.   :9.760e+02   Min.   :1.150e+03
##  1st Qu.:4.405e+04   1st Qu.:6.502e+04   1st Qu.:1.230e+05
##  Median :3.810e+05   Median :5.130e+05   Median :8.390e+05
##  Mean   :1.617e+07   Mean   :1.770e+07   Mean   :3.173e+07
##  3rd Qu.:2.152e+06   3rd Qu.:2.998e+06   3rd Qu.:4.460e+06
##  Max.   :1.050e+10   Max.   :1.240e+10   Max.   :2.040e+10
##  NA's   :64          NA's   :64          NA's   :41
##       c12                 c13                 c14
##  Min.   :9.420e+02   Min.   :1.080e+03   Min.   :8.960e+02
##  1st Qu.:1.638e+05   1st Qu.:6.090e+04   1st Qu.:7.888e+04
##  Median :9.460e+05   Median :5.095e+05   Median :5.900e+05
##  Mean   :4.033e+07   Mean   :1.976e+07   Mean   :1.907e+07
##  3rd Qu.:4.918e+06   3rd Qu.:2.628e+06   3rd Qu.:2.932e+06
##  Max.   :2.380e+10   Max.   :1.120e+10   Max.   :9.930e+09
##  NA's   :38          NA's   :56          NA's   :50
```

```
##       c15                 c16                 c17
##  Min.   :9.970e+02   Min.   :7.140e+02   Min.   :6.900e+02
##  1st Qu.:1.132e+05   1st Qu.:4.012e+04   1st Qu.:2.282e+04
##  Median :7.330e+05   Median :3.640e+05   Median :2.410e+05
##  Mean   :3.485e+07   Mean   :1.658e+07   Mean   :1.296e+07
##  3rd Qu.:4.415e+06   3rd Qu.:1.958e+06   3rd Qu.:1.540e+06
##  Max.   :2.140e+10   Max.   :7.550e+09   Max.   :5.920e+09
##  NA's   :40          NA's   :68          NA's   :76
##        c2                  c4                  c6
##  Min.   :1.030e+03   Min.   :8.880e+02   Min.   :1.160e+03
##  1st Qu.:1.002e+05   1st Qu.:3.945e+04   1st Qu.:1.120e+05
##  Median :7.500e+05   Median :4.040e+05   Median :7.450e+05
##  Mean   :3.653e+07   Mean   :1.385e+07   Mean   :3.059e+07
##  3rd Qu.:3.870e+06   3rd Qu.:2.300e+06   3rd Qu.:3.835e+06
##  Max.   :2.110e+10   Max.   :1.060e+10   Max.   :1.710e+10
##  NA's   :44          NA's   :67          NA's   :47
##        c7                  c8                  c9
##  Min.   :6.380e+02   Min.   :1.040e+03   Min.   :8.310e+02
##  1st Qu.:2.312e+04   1st Qu.:8.805e+04   1st Qu.:2.505e+04
##  Median :2.825e+05   Median :5.890e+05   Median :2.550e+05
##  Mean   :1.768e+07   Mean   :2.277e+07   Mean   :1.396e+07
##  3rd Qu.:1.840e+06   3rd Qu.:3.020e+06   3rd Qu.:1.770e+06
##  Max.   :9.750e+09   Max.   :1.480e+10   Max.   :7.420e+09
##  NA's   :84          NA's   :50          NA's   :79
```

El análisis de los datos de metabolitos de AN000464 revela una variabilidad considerable en los valores de concentración de metabolitos en diferentes muestras. Podemos destacar algunos aspectos clave del resumen:

- Distribución de valores: Cada muestra presenta valores mínimos relativamente bajos, alrededor de 600 a 1200 unidades, mientras que los máximos alcanzan hasta el orden de 10^10, lo cual indica una gran variabilidad en las concentraciones entre los metabolitos. Los valores de media también son elevados, oscilando entre 10^6 y 10^7, lo que sugiere la presencia de algunos valores atípicos altos.

- Valores medios y medianas: Las medias son considerablemente más altas que las medianas, especialmente en las muestras a10 a a13 y b1 a b13, lo que sugiere una distribución sesgada hacia valores altos en estas muestras, debido probablemente a unos pocos metabolitos presentes en altas concentraciones.

- Distribución percentil y rango intercuartil: La mayoría de los valores están concentrados entre el primer y tercer cuartil, mientras que los valores extremos se alejan mucho de este rango, lo que refuerza el sesgo hacia concentraciones altas.

- Datos faltantes (NA's): Hay una cantidad notable de datos faltantes (NA's) en las muestras, especialmente en algunas como a14 (123 valores faltantes) y c1 (64 valores faltantes), lo cual podría ser debido a la ausencia de detección en ciertos metabolitos o errores de medición.

Agrupaciones y relaciones entre muestras y grupos de tratamientos

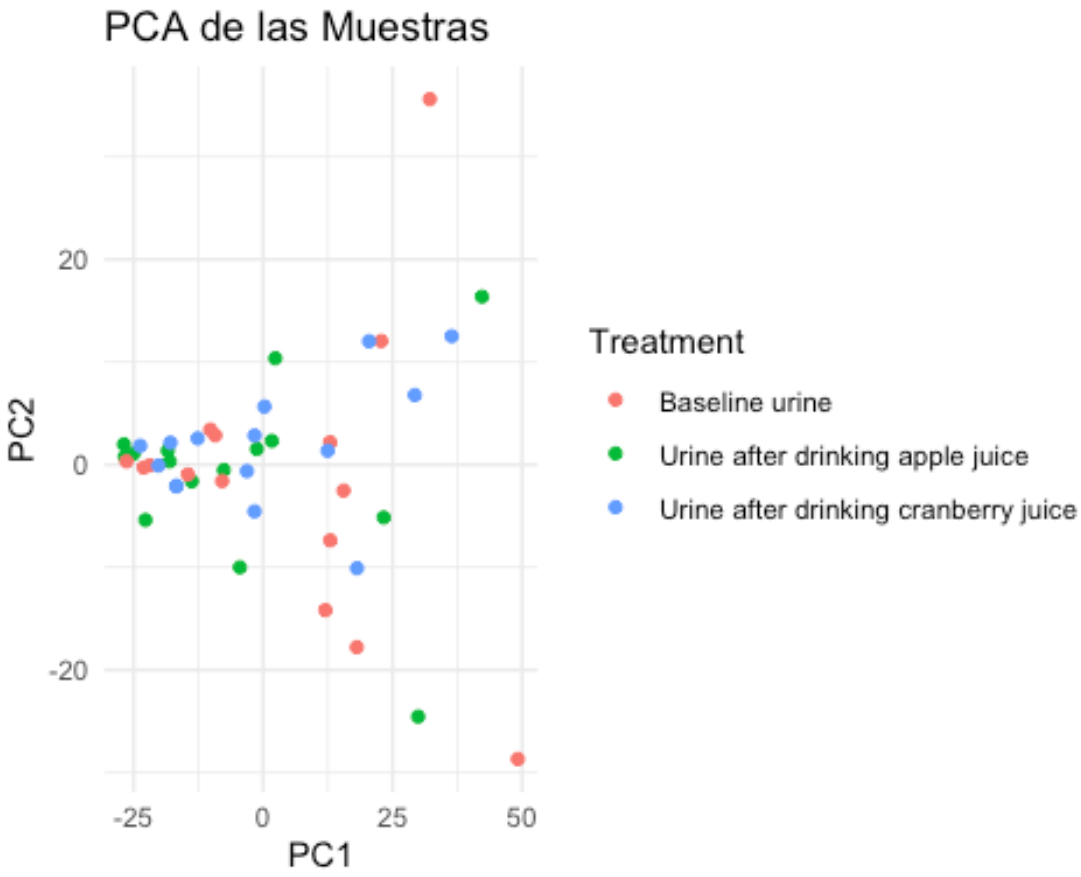
Para visualizar las relaciones entre muestras y grupos de tratamientos, podemos elaborar un análisis de componentes principales (PCA), que nos permitirá explorar la variabilidad general de los datos y visualizar posibles agrupaciones o patrones en las muestras según sus perfiles metabolómicos. Este análisis es conveniente en nuestro caso dado el número elevado de metabolitos del que se dispone en el estudio:

```
# Librería para el PCA y visualización
library(ggplot2)

metabolite_data <- metabolite_data[complete.cases(metabolite_data), ]

# Calculamos el PCA
pca <- prcomp(t(metabolite_data), scale. = TRUE)
pca_df <- as.data.frame(pca$x)
pca_df$Treatment <- col_data$Treatment_

# Visualización PCA
ggplot(pca_df, aes(x = PC1, y = PC2, color = Treatment)) +
  geom_point() +
  labs(title = "PCA de las Muestras", x = "PC1", y = "PC2") +
  theme_minimal()
```



La gráfica muestra una representación bidimensional de la variabilidad en los datos de intensidad de metabolitos en función de las muestras (orina antes y después de consumir jugo de manzana y jugo de arándano). Utiliza los primeros dos componentes principales (PC1 y PC2) para resumir la mayor parte de la variación en los datos originales. En otras palabras, PC1 y PC2 representan combinaciones lineales de las intensidades de los metabolitos, que maximizan la variabilidad total en los datos. PC1 (eje x) es el componente con la mayor variabilidad, y PC2 (eje y) es el siguiente en importancia. Cada punto en la gráfica representa una muestra de orina y está coloreado según el tratamiento correspondiente (orina basal, orina después de beber jugo de manzana y orina después de beber jugo de arándano). Fijándonos en la agrupación de las muestras por tratamiento, parece que hay cierto solapamiento, pero también hay algunas muestras que se separan, especialmente en el eje de PC1. Así, el perfil de metabolitos de las muestras no está claramente diferenciada en función dde los tratamientos.

Discusión, limitaciones y conclusiones del estudio

Este análisis exploratorio inicial nos ha proporcionado una comprensión general de la estructura y características del dataset. A través de la descarga y la preparación de los datos desde Metabolomics Workbench, la organización en sendos contenedores `SummarizedExperiment`, y el análisis de los datos (utilizando técnicas como visualizaciones por PCA), hemos podido identificar las características fundamentales de las muestras y los

metabolitos, explorando de esta froma la estructura y organización de un conjunto de datos grande y complejo.

Algunos aspectos y limitaciones observados durante el análisis son los siguientes:

- Al descargar los datos desde Metabolomics Workbench, encontramos que el estudio está dividido en dos análisis (analysis_id), cada uno de los cuales captura diferentes modos (positivo y negativo) de ionización de metabolitos. Esta división impica trabajar bien con uno de los subconjuntos (el positivo por ejemplo al ser más completo) o con los dos subconjuntos de datos. Esto puede generar inconsistencias si por ejemplo se mantienen criterios homogéneos en el procesamiento de los dos subjuntos de datos.
- El dataset contiene una gran cantidad de variables (metabolitos), por lo que se requiere de un procesamiento profundo con técnicas analíticas para poder detectar patrones concluyentes en el comportamiento de las mismas y extrapolar interpretaciones funcionales o metabólicas de los resultados.

No obstante, este análisis ha proporcionado una primera experiencia práctica en la exploración de datos metabolómicos y en el uso de paquetes específicos de R para bioinformática. Se alcanzaron de esa forma los objetivos de familiarizarse con la descarga de datos, la creación de un contenedor de datos, y la exploración inicial del dataset. Estos análisis deben en cualquier caso profundizarse y completarse con técnicas de normalización y análisis estadísticos avanzados como estudios de correlación, análisis de varianza (ANOVA) o incluso análisis multivariantes para obtener conclusiones más fiables y extraer conclusiones biológicamente más relevantes.

Reposición de los datos en GitHub

Tras la realización del presente informe y, con ello, la generación del archivo .Rda, el código de exploración en R, los datos en formato texto y los metadatos en Markdown, se ha confirmado que estos archivos están en el directorio del proyecto con `system("git status")`. Con `system("git add .")` se han agregado al área de control de versiones de Git, y con `system("git commit -m 'Archivos informe análisis de datos metabolómicos'")` se han registrado los cambios en el repositorio local. Finalmente, se ha sincronizado el repositorio local con el repositorio remoto de GitHub usando `system("git push origin")`. De esta manera, el informe completo, junto con otros archivos, están disponibles en el siguiente repositorio de GitHub: https://github.com/PaschalCO/Ogbogu-Emeghalu-Paschal-PEC1.