PEC 1 - Análisis de datos ómicos

Virginia Sestelo Prado

2024-11-05

Actividad 1

Seleccionar un dataset de metabolómica.

He elegido el dataset del repositorio de github denominado "2024-fobitools-UseCase_1". En la descripción se indica que se trata del estudio con identificador ST000291 del repositorio metabolomicsWorkbench. El estudio se titula "LC-MS Based Approaches to Investigate Metabolomic Differences in the Urine of Young Women after Drinking Cranberry Juice or Apple Juice" y tiene como objetivo investigar los cambios metabólicos generales causados por los concentrados de proantocianidinas de los arándanos y las manzanas en muestras de orina de mujeres jóvenes saludables.

En el repositorio de github se encuentran los archivos de tanto los datos como de los metadatos con la información acerca de las filas y las columnas del dataset:

- Archivo de datos (features.csv): contiene los datos de los metabolitos (1541 variables) para cada una de las 45 muestras (tratamientos).
- Archivo de metadatos (metadata.csv): contiene la información de las columnas: identificador y nombre de cada tratamiento.
- Archivo de nombres de metabolitos (metaboliteNames.csv): contiene información de las filas: los nombres y los identificadores de PubChem y KEGG de los metabolitos.

Como no está disponible un archivo con la información del experimento, lo creé yo misma a partir de la información que se encuentra en metabolomicsWorkbench y lo llamé experimental_metadata.txt.

Se descargaron y se importaron todos los archivos a R:

```r
library(readr)

# Archivo de metadatos
metadata <- read_delim("C:/Users/virse/Documents/BIOINFORMÁTICA Y BIOESTADÍSTICA/SEGUNDO SEMESTRE/ANÁLIS
    delim = ";", escape_double = FALSE, trim_ws = TRUE)
```

```
## New names:
## Rows: 45 Columns: 3
## -- Column specification
## -------------------------------------------------------- Delimiter: ";" chr
## (2): ID, Treatment dbl (1): ...1
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```r
View(metadata)

# Archivo de datos
features <- read_delim("C:/Users/virse/Documents/BIOINFORMÁTICA Y BIOESTADÍSTICA/SEGUNDO SEMESTRE/ANÁLIS
    delim = ";", escape_double = FALSE, col_types = cols(b1 = col_number()),
    trim_ws = TRUE)
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```r
View(features)

# Archivo de los nombres de los metabolitos
metaboliteNames <- read_delim("C:/Users/virse/Documents/BIOINFORMÁTICA Y BIOESTADÍSTICA/SEGUNDO SEMESTRE
    delim = ";", escape_double = FALSE, trim_ws = TRUE)
```

```
## New names:
## Rows: 1541 Columns: 4
## -- Column specification
## -------------------------------------------------------- Delimiter: ";" chr
## (3): names, PubChem, KEGG dbl (1): ...1
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```r
View(metaboliteNames)

# Archivo de la información del experimento
experiment_metadata <- read_delim("C:/Users/virse/Documents/BIOINFORMÁTICA Y BIOESTADÍSTICA/SEGUNDO SEME
    delim = "\t", escape_double = FALSE,
    trim_ws = TRUE)
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 128 Columns: 1
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): #METABOLOMICS WORKBENCH amitch_20151211_9581341_mwtab.txt DATATRACK...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
View(experiment_metadata)
```

Actividad 2

Una vez descargados los datos cread un contenedor del tipo SummarizedExperiment que contenga los datos y los metadatos (información acerca del dataset, las filas y las columnas).

```r
library(SummarizedExperiment)
```

```
## Cargando paquete requerido: MatrixGenerics

## Cargando paquete requerido: matrixStats

##
## Adjuntando el paquete: '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

## Cargando paquete requerido: GenomicRanges

## Cargando paquete requerido: stats4

## Cargando paquete requerido: BiocGenerics

##
## Adjuntando el paquete: '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
```

```
## Cargando paquete requerido: S4Vectors

##
## Adjuntando el paquete: 'S4Vectors'

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

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

## Cargando paquete requerido: IRanges

##
## Adjuntando el paquete: 'IRanges'

## The following object is masked from 'package:grDevices':
##
##     windows

## Cargando paquete requerido: GenomeInfoDb

## Cargando paquete requerido: Biobase

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

##
## Adjuntando el paquete: 'Biobase'

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

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

```r
# Se crea el data frame de los datos de las filas
rowData <- DataFrame(metaboliteNames[2:4], row.names = rownames(features))

# Se crea el data frame de los datos de las columnas
colData <- DataFrame(metadata[2:3])

# Se crea el contenedor SummarizedExperiment
se <- SummarizedExperiment(assays = list(counts = as.matrix(features)),
                           rowData = rowData,
                           colData = colData,
                           metadata = experiment_metadata)
se
```

```
## class: SummarizedExperiment
## dim: 1541 45
## metadata(1): #METABOLOMICS WORKBENCH amitch_20151211_9581341_mwtab.txt
##   DATATRACK_ID:450 efahy_20151227_122651 STUDY_ID:ST000291
##   ANALYSIS_ID:AN000464 PROJECT_ID:PR000233
## assays(1): counts
## rownames(1541): 1 2 ... 1540 1541
## rowData names(3): names PubChem KEGG
## colnames(45): b1 b10 ... c8 c9
## colData names(2): ID Treatment
```

Actividad 3

Llevad a cabo una exploración del dataset que os proporcione una visión general del mismo en la línea de lo que hemos visto en las actividades.

```r
# Matriz de los datos
matriz <- assays(se)[[1]]

# Ver las primeras filas y columnas de la matriz
head(matriz)
```

```
##         b1     b10     b11      b12    b13     b14     b15     b16     b17
## 1   443489  941000  757000   612000 858000  185000  671000 1140000  108000
## 2   107754 8300000 6790000 20800000 320000 1290000 1580000 2340000 1180000
## 3  9543071    1500     890 16200000   1250     968     657     809     767
## 4 11011465  276000   35700   631000 369000  242000  472000 5320000   18000
## 5  5281160  706000  121000 11600000 164000  424000  749000  267000 3050000
## 6   440341    6340   34100    31900   9440   92600    6740   14400    8180
##        b2       b4      b6     b7      b8      b9     a1     a10     a11
## 1  383000   593000 7240000 494000  812000 1290000  66000  215000  310000
## 2 1260000 15000000  495000  58100 1350000 1860000 698000 1220000 6920000
## 3     826     2810    1140   1010     635    1280    664     644    1060
## 4  243000   131000  158000 208000  228000  119000  58000   17700  394000
## 5   99100   136000  452000  75600  132000  341000 119000   51600   54900
## 6    8980     4610   10100   8180    5920    1950   1810    4350    1450
##        a12     a13     a14     a15     a16    a17     a2      a4       a6
## 1   798000 1070000  228000  241000 1180000  15100 255000  411000   463000
## 2 18700000 1320000 1230000 1980000 6980000 716000 761000 2910000 11300000
## 3     1500       0     818     660     754    695    562     851      766
## 4  4230000 3740000  361000   63300 2090000  37400  13400  260000   347000
## 5 26700000  323000  152000  208000 1400000  76100  16500  374000   491000
## 6        0   56200    3700    8360    2590   1390   1090   30400     2340
##       a7      a8      a9     c1     c10     c11     c12    c13     c14      c15
## 1 242000 1010000  702000  44600  136000 1060000 1050000 464000 1460000   636000
## 2 689000 1350000 1130000 479000  652000 2200000 7380000 187000 1430000  9730000
## 3    637     846       0    618     546     926  800000      0       0      809
## 4 151000 1080000    5080   2140  266000  627000 3140000 127000  197000   286000
## 5  81200 1000000 7000000  22400 1500000  171000  331000 198000  110000   178000
## 6   2930    7060    2830      0       0    4460       0   3190   22900 49200000
##        c16     c17      c2       c4     c6     c7      c8      c9
## 1  4510000  146000  400000   783000 213000 816000  587000  319000
## 2 11200000 6660000 1830000 15100000 971000 574000 4590000 9730000
```

```
## 3     1380     982     625     1790    626    991    1600     949
## 4   545000   35800   23200   230000  59600  48100   44000  576000
## 5   791000   44100   57100   150000  29500 126000  646000  291000
## 6        0    6930    1730     2400   3450   2880    2450   11200
```

```r
# Resumen estadístico de cada tratamiento
summary(matriz)
```

```
##        b1                 b10                  b11                  b12
##  Min.   :      16   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  1st Qu.:   10664   1st Qu.:1.235e+05   1st Qu.:8.145e+04   1st Qu.:2.240e+05
##  Median :  151126   Median :9.100e+05   Median :7.200e+05   Median :1.450e+06
##  Mean   : 4611998   Mean   :3.245e+07   Mean   :2.800e+07   Mean   :4.107e+07
##  3rd Qu.: 4772624   3rd Qu.:4.980e+06   3rd Qu.:4.500e+06   3rd Qu.:8.050e+06
##  Max.   :92042784   Max.   :1.920e+10   Max.   :1.550e+10   Max.   :2.070e+10
##  NA's   :1          NA's   :182         NA's   :182         NA's   :182
##       b13                  b14                  b15
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  1st Qu.:1.390e+05   1st Qu.:5.810e+04   1st Qu.:4.435e+04
##  Median :1.010e+06   Median :5.380e+05   Median :4.890e+05
##  Mean   :3.606e+07   Mean   :2.452e+07   Mean   :2.227e+07
##  3rd Qu.:5.545e+06   3rd Qu.:3.095e+06   3rd Qu.:2.925e+06
##  Max.   :2.300e+10   Max.   :1.130e+10   Max.   :1.240e+10
##  NA's   :182         NA's   :182         NA's   :182
##       b16                  b17                  b2
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  1st Qu.:1.070e+05   1st Qu.:2.560e+04   1st Qu.:2.635e+04
##  Median :8.560e+05   Median :2.940e+05   Median :2.910e+05
##  Mean   :3.086e+07   Mean   :1.551e+07   Mean   :1.270e+07
##  3rd Qu.:4.920e+06   3rd Qu.:1.940e+06   3rd Qu.:1.675e+06
##  Max.   :1.650e+10   Max.   :7.320e+09   Max.   :6.810e+09
##  NA's   :182         NA's   :182         NA's   :182
##        b4                   b6                   b7
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  1st Qu.:1.565e+05   1st Qu.:1.620e+05   1st Qu.:3.615e+04
##  Median :9.160e+05   Median :1.250e+06   Median :3.740e+05
##  Mean   :3.124e+07   Mean   :4.695e+07   Mean   :2.116e+07
##  3rd Qu.:5.070e+06   3rd Qu.:7.130e+06   3rd Qu.:2.830e+06
##  Max.   :1.700e+10   Max.   :2.610e+10   Max.   :1.270e+10
##  NA's   :182         NA's   :182         NA's   :182
##        b8                   b9                   a1
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  1st Qu.:4.225e+04   1st Qu.:1.435e+05   1st Qu.:1.520e+04
##  Median :4.540e+05   Median :1.070e+06   Median :2.090e+05
##  Mean   :1.688e+07   Mean   :2.841e+07   Mean   :1.288e+07
##  3rd Qu.:2.710e+06   3rd Qu.:5.615e+06   3rd Qu.:1.475e+06
##  Max.   :7.870e+09   Max.   :1.430e+10   Max.   :6.970e+09
##  NA's   :182         NA's   :182         NA's   :182
##       a10                  a11                  a12
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  1st Qu.:3.375e+04   1st Qu.:3.025e+04   1st Qu.:1.480e+05
##  Median :3.000e+05   Median :2.860e+05   Median :1.180e+06
##  Mean   :1.371e+07   Mean   :2.108e+07   Mean   :3.527e+07
##  3rd Qu.:2.005e+06   3rd Qu.:2.475e+06   3rd Qu.:5.895e+06
```

```
##  Max.   :6.050e+09   Max.   :1.210e+10   Max.   :2.190e+10
##  NA's   :182         NA's   :182         NA's   :182
##       a13                 a14                 a15
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  1st Qu.:1.430e+05   1st Qu.:5.950e+04   1st Qu.:9.860e+03
##  Median :1.030e+06   Median :5.350e+05   Median :1.640e+05
##  Mean   :4.372e+07   Mean   :2.336e+07   Mean   :1.336e+07
##  3rd Qu.:7.840e+06   3rd Qu.:3.155e+06   3rd Qu.:1.400e+06
##  Max.   :2.780e+10   Max.   :1.230e+10   Max.   :6.570e+09
##  NA's   :182         NA's   :182         NA's   :182
##       a16                 a17                  a2
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  1st Qu.:7.320e+04   1st Qu.:2.300e+04   1st Qu.:1.085e+04
##  Median :5.380e+05   Median :2.790e+05   Median :1.520e+05
##  Mean   :2.305e+07   Mean   :1.758e+07   Mean   :1.145e+07
##  3rd Qu.:3.230e+06   3rd Qu.:2.150e+06   3rd Qu.:1.475e+06
##  Max.   :1.430e+10   Max.   :9.770e+09   Max.   :5.330e+09
##  NA's   :182         NA's   :182         NA's   :182
##        a4                  a6                  a7
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  1st Qu.:1.670e+05   1st Qu.:9.705e+04   1st Qu.:1.870e+04
##  Median :1.150e+06   Median :7.080e+05   Median :2.420e+05
##  Mean   :3.704e+07   Mean   :2.767e+07   Mean   :1.619e+07
##  3rd Qu.:6.380e+06   3rd Qu.:4.140e+06   3rd Qu.:2.215e+06
##  Max.   :2.090e+10   Max.   :1.530e+10   Max.   :8.930e+09
##  NA's   :182         NA's   :182         NA's   :182
##        a8                  a9                  c1
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  1st Qu.:7.105e+04   1st Qu.:1.005e+05   1st Qu.:1.135e+04
##  Median :6.410e+05   Median :8.590e+05   Median :1.740e+05
##  Mean   :2.039e+07   Mean   :3.144e+07   Mean   :1.317e+07
##  3rd Qu.:3.355e+06   3rd Qu.:4.900e+06   3rd Qu.:1.470e+06
##  Max.   :9.710e+09   Max.   :1.640e+10   Max.   :5.900e+09
##  NA's   :182         NA's   :182         NA's   :182
##       c10                 c11                 c12
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  1st Qu.:4.840e+04   1st Qu.:6.490e+04   1st Qu.:1.305e+05
##  Median :4.920e+05   Median :5.980e+05   Median :1.050e+06
##  Mean   :2.021e+07   Mean   :2.269e+07   Mean   :4.042e+07
##  3rd Qu.:2.680e+06   3rd Qu.:3.880e+06   3rd Qu.:6.120e+06
##  Max.   :1.050e+10   Max.   :1.240e+10   Max.   :2.040e+10
##  NA's   :182         NA's   :182         NA's   :182
##       c13                 c14                 c15
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  1st Qu.:1.625e+05   1st Qu.:6.530e+04   1st Qu.:8.165e+04
##  Median :1.120e+06   Median :5.770e+05   Median :7.260e+05
##  Mean   :5.147e+07   Mean   :2.468e+07   Mean   :2.447e+07
##  3rd Qu.:7.135e+06   3rd Qu.:3.415e+06   3rd Qu.:3.810e+06
##  Max.   :2.380e+10   Max.   :1.120e+10   Max.   :9.930e+09
##  NA's   :182         NA's   :182         NA's   :182
##       c16                 c17                  c2
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00
##  1st Qu.:1.150e+05   1st Qu.:4.185e+04   1st Qu.:2.320e+04
##  Median :9.290e+05   Median :4.350e+05   Median :2.700e+05
```

```
##  Mean   :4.430e+07   Mean   :2.168e+07   Mean   :1.682e+07
##  3rd Qu.:6.195e+06   3rd Qu.:2.655e+06   3rd Qu.:2.055e+06
##  Max.   :2.140e+10   Max.   :7.550e+09   Max.   :5.920e+09
##  NA's   :182         NA's   :182         NA's   :182
##        c4                  c6                  c7                  c8
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.000e+00   Min.   :0.00e+00
##  1st Qu.:1.095e+05   1st Qu.:4.110e+04   1st Qu.:1.180e+05   1st Qu.:2.31e+04
##  Median :8.590e+05   Median :5.200e+05   Median :8.790e+05   Median :3.06e+05
##  Mean   :4.522e+07   Mean   :1.774e+07   Mean   :3.881e+07   Mean   :2.12e+07
##  3rd Qu.:5.310e+06   3rd Qu.:2.865e+06   3rd Qu.:5.245e+06   3rd Qu.:2.35e+06
##  Max.   :2.110e+10   Max.   :1.060e+10   Max.   :1.710e+10   Max.   :9.75e+09
##  NA's   :182         NA's   :182         NA's   :182         NA's   :182
##        c9
##  Min.   :0.000e+00
##  1st Qu.:8.745e+04
##  Median :7.260e+05
##  Mean   :2.937e+07
##  3rd Qu.:4.345e+06
##  Max.   :1.480e+10
##  NA's   :182
```

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

```
## DataFrame with 45 rows and 2 columns
##              ID   Treatment
##     <character> <character>
## b1           b1    Baseline
## b10         b10    Baseline
## b11         b11    Baseline
## b12         b12    Baseline
## b13         b13    Baseline
## ...         ...         ...
## c4           c4   Cranberry
## c6           c6   Cranberry
## c7           c7   Cranberry
## c8           c8   Cranberry
## c9           c9   Cranberry
```

```r
# Metadatos de las filas (información de los atributos)
rowData(se)
```

```
## DataFrame with 1541 rows and 3 columns
##                        names     PubChem        KEGG
##                  <character> <character> <character>
## 1    10-Desacetyltaxuyunn..     5460449      C15538
## 2    10-Hydroxydecanoic a..       74300      C02774
## 3         10-Oxodecanoate_1    19734156      C02217
## 4    11beta,21-Dihydroxy-..    21145110      C05475
## 5    1,1-Dichloroethylene..      119521      C14857
## ...                     ...         ...         ...
## 1537             Ungeremine      159646      C12189
## 1538           Valacyclovir       60773      C07184
```

```
## 1539             Versiconal    25203618      C20507
## 1540          Zizybeoside I    11972301      C17564
## 1541           Zoxazolamine        6103      C13841
```

```
# Metadatos del experimento
metadata(se)
```

```
## $`#METABOLOMICS WORKBENCH amitch_20151211_9581341_mwtab.txt DATATRACK_ID:450 efahy_20151227_122651 S
##   [1] "VERSION             \t1"
##   [2] "CREATED_ON             \tDecember 27, 2015, 12:26 pm"
##   [3] "#PROJECT"
##   [4] "PR:PROJECT_TITLE                 \tLC-MS Based Approaches to Investigate Metabolomic Differen
##   [5] "PR:PROJECT_TITLE                 \tPlasma of Young Women after Drinking Cranberry Juice or Ap
##   [6] "PR:PROJECT_SUMMARY               \tThe present study aimed to investigate overall metabolic c
##   [7] "PR:PROJECT_SUMMARY               \tcranberry juice or apple juice consumption using a global 
MS based"
##   [8] "PR:PROJECT_SUMMARY               \tmetabolomics approach."
##   [9] "PR:INSTITUTE                     \tUniversity of Florida"
##  [10] "PR:DEPARTMENT                    \tFood Science and Nutrition"
##  [11] "PR:LABORATORY                    \tGu"
##  [12] "PR:LAST_NAME                     \tLiu"
##  [13] "PR:FIRST_NAME                    \tHaiyan"
##  [14] "PR:ADDRESS                       \t--"
##  [15] "PR:EMAIL                         \thaiyan66@ufl.edu"
##  [16] "PR:PHONE                         \t352-392-1991x210"
##  [17] "#STUDY"
##  [18] "ST:STUDY_TITLE                   \tLC-MS Based Approaches to Investigate Metabolomic Differen
##  [19] "ST:STUDY_TITLE                   \tYoung Women after Drinking Cranberry Juice or Apple Juice"
##  [20] "ST:STUDY_TYPE                    \tdrug dosage"
##  [21] "ST:STUDY_SUMMARY                 \tEighteen healthy female college students between 21-
29 years old with a normal"
##  [22] "ST:STUDY_SUMMARY                 \tBMI of 18.5-25 were recruited. Each subject was provided w
##  [23] "ST:STUDY_SUMMARY                 \tthat contained significant amount of procyanidins, such as
##  [24] "ST:STUDY_SUMMARY                 \tgrapes, blueberries, chocolate and plums. They were advise
##  [25] "ST:STUDY_SUMMARY                 \tduring the 1-6th day and the rest of the study. On the mor
##  [26] "ST:STUDY_SUMMARY                 \tfirst-morning baseline urine sample and blood sample were 
##  [27] "ST:STUDY_SUMMARY                 \thuman subjects after overnight fasting. Participants were 
##  [28] "ST:STUDY_SUMMARY                 \tallocated into two groups (n=9) to consume cranberry juice
##  [29] "ST:STUDY_SUMMARY                 \tbottles (250 ml/bottle) of juice were given to participant
##  [30] "ST:STUDY_SUMMARY                 \tmorning and evening of the 7th, 8th, and 9th day. On the m
##  [31] "ST:STUDY_SUMMARY                 \tall subjects returned to the clinical unit to provide a fi
morning urine"
##  [32] "ST:STUDY_SUMMARY                 \tsample after overnight fasting. The blood sample was also 
##  [33] "ST:STUDY_SUMMARY                 \tparticipants 30 min later after they drank another bottle 
##  [34] "ST:STUDY_SUMMARY                 \tmorning. After two-weeks of wash out period, participants 
##  [35] "ST:STUDY_SUMMARY                 \talternative regimen and repeated the protocol. One human s
##  [36] "ST:STUDY_SUMMARY                 \tthis study because she missed part of her appointments. An
##  [37] "ST:STUDY_SUMMARY                 \tsubjects were removed from urine metabolomics analyses bec
##  [38] "ST:STUDY_SUMMARY                 \tprovide required urine samples after juice drinking.The pr
##  [39] "ST:STUDY_SUMMARY                 \tinvestigate overall metabolic changes caused by procyanidi
##  [40] "ST:STUDY_SUMMARY                 \tcranberries and apples using a global LCMS based metabolom
##  [41] "ST:STUDY_SUMMARY                 \tplasma and urine samples were stored at -
80oC until analysis."
##  [42] "ST:INSTITUTE                     \tUniversity of Florida"
```

```
## [43] "ST:DEPARTMENT                    \tSECIM"
## [44] "ST:LABORATORY                    \tGu"
## [45] "ST:LAST_NAME                     \tLiu"
## [46] "ST:FIRST_NAME                    \tHaiyan"
## [47] "ST:ADDRESS                       \t--"
## [48] "ST:EMAIL                         \thaiyan66@ufl.edu"
## [49] "ST:PHONE                         \t352-392-1991x210"
## [50] "ST:NUM_GROUPS                    \t3"
## [51] "ST:TOTAL_SUBJECTS                \t45"
## [52] "#SUBJECT"
## [53] "SU:SUBJECT_TYPE                  \tHuman"
## [54] "SU:SUBJECT_SPECIES               \tHomo sapiens"
## [55] "SU:SUBJECT_COMMENTS              \tGu_subjects_human urine.txt"
## [56] "#SUBJECT_SAMPLE_FACTORS:         \tSUBJECT(optional)[tab]SAMPLE[tab]FACTORS(NAME:VALUE pairs s
## [57] "SUBJECT_SAMPLE_FACTORS           \t-\tb1\tTreatment :Baseline urine"
## [58] "SUBJECT_SAMPLE_FACTORS           \t-\tb2\tTreatment :Baseline urine"
## [59] "SUBJECT_SAMPLE_FACTORS           \t-\tb4\tTreatment :Baseline urine"
## [60] "SUBJECT_SAMPLE_FACTORS           \t-\tb6\tTreatment :Baseline urine"
## [61] "SUBJECT_SAMPLE_FACTORS           \t-\tb7\tTreatment :Baseline urine"
## [62] "SUBJECT_SAMPLE_FACTORS           \t-\tb8\tTreatment :Baseline urine"
## [63] "SUBJECT_SAMPLE_FACTORS           \t-\tb9\tTreatment :Baseline urine"
## [64] "SUBJECT_SAMPLE_FACTORS           \t-\tb10\tTreatment :Baseline urine"
## [65] "SUBJECT_SAMPLE_FACTORS           \t-\tb11\tTreatment :Baseline urine"
## [66] "SUBJECT_SAMPLE_FACTORS           \t-\tb12\tTreatment :Baseline urine"
## [67] "SUBJECT_SAMPLE_FACTORS           \t-\tb13\tTreatment :Baseline urine"
## [68] "SUBJECT_SAMPLE_FACTORS           \t-\tb14\tTreatment :Baseline urine"
## [69] "SUBJECT_SAMPLE_FACTORS           \t-\tb15\tTreatment :Baseline urine"
## [70] "SUBJECT_SAMPLE_FACTORS           \t-\tb16\tTreatment :Baseline urine"
## [71] "SUBJECT_SAMPLE_FACTORS           \t-\tb17\tTreatment :Baseline urine"
## [72] "SUBJECT_SAMPLE_FACTORS           \t-\tc1\tTreatment :Urine after drinking cranberry juice"
## [73] "SUBJECT_SAMPLE_FACTORS           \t-\tc2\tTreatment :Urine after drinking cranberry juice"
## [74] "SUBJECT_SAMPLE_FACTORS           \t-\tc4\tTreatment :Urine after drinking cranberry juice"
## [75] "SUBJECT_SAMPLE_FACTORS           \t-\tc6\tTreatment :Urine after drinking cranberry juice"
## [76] "SUBJECT_SAMPLE_FACTORS           \t-\tc7\tTreatment :Urine after drinking cranberry juice"
## [77] "SUBJECT_SAMPLE_FACTORS           \t-\tc8\tTreatment :Urine after drinking cranberry juice"
## [78] "SUBJECT_SAMPLE_FACTORS           \t-\tc9\tTreatment :Urine after drinking cranberry juice"
## [79] "SUBJECT_SAMPLE_FACTORS           \t-\tc10\tTreatment :Urine after drinking cranberry juice"
## [80] "SUBJECT_SAMPLE_FACTORS           \t-\tc11\tTreatment :Urine after drinking cranberry juice"
## [81] "SUBJECT_SAMPLE_FACTORS           \t-\tc12\tTreatment :Urine after drinking cranberry juice"
## [82] "SUBJECT_SAMPLE_FACTORS           \t-\tc13\tTreatment :Urine after drinking cranberry juice"
## [83] "SUBJECT_SAMPLE_FACTORS           \t-\tc14\tTreatment :Urine after drinking cranberry juice"
## [84] "SUBJECT_SAMPLE_FACTORS           \t-\tc15\tTreatment :Urine after drinking cranberry juice"
## [85] "SUBJECT_SAMPLE_FACTORS           \t-\tc16\tTreatment :Urine after drinking cranberry juice"
## [86] "SUBJECT_SAMPLE_FACTORS           \t-\tc17\tTreatment :Urine after drinking cranberry juice"
## [87] "SUBJECT_SAMPLE_FACTORS           \t-\ta1\tTreatment :Urine after drinking apple juice"
## [88] "SUBJECT_SAMPLE_FACTORS           \t-\ta2\tTreatment :Urine after drinking apple juice"
## [89] "SUBJECT_SAMPLE_FACTORS           \t-\ta4\tTreatment :Urine after drinking apple juice"
## [90] "SUBJECT_SAMPLE_FACTORS           \t-\ta6\tTreatment :Urine after drinking apple juice"
## [91] "SUBJECT_SAMPLE_FACTORS           \t-\ta7\tTreatment :Urine after drinking apple juice"
## [92] "SUBJECT_SAMPLE_FACTORS           \t-\ta8\tTreatment :Urine after drinking apple juice"
## [93] "SUBJECT_SAMPLE_FACTORS           \t-\ta9\tTreatment :Urine after drinking apple juice"
## [94] "SUBJECT_SAMPLE_FACTORS           \t-\ta10\tTreatment :Urine after drinking apple juice"
## [95] "SUBJECT_SAMPLE_FACTORS           \t-\ta11\tTreatment :Urine after drinking apple juice"
## [96] "SUBJECT_SAMPLE_FACTORS           \t-\ta12\tTreatment :Urine after drinking apple juice"
```

```
##  [97] "SUBJECT_SAMPLE_FACTORS           \t-\ta13\tTreatment :Urine after drinking apple juice"
##  [98] "SUBJECT_SAMPLE_FACTORS           \t-\ta14\tTreatment :Urine after drinking apple juice"
##  [99] "SUBJECT_SAMPLE_FACTORS           \t-\ta15\tTreatment :Urine after drinking apple juice"
## [100] "SUBJECT_SAMPLE_FACTORS           \t-\ta16\tTreatment :Urine after drinking apple juice"
## [101] "SUBJECT_SAMPLE_FACTORS           \t-\ta17\tTreatment :Urine after drinking apple juice"
## [102] "#COLLECTION"
## [103] "CO:COLLECTION_SUMMARY            \tn/a"
## [104] "CO:COLLECTION_PROTOCOL_FILENAME  \tGu_collection_human urine.txt"
## [105] "#TREATMENT"
## [106] "TR:TREATMENT_SUMMARY             \tn/a"
## [107] "TR:TREATMENT_PROTOCOL_FILENAME   \tGu_treatment_human urine.txt"
## [108] "#SAMPLEPREP"
## [109] "SP:SAMPLEPREP_SUMMARY            \tn/a"
## [110] "SP:SAMPLEPREP_PROTOCOL_FILENAME  \tMetabolomics_LCMSProtocol_urine.pdf"
## [111] "#CHROMATOGRAPHY"
## [112] "CH:CHROMATOGRAPHY_TYPE           \tReversed phase"
## [113] "CH:INSTRUMENT_NAME               \tThermo Scientific-Dionex Ultimate 3000"
## [114] "CH:COLUMN_NAME                   \tACE Excel 2 C18-PFP (100 x 2.1mm, 2um)"
## [115] "CH:METHODS_FILENAME              \tMetabolomics_LCMSProtocol_urine.pdf"
## [116] "CH:INTERNAL_STANDARD             \tAppendix A - Internal Standard Prep GLCMS.pdf"
## [117] "#ANALYSIS"
## [118] "AN:ANALYSIS_TYPE                 \tMS"
## [119] "#MS"
## [120] "MS:MS_COMMENTS                   \t-"
## [121] "MS:INSTRUMENT_TYPE               \tOrbitrap"
## [122] "MS:MS_TYPE                       \tESI"
## [123] "MS:ION_MODE                      \tPOSITIVE"
## [124] "MS:INSTRUMENT_NAME               \tThermo Q Exactive Orbitrap"
## [125] "MS:INSTRUMENT_NAME               \tThermo Scientific Q-Exactive"
## [126] "#MS_METABOLITE_DATA"
## [127] "MS_METABOLITE_DATA:UNITS         \tPeak area"
## [128] "MS_METABOLITE_DATA_START"
##
## attr(,"spec")
## cols(
##   `#METABOLOMICS WORKBENCH amitch_20151211_9581341_mwtab.txt DATATRACK_ID:450 efahy_20151227_122651 S
## )
## attr(,"problems")
## <pointer: 0x000002533baffb80>
```

```r
# Dimensiones del objeto SummarizedExperiment
dim(se)
```

```
## [1] 1541   45
```

```r
# Nombres de las columnas
colnames(se)
```

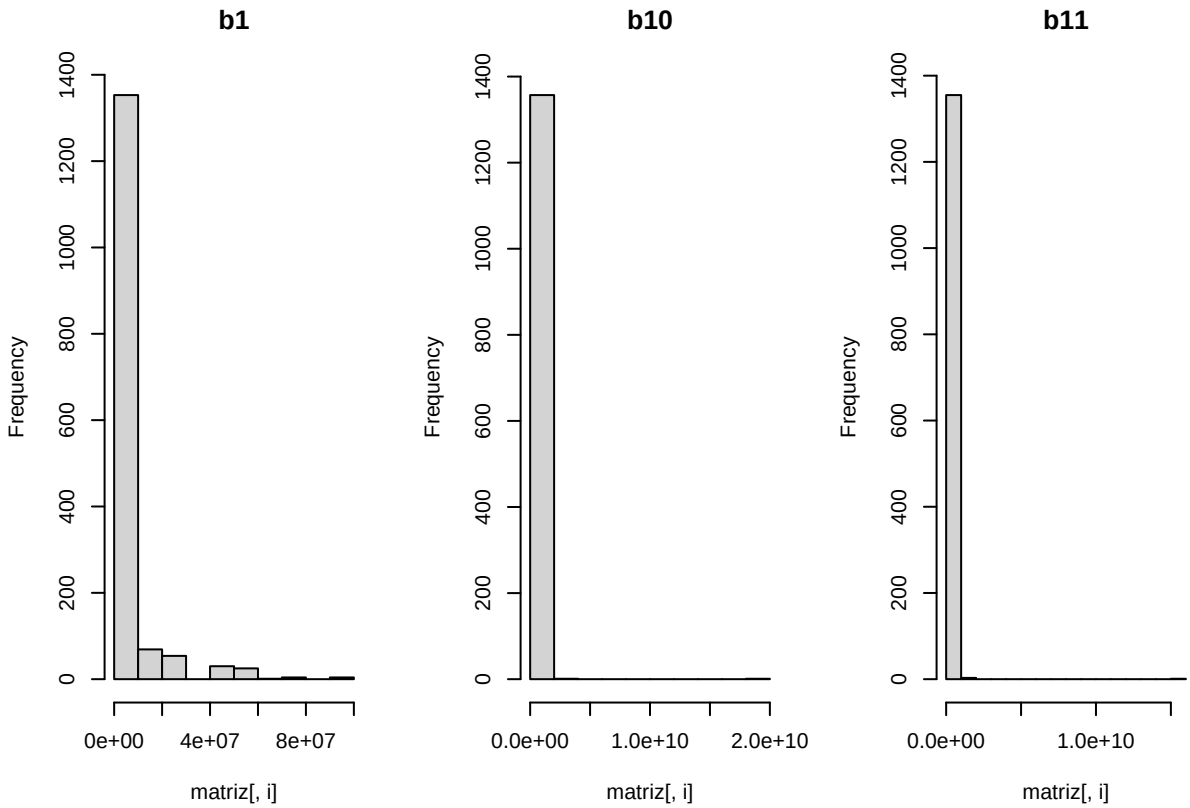
```
##  [1] "b1"  "b10" "b11" "b12" "b13" "b14" "b15" "b16" "b17" "b2"  "b4"  "b6"
## [13] "b7"  "b8"  "b9"  "a1"  "a10" "a11" "a12" "a13" "a14" "a15" "a16" "a17"
## [25] "a2"  "a4"  "a6"  "a7"  "a8"  "a9"  "c1"  "c10" "c11" "c12" "c13" "c14"
## [37] "c15" "c16" "c17" "c2"  "c4"  "c6"  "c7"  "c8"  "c9"
```

```r
# Nombres de las primeras 6 filas
head(rownames(se))
```

```
## [1] "1" "2" "3" "4" "5" "6"
```

Se pueden elaborar los histogramas de cada tratamiento, se representan los tres primeros:

```r
opt <- par(mfrow=c(1,3))
for (i in 1:3)
  hist(matriz[,i], main = colnames(matriz)[i])
```

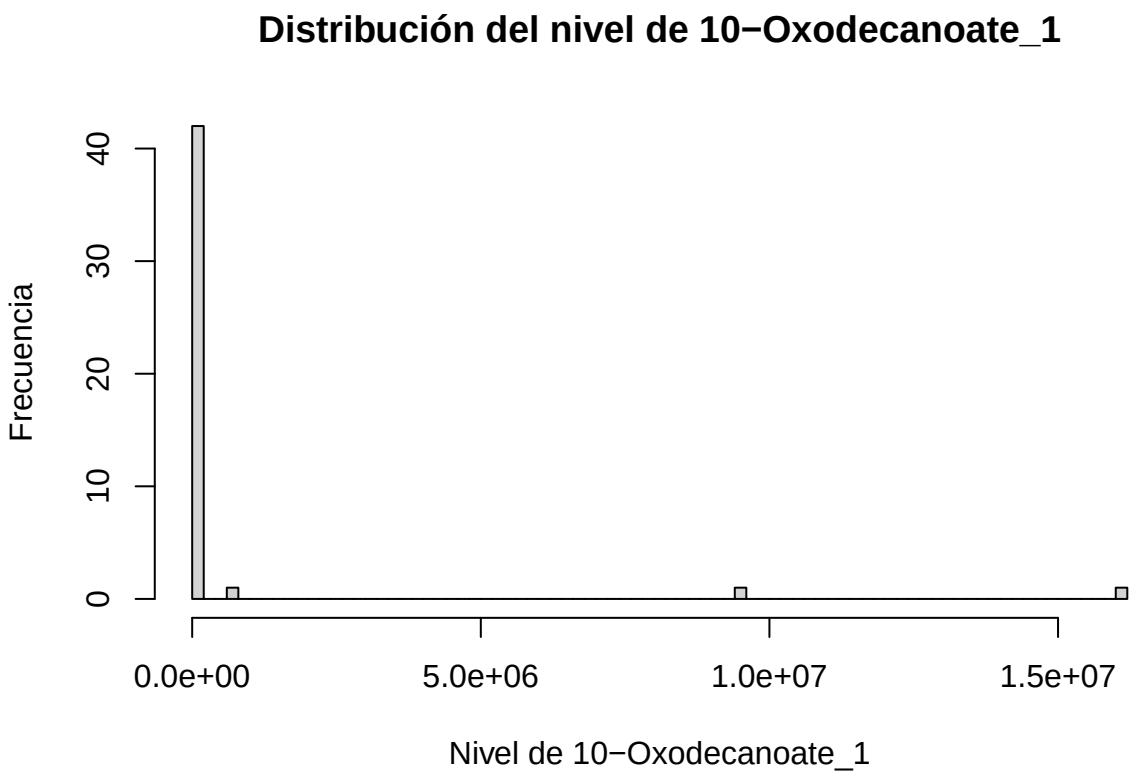



```r
par(opt)
```

Se observa que en el tratamiento b1 hay más metabolitos que presentan un nivel de señal alto, mientras que en b10 y b11 todos presentan niveles bajos.

Se prueba también a obtener, por ejemplo, la distribución de los niveles del tercer metabolito (10-Oxodecanoate_1).

```r
subse <- se[3,]
msubse <- assays(subse)[[1]]

hist(msubse, main="Distribución del nivel de 10-Oxodecanoate_1",
     xlab="Nivel de 10-Oxodecanoate_1",
     ylab="Frecuencia",
     breaks=100)
```



Se observa que en la mayoría de tratamientos el nivel de 10-Oxodecanoate_1 es bastante bajo excepto en dos casos. Para saber qué muestras son se usa la función which:

```
which(msubse > 5000000)
```

```
## [1] 1 4
```

Para ver qué tratamiento recibieron la primera y cuarta muestra vamos a los datos de las columnas:

```
colData(se)[1,]
```

```
## DataFrame with 1 row and 2 columns
##              ID   Treatment
##    <character> <character>
## b1          b1    Baseline
```

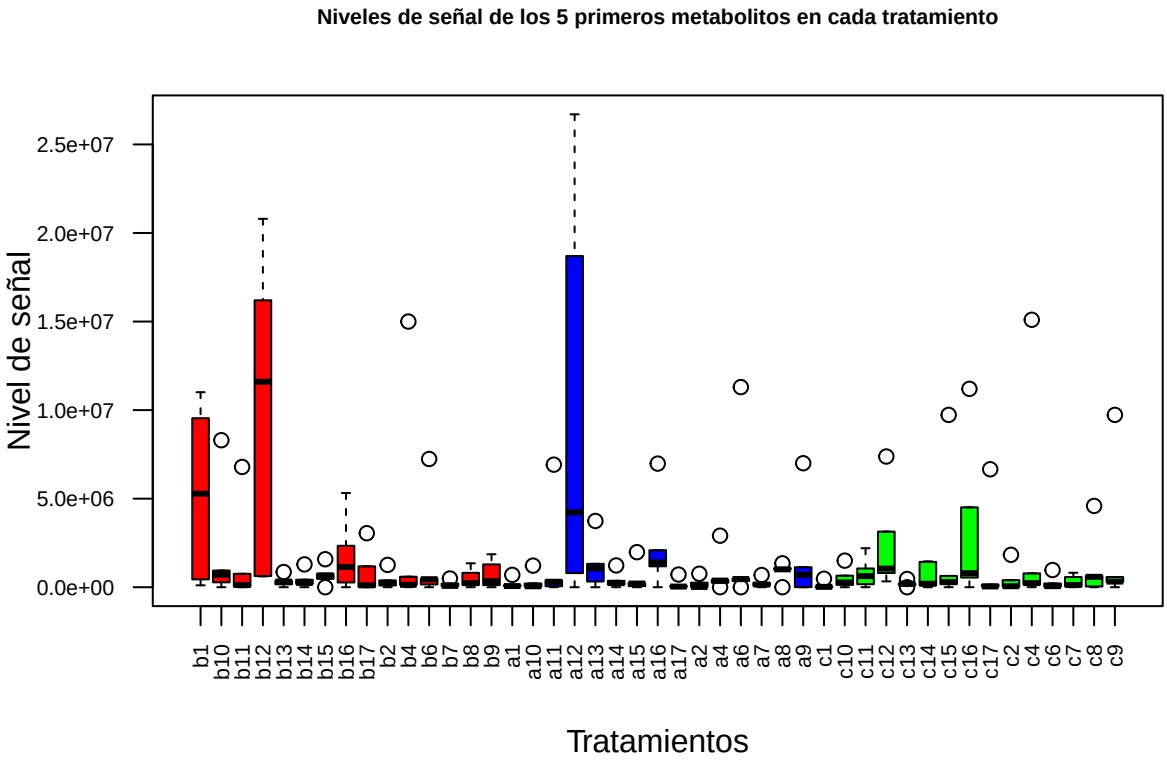
```
colData(se)[4,]
```

```
## DataFrame with 1 row and 2 columns
##               ID   Treatment
##     <character> <character>
## b12         b12    Baseline
```

Los dos son del grupo de baseline urine, es decir, pertenecen al grupo control, lo que a primera vista sugiere que ninguno de los tratamientos aumenta el nivel de 10-Oxodecanoate_1, si no que más bien podrían reducirlo.

Se pueden extraer los 5 primeros metabolitos y representar un diagrama de caja por cada tratamiento (controles en rojo, zumo de manzana en azul y de arándanos en verde).

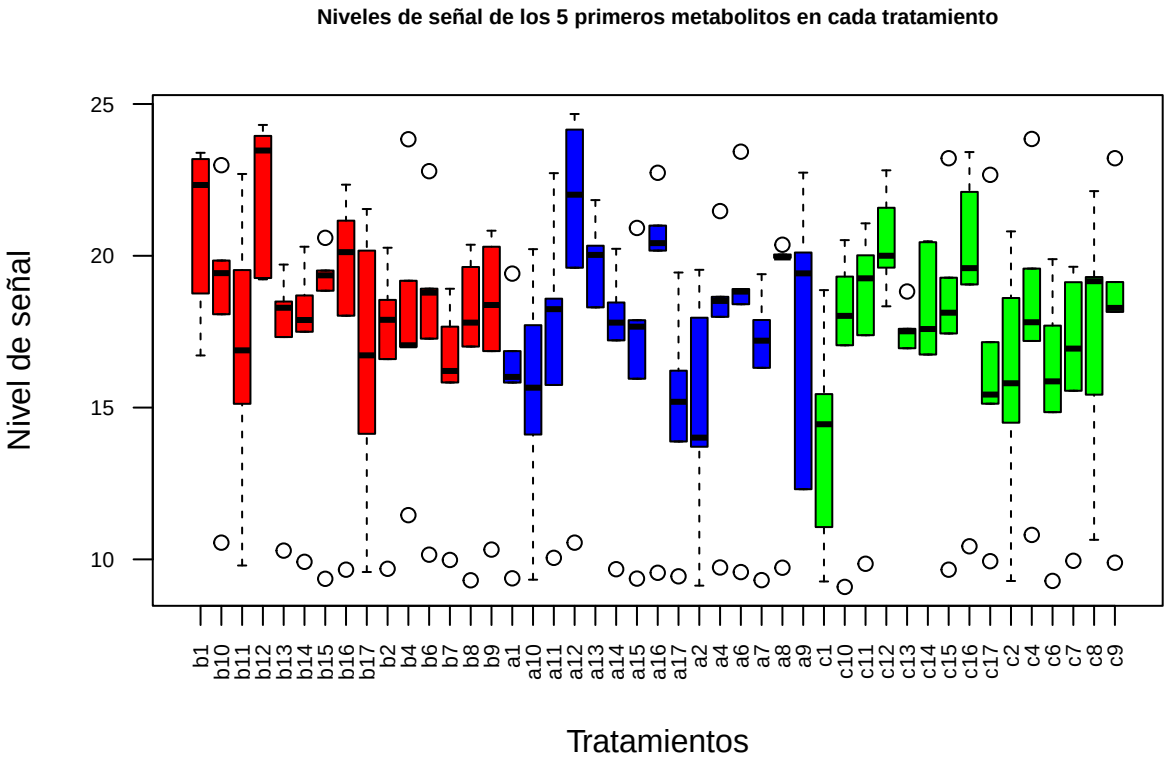
```
subse2 <- se[1:5,]
msubse2 <- assays(subse2)[[1]]
groupColors <- c(rep("red", 15), rep("blue", 15),rep("green", 15) )
boxplot(msubse2, col=groupColors, main="Niveles de señal de los 5 primeros metabolitos en cada tratamien
    xlab="Tratamientos",
    ylab="Nivel de señal", las=2, cex.axis=0.7, cex.main=0.7)
```



Así los datos no se interpretan bien, es mejor usar el logaritmo de los datos:

```
groupColors <- c(rep("red", 15), rep("blue", 15),rep("green", 15) )
boxplot(log2(msubse2), col=groupColors, main="Niveles de señal de los 5 primeros metabolitos en cada tra
    xlab="Tratamientos",
    ylab="Nivel de señal", las=2, cex.axis=0.7, cex.main=0.7)
```

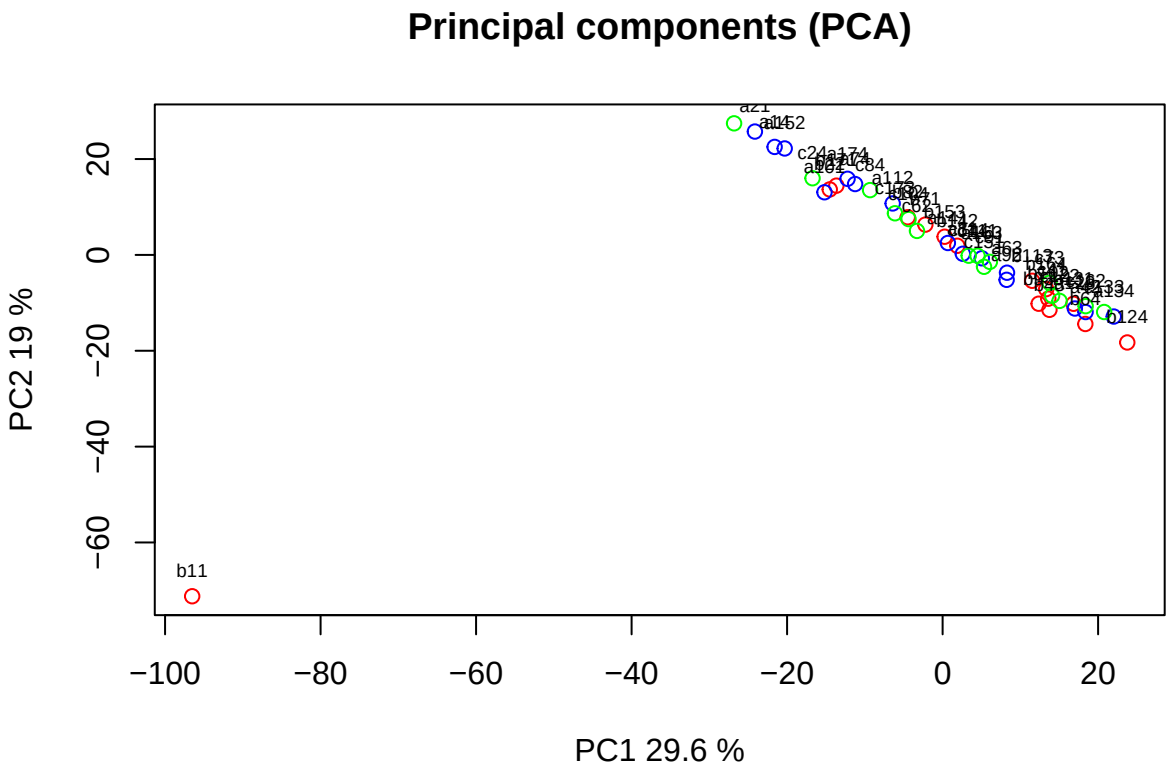
```
## Warning in bplt(at[i], wid = width[i], stats = z$stats[, i], out =
## z$out[z$group == : Outlier (-Inf) in boxplot 20 is not drawn

## Warning in bplt(at[i], wid = width[i], stats = z$stats[, i], out =
## z$out[z$group == : Outlier (-Inf) in boxplot 30 is not drawn

## Warning in bplt(at[i], wid = width[i], stats = z$stats[, i], out =
## z$out[z$group == : Outlier (-Inf) in boxplot 35 is not drawn

## Warning in bplt(at[i], wid = width[i], stats = z$stats[, i], out =
## z$out[z$group == : Outlier (-Inf) in boxplot 36 is not drawn
```



A simple vista, no parece que haya diferencias muy claras entre los niveles de señal de los metabolitos de los diferentes tratamientos.

Se elabora un análisis de componentes principales (PCA):

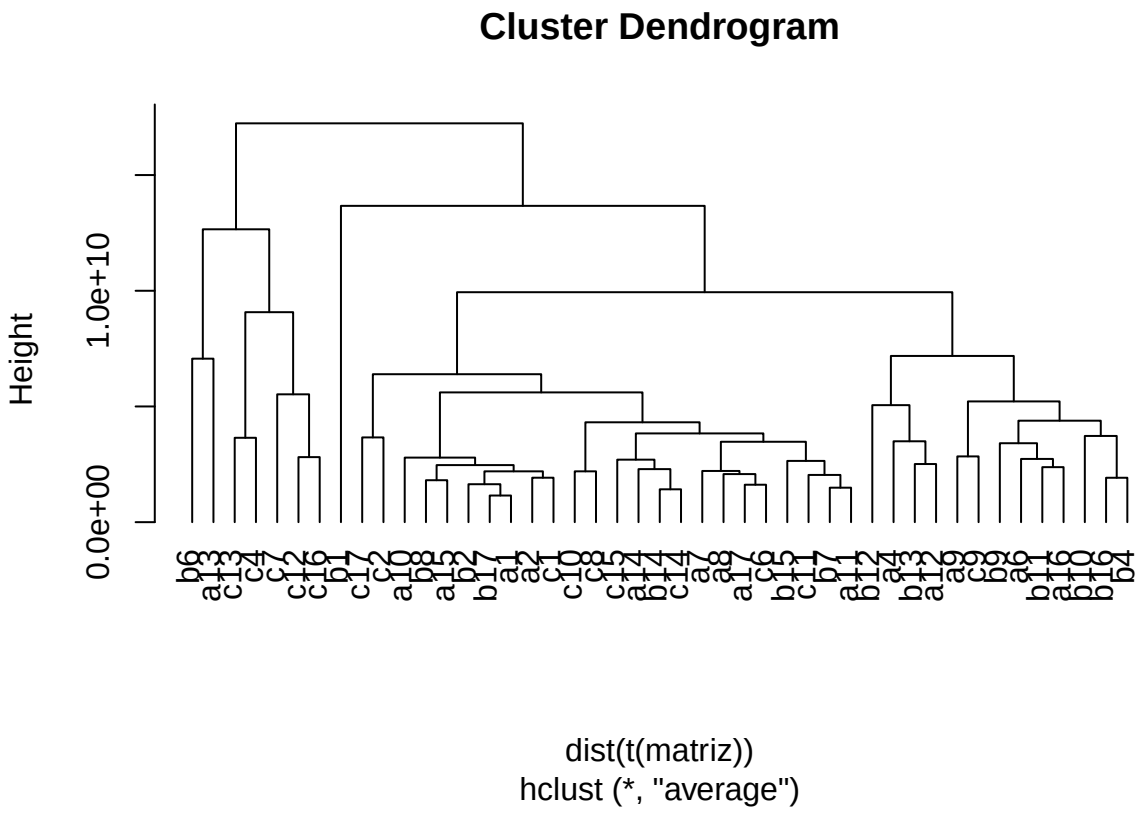
```
# Se transforma logarítmicamente la matriz (se suma 1 para evitar log(0))
logX <- log2(matriz+1)

# Se elimina cualquier fila/columna con NA tras la transformación
logX <- na.omit(logX)

# Se realiza el PCA en los datos transpuestos y escalados
pcX <- prcomp(t(logX), scale = TRUE)

# Se calcula la varianza explicada de cada componente
loads <- round(pcX$sdev^2 / sum(pcX$sdev^2) * 100, 1)

# Se representan los resultados de los dos primeros componentes
xlab<-c(paste("PC1",loads[1],"%"))
ylab<-c(paste("PC2",loads[2],"%"))
plot(pcX$x[,1:2],xlab=xlab,ylab=ylab, col=groupColors,
     main ="Principal components (PCA)")
names2plot<-paste0(substr(colnames(matriz),1,3), 1:4)

text(pcX$x[,1],pcX$x[,2],names2plot, pos=3, cex=.6)
```



El análisis de componentes principales (PCA) revela que los dos primeros componentes (PC1 y PC2) explican un 49.6% de la varianza total en los datos (29.6% para PC1 y 19% para PC2). Sin embargo, la distribución de las muestras en el gráfico de PCA muestra que todas las muestras, excepto una (podría ser un outlier), se agrupan de manera cercana, lo que sugiere una baja variabilidad entre las condiciones experimentales en los primeros componentes. Esto podría indicar que los metabolitos medidos no presentan una diferenciación clara entre los tratamientos con zumo de arándanos y de manzana en base a los componentes principales considerados.

También se lleva a cabo un clustering jerárquico para ver el agrupamiento de las muestras:

```
clust.euclid.average <- hclust(dist(t(matriz)),method="average")
plot(clust.euclid.average, hang=-1)
```



El clustering jerárquico revela una estructura compleja con múltiples ramificaciones. La agrupación de las muestras no separa claramente los tratamientos, lo que sugiere que no existen grandes diferencias en los perfiles metabólicos de estos. Esto puede indicar que los efectos de los tratamientos en los metabolitos no son suficientemente significativos.

Actividad 4

Elaborad un informe que describa el proceso que habéis realizado, incluyendo la descarga de los datos, la creación del contenedor, la exploración de los datos y la reposición de los datos en github. El nombre del repositorio tiene que ser el siguiente: APELLIDO1-Apellido2-Nombre-PEC1.

Es este documento.

Actividad 5

Cread un repositorio de github2 que contenga: el informe, el objeto contenedor con los datos y los metadatos en formato binario (.Rda), el código R para la exploración de los datos, los datos en formato texto y los metadatos acerca del dataset en un archivo markdown.

Se guarda el objeto contenedor en formato binario:

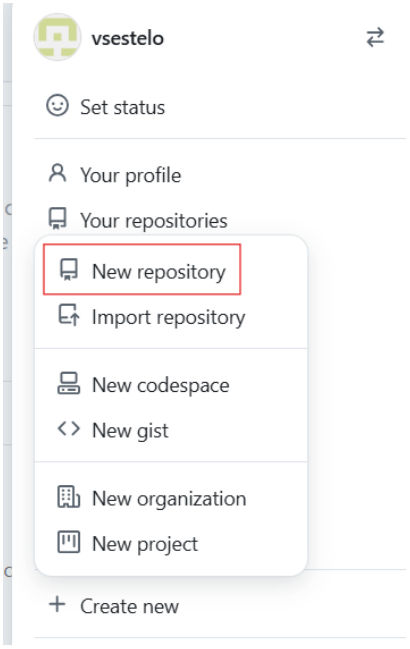
```
save(se, file = "C:/Users/virse/Documents/BIOINFORMÁTICA Y BIOESTADÍSTICA/SEGUNDO SEMESTRE/ANÁLISIS DE
```

Se copia el código empleado en la actividad 3 para la exploración de los datos en un nuevo archivo .R.

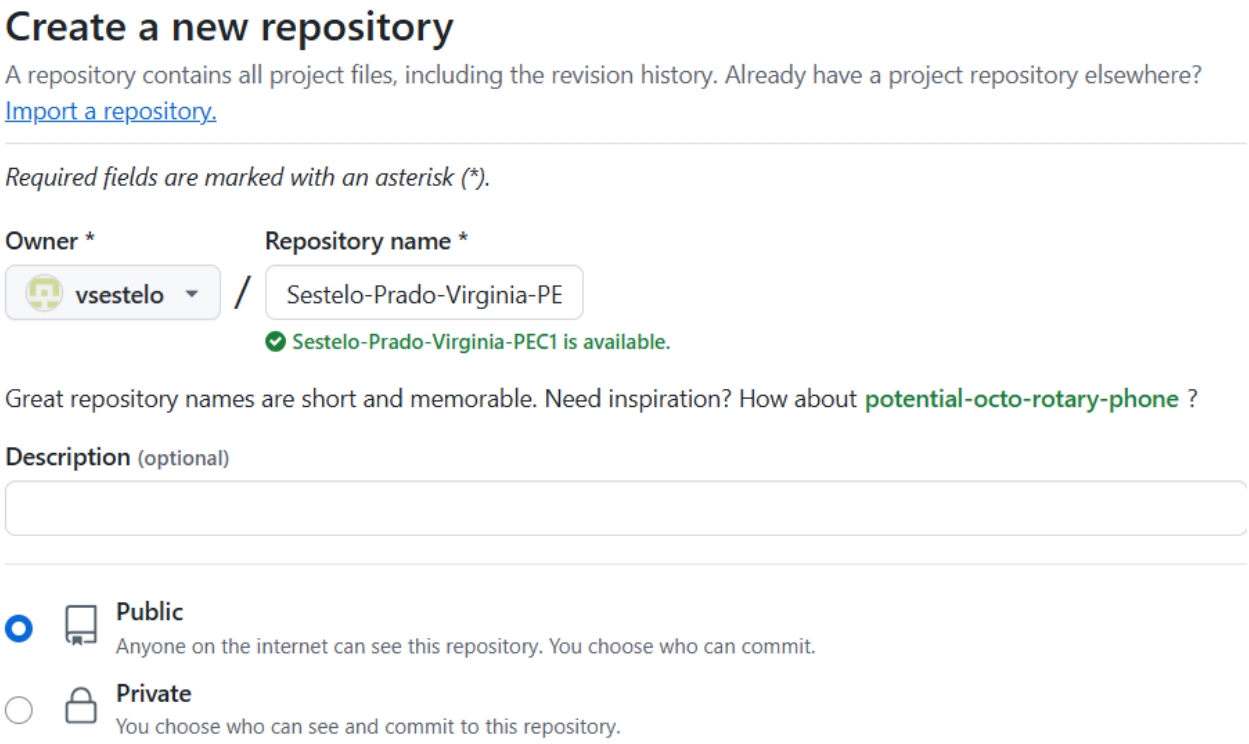
Se guarda el archivo de datos en formato txt y se crea un archivo RMarkdown en el que se introducen los metadatos acerca del dataset.

Ahora, se crea el repositorio de github:

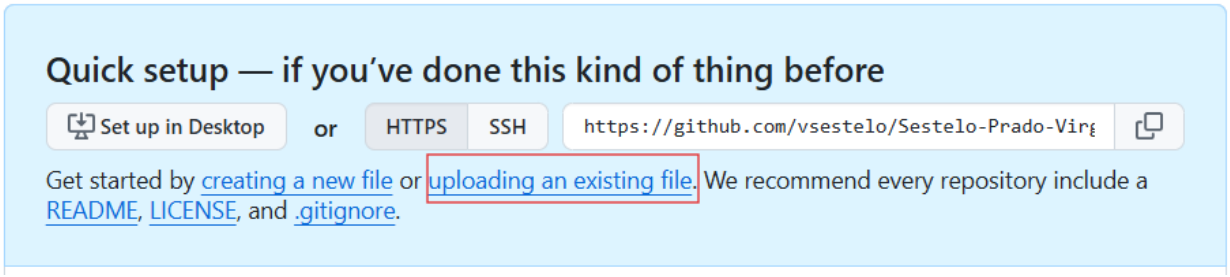
Se accede al menú clicando en el icono de mi perfil. Se pincha en “Create new” y después en “New repository”.



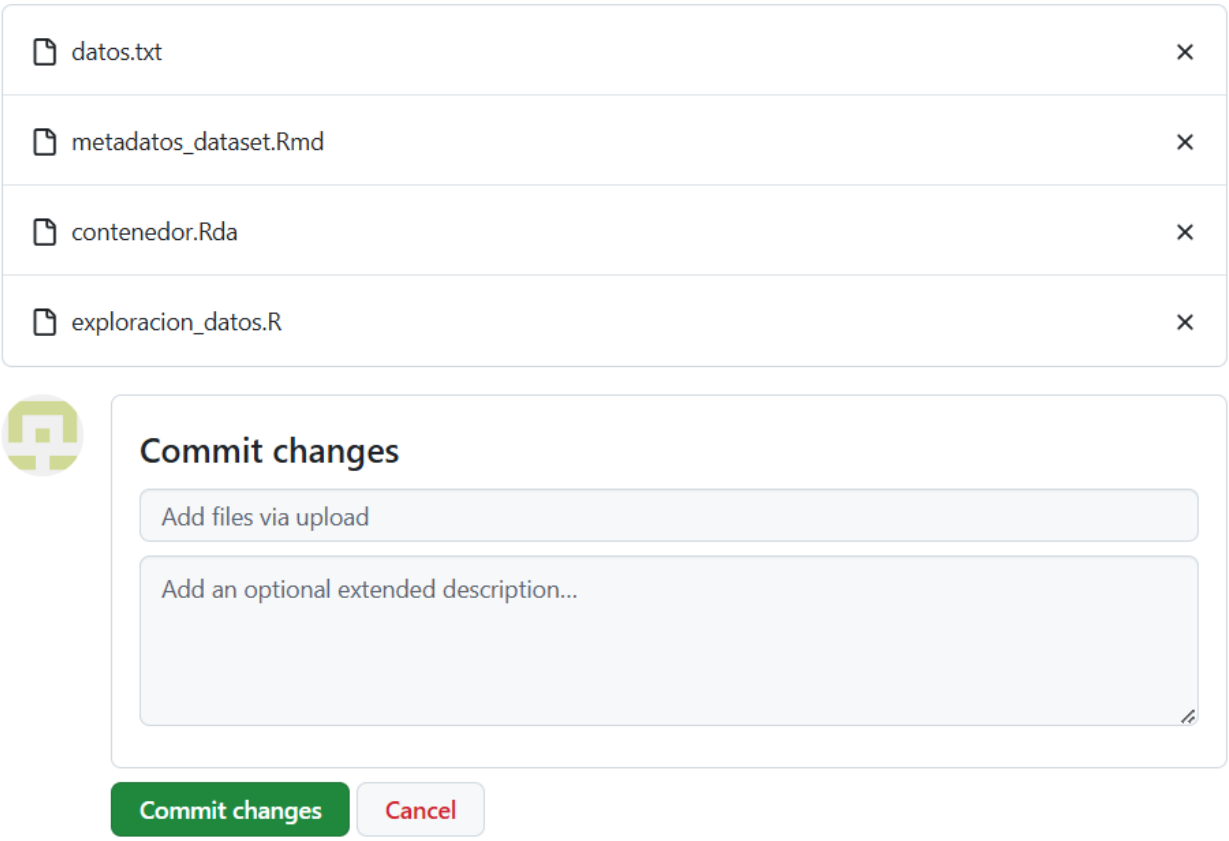
A continuación se introduce el nombre del repositorio (Sestelo-Prado-Virginia-PEC1) y se pulsa en “Create repository”.



En la sección “Quick setup” se pincha en “uploading an existing file”:



Y por último se cargan los archivos y se pincha en "Commit changes".



(Se añadirá también el pdf de este documento)

El enlace al repositorio de github es el siguiente:

https://github.com/vsestelo/Sestelo-Prado-Virginia-PEC1.git