PEC1 - Análisis de datos ómicos

Alumna: Julia Escudero Feliu

1) Selección y carga del dataset

Para esta práctica, he seleccionado un dataset disponible en el repositorio público **Metabolomics Workbench**, que es una fuente confiable y ampliamente utilizada para datos de metabolómica. He seleccionado específicamente el conjunto de datos relacionado con la enfermedad hepática grasa no alcohólica (NAFLD), dado que entra dentro de mi campo de estudio y me interesa mucho, titulado **"Biomarkers of NAFLD progression: a lipidomics approach to an epidemic"**.

Este dataset se enfoca en identificar biomarcadores lipídicos asociados con la progresión de NAFLD a través de distintas fases, incluyendo la esteatosis y la esteatohepatitis no alcohólica (NASH). He seleccionado este dataset debido a su relevancia clínica en el diagnóstico y tratamiento de una enfermedad con alta prevalencia en la población. Además, nos proporciona un conjunto de datos transcriptómicos obtenidos mediante secuenciación de ARN (RNA-Seq), ideal para el análisis multivariante de datos ómicos que se solicita en esta práctica.

Para cargar los datos en R, he descargado el archivo NASHALL-Count.txt, que contiene los conteos de expresión génica por muestra en formato de texto delimitado por tabulaciones. Luego, he usado el siguiente código para cargar el archivo en R, especificando el nombre de la primera columna como identificador de fila (row.names = 1), con el fin de tener los transcritos correctamente identificados en el dataframe de datos.

```
# Cargamos los datos descargados en en R
data <- read.table("/Users/iei/NASH-ALL-Count.txt", header = TRUE, sep =
"\t", row.names = 1)

# Verificamos los datos cargados
head(data)
```

```
##                   T020    X025    X028   T077    T084    T115    T123
X021
## ENST00000456328  0.3132  0.0000  0.0000 0.0000  0.0000   0.000  0.0000
0.0000
## ENST00000515242  0.5079  0.0000  0.0000 0.0000  0.0000   0.000  0.0000
0.0000
## ENST00000518655  0.0058  0.0000  0.0000 0.0000  0.0000   0.000  0.0000
0.0000
## ENST00000450305  0.0000  0.0000  2.9512 0.0000  0.0000   0.000  0.0000
0.0000
```

```
## ENST00000438504  0.1381  0.0001  0.0073 0.2968  0.1627 107.698  3.7960
0.1443
## ENST00000541675 36.2513 66.5433 79.8942 0.1249 51.0956  31.276 56.8687
33.4646
##                    X019 X020    X022    X023    X024    X026    X027
X029
## ENST00000456328  0.0000 0.00  0.0000  1.0068  0.0000  0.0000  0.0000
0.0000
## ENST00000515242  0.0000 0.00  0.0000  0.0000  0.0000  0.0000  0.0000
0.0000
## ENST00000518655  0.0000 0.00  0.0000  0.0000  2.6577  0.0000  0.0000
0.0000
## ENST00000450305  0.0000 0.00  0.0000  0.0000  0.0000  0.0000  0.0000
0.0000
## ENST00000438504 47.5349 0.00 16.4696 24.9874  2.9335 34.3704 61.9019
0.0739
## ENST00000541675 29.8262 0.36 35.7561 35.6466 81.2375 26.1552 50.4914 1
30.3709
##                     X011     X015    X013    X014    X012    X016    X
017
## ENST00000456328   0.0000   0.0000  0.0000  0.0000  0.0000  0.0000  0.0
000
## ENST00000515242   0.0000   0.0000  0.0000  0.0000  0.0000  0.0000  0.0
000
## ENST00000518655   0.0000   0.0000  1.1165  6.1597  0.0000  1.9680  0.0
000
## ENST00000450305   0.0000   0.0000  0.0000  0.0000  0.0000  0.0000  0.0
000
## ENST00000438504   0.2466   0.0033  0.2031 39.2891  3.5712 13.0745 49.9
397
## ENST00000541675 150.5204 135.6107 46.3577 49.8722 22.0722 16.6855  1.1
017
##                    X018    T015    T051    T031    T047     T054    T0
96
## ENST00000456328  0.0000  0.0000  0.0000  0.0000  0.0000   0.0000  0.00
00
## ENST00000515242  0.0000  0.0000  0.0000  0.0000  0.0000   0.0000  0.00
00
## ENST00000518655  0.4950  0.0096  0.0000  0.0000  0.0000   0.0000  0.00
00
## ENST00000450305  0.0000  0.0000  0.0000  0.0000  0.0000   0.0000  0.00
00
## ENST00000438504 32.0613  0.0001  0.0171 36.3724  0.0865   0.0019  3.96
96
## ENST00000541675 35.0410 52.8008 32.8493  0.1524 43.7852 148.1597 70.38
29
##                    T135    T122    T143    T138    T144  T020.1  T123.
1
## ENST00000456328  0.0000  0.0000  0.0000  0.5120  0.0000  0.0000  0.000
0
```

```
## ENST00000515242  0.0000  0.0000  0.0000  0.8302  0.0000  0.0000  0.000
0
## ENST00000518655  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.000
0
## ENST00000450305  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.000
0
## ENST00000438504 10.5447  0.0001 37.8077 15.4663  0.0005 22.1333 37.405
7
## ENST00000541675 16.6850 59.6944 12.0664 94.8735 25.8153 15.5936  1.913
6
##                   X023.1  X024.1  X019.1    T004    T013  T028    T094
T095
## ENST00000456328   0.0000  0.0000  0.0000  0.0000  0.0000 0.000  0.0000
0.0000
## ENST00000515242   0.0000  0.0000  0.0000  0.0000  0.0000 0.000  0.0000
0.0000
## ENST00000518655   0.0000  0.0000  0.0000  0.0000  0.0000 0.000  0.0000
0.0000
## ENST00000450305   0.0000  0.0000  0.0000  0.0000  0.0000 0.000  0.0000
0.0000
## ENST00000438504 112.2855  1.0054 24.2184  0.0139  0.2809 0.000  0.0728
2.8651
## ENST00000541675  21.3641 62.5615 35.7641 47.5904 24.9761 8.289 12.1115
27.0981
##                     T104 X042    P130    X030     X149 P030    P031
P110
## ENST00000456328  0.0000    0  0.0000  0.0000   0.0000    0  0.0000  0.
0000
## ENST00000515242  0.0000    0  0.0000  0.0000   0.0000    0  0.0000  0.
0000
## ENST00000518655  0.0000    0  1.3244  0.0000   0.0068    0  0.0000  0.
0000
## ENST00000450305  0.0000    0  0.0000  0.0000   0.0000    0  0.0000  0.
0000
## ENST00000438504 10.1162    0 30.1823  9.6602   3.6699    0  0.1524  0.
0000
## ENST00000541675 51.1424    0  3.6306 30.4615 154.0514    0 10.0889 60.
4758
##                    P059    P015   P122   T011    T009    T032    T050
T081
## ENST00000456328   0.0000  0.0000 0.0000 0.0000  0.0000  0.0000  0.0000
0.0000
## ENST00000515242   0.0000  0.0000 0.0000 0.0000  0.0000  0.0000  0.0000
0.0000
## ENST00000518655   0.0000  0.0000 0.0000 0.0000  0.0000  3.9990  0.0000
0.0000
## ENST00000450305   0.0000  0.0000 0.0000 0.0000  0.0000  0.0000  0.0000
0.0000
## ENST00000438504   0.1114  8.8025 0.0000 0.0000  0.5928  0.0128  0.0000
1.2815
```

```
## ENST00000541675 123.5109 52.0953 0.0089 0.2446 63.0203 50.8777 29.9643
0.0087
##                    T075    T091    T103    P075    T111    T042    T06
1   T001
## ENST00000456328  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.000
0 0.0000
## ENST00000515242  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.000
0 0.0000
## ENST00000518655  2.6489  0.0000  1.3244  0.0000  0.0000  0.0000  0.000
0 0.0000
## ENST00000450305  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.000
0 0.0000
## ENST00000438504  0.0000  0.0315  0.4263  0.0000  0.0008  0.0000  0.006
8 4.9623
## ENST00000541675 53.7057 72.9416 17.0413 21.8244 40.9747 58.0468 39.882
9 6.1964
##                     T045    T112 T113     T005    T136    T023    P011
T006
## ENST00000456328   0.0000  0.0000    0   0.0000  0.0000  0.0000  0.0000
0.0000
## ENST00000515242   0.0000  0.0000    0   0.0000  0.0000  0.0000  0.0000
0.0000
## ENST00000518655   0.0000  0.0000    0   0.0000  0.0000  0.0000  3.6596
0.0000
## ENST00000450305   0.0000  0.0000    0   0.0000  0.0000  0.0000  0.0000
0.0000
## ENST00000438504   0.8605  0.0000    0   0.2035  0.0002  0.0022  3.4895
0.0001
## ENST00000541675 117.1743 10.6708    0 105.0894 82.0252 35.6317 33.3988
64.3889
##                    T150    T022    T107    P003    X001    X002    X00
3
## ENST00000456328  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.809
3
## ENST00000515242  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  1.312
3
## ENST00000518655  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.000
0
## ENST00000450305  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.000
0
## ENST00000438504  0.0161  0.0001  4.1179  5.7120  0.0001  0.4423  0.600
6
## ENST00000541675 20.1463 35.6892 71.9851 53.1767 11.5718 70.2896 85.982
3
##                     X004     X005     X006    X007    X008     X009
X010
## ENST00000456328   0.0000   0.0000   0.0000  0.0000  0.0000   0.0000
0.0000
## ENST00000515242   0.0000   0.0001   0.0000  0.0000  0.0000   0.0000
0.0000
```

```
## ENST00000518655   0.0000   4.0061   0.0000  0.0000  0.0000   0.0000
0.0000
## ENST00000450305   0.0000   0.0000   0.0000  0.0000  0.0000   0.0000
0.0000
## ENST00000438504   0.0001   0.2840   0.0526  2.0477 23.0550   0.1685
2.4704
## ENST00000541675 132.0632 100.0605 102.9836 64.8998 54.6579 131.0907 11
3.5297
##                 P130.1     T002     T012   T016     T017    T070     T
076
## ENST00000456328 0.0000   0.0000   0.0000 0.0000   0.0000  0.0000   0.0
000
## ENST00000515242 0.0000   0.0000   0.0000 0.0000   0.0000  0.0000   0.0
000
## ENST00000518655 0.0000   0.0000   0.0000 0.0000   0.0000  0.0000   4.0
083
## ENST00000450305 0.0000   0.0000   0.0000 0.0000   0.0000  0.0000   0.0
000
## ENST00000438504 2.2565  38.8024   0.0000 0.0000   0.0020 11.0074   0.3
359
## ENST00000541675 2.4642 243.0274 211.0969 5.6373 110.1146 19.7457 233.0
158
##                    T128     T132     T134    T149 P031.1 P110.1  P059
.1
## ENST00000456328   0.0000   0.0000   0.0000  0.0000      0 0.0000  0.00
00
## ENST00000515242   0.0000   0.0000   0.0000  0.0000      0 0.0000  0.00
00
## ENST00000518655   0.0000   0.0000   0.0000  0.0000      0 0.0000  0.00
00
## ENST00000450305   0.0000   0.0000   0.0000  0.0000      0 0.0000  0.00
00
## ENST00000438504 107.6856  21.0786   0.0420  0.0004      0 0.0000  0.00
00
## ENST00000541675  22.5870 178.0620 252.1055 81.9792      0 0.0285 66.42
86
##                 P015.1 P122.1  T011.1  T009.1  T032.1 T050.1  T081.1
T075.1
## ENST00000456328  0.0000 0.0000  0.0000  0.0000  0.0000 0.0000  0.0000
0.000
## ENST00000515242  0.0000 0.0000  0.0000  0.0000  0.0000 0.0000  0.0000
0.000
## ENST00000518655  0.0000 0.0000  0.0000  0.0000  0.0000 0.0000  0.0000
0.000
## ENST00000450305  0.0000 0.0000  0.0000  0.0000  0.0000 0.0000  0.0000
0.000
## ENST00000438504  1.5272 0.0000  9.0277  7.7016  0.0012 0.0001  4.0946
0.000
## ENST00000541675 21.0449 3.2047 31.8206 34.7627 34.0171 6.5594 15.8364
15.883
```

```
##                  T091.1  T103.1  P075.1  T111.1  T045.1 T112.1 T113.1
T005.1
## ENST00000456328  0.0000  0.0000  0.0000  0.0000  0.0000 0.1585 0.0000
0.0000
## ENST00000515242  0.0000  0.0000  0.0000  0.0000  0.0000 0.2570 0.0000
0.0000
## ENST00000518655  0.0000  0.0000  0.0000  0.0000  0.0000 0.0000 0.0000
0.9784
## ENST00000450305  0.0000  0.0000  0.0000  0.0000  0.0000 0.0000 0.0000
0.0000
## ENST00000438504  9.9322 13.3457  0.0077  0.0097  0.0000 2.1908 0.6393
0.0000
## ENST00000541675 63.0028 16.9807 20.8092 52.2020 19.7579 7.3970 8.4689
0.0023
##                  T136.1 T023.1 P011.1 T006.1 T150.1 T022.1 T107.1 P003
.1
## ENST00000456328  0.0000 0.0000      0  0.000      0 0.0000 0.0000 0.00
00
## ENST00000515242  0.0000 0.0000      0  0.000      0 0.0000 0.0000 0.00
00
## ENST00000518655  0.0000 0.0000      0  0.000      0 0.0000 0.0000 0.00
00
## ENST00000450305  0.0000 0.0000      0  0.000      0 0.0000 0.0000 0.00
00
## ENST00000438504 18.5314 3.3288      0  0.000      0 0.0000 1.5664 1.87
24
## ENST00000541675  6.1606 7.2131      0 29.985      0 3.8534 7.0881 1.00
34
```

2) Carga de metadatos y creación del objeto SummarizedExperiment

Para complementar la matriz de datos de expresión, es necesario cargar también los metadatos asociados a cada muestra. Para ello, utilicé el archivo RNAseq_codes.txt, que contiene identificadores únicos de cada muestra junto con el estado de la enfermedad correspondiente en la columna NASH_IDENTIFIER. Este archivo se me descargó en el paquete dónde estaba el dataset. Esta información es fundamental para realizar análisis diferenciales, ya que nos permite distinguir entre las distintas condiciones presentes en el estudio de NAFLD.

De la misma forma que en el apartado anterior, los datos se cargaron en R utilizando la función read.table() y se verificaron mediante la función head() para asegurarnos que los identificadores de muestra y sus respectivos estados se hayan cargado correctamente.

```
# Cargamos los metadatos de las muestras
sample_metadata <- read.table("/Users/iei/RNAseq_codes.txt", header = TRU
```

```
E, sep = "\t")

# Verificamos, igual que antes, los datos cargados
head(sample_metadata)

##   Sample_ID NASH_IDENTIFIER
## 1       005            C001
## 2       007            C002
## 3       009            C003
## 4       013            C004
## 5       016            C005
## 6       022            C006
```

- ## Creación del objeto:

Una vez cargados los datos de expresión y los metadatos de las muestras, el siguiente paso es la creación del objeto **SummarizedExperiment**. Este tipo de objeto es fundamental para organizar y manipular datos ómicos en R, ya que permite almacenar tanto la matriz de datos como la información adicional de las muestras en un único contenedor estructurado, facilitando el análisis de datos multivariados.

Primero, he verificado, con el código que aparece a continuación, que los nombres de las columnas en la matriz de datos coincidieran con los IDs de las muestras en el archivo de metadatos. Había observado que algunas columnas en la matriz de datos y filas en los metadatos contenían "NA" en los nombres, lo cual podía causar conflictos al crear el objeto SummarizedExperiment. Para resolver esto, se han filtrado tanto las columnas de la matriz de datos como las filas de los metadatos, eliminando aquellos elementos con "NA" en sus nombres y quedándome únicamente con las muestras claramente identificadas.

A continuación, hay qeu asegurarse de que tanto la matriz de datos como los metadatos contengan solo las muestras en común. Además, también se deben ordenar los metadatos para que coincidan con el orden de las columnas en la matriz de expresión, de modo que cada muestra en la matriz de datos tenga su correspondiente información en colData.

Finalmente, con la función SummarizedExperiment() del paquete del mismo nombre podemos crear el objeto, asignando la matriz de expresión de datos como assays y los metadatos de las muestras como colData. Tras ejecutar el código y verificar el objeto creado, debemos confirmar que el objeto SummarizedExperiment contiene correctamente los datos de expresión y los metadatos de las muestras, y así proceder con el análisis.

```r
# Cargamos el paquete SummarizedExperiment
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, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min
```

```
## Loading required package: S4Vectors
```

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

## 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

# Configuramos los nombres de las filas en los metadatos
rownames(sample_metadata) <- sample_metadata$Sample_ID

# Filtramos columnas de 'data' y filas de 'sample_metadata' para eliminar
elementos "NA"
data <- data[, !grepl("^NA", colnames(data))]
sample_metadata <- sample_metadata[!grepl("^NA", rownames(sample_metadata
)), ]

# Filtramos 'data' y 'sample_metadata' para que solo contengan las muestr
as en común
data <- data[, colnames(data) %in% rownames(sample_metadata)]
sample_metadata <- sample_metadata[rownames(sample_metadata) %in% colname
s(data), ]

# Antes de crear el objeto, nos aseguramos de que las muestras estén en e
l mismo orden
sample_metadata <- sample_metadata[match(colnames(data), rownames(sample_
metadata)), ]

# Creamos el objeto SummarizedExperiment
se <- SummarizedExperiment(
```

```
    assays = list(counts = as.matrix(data)),
    colData = sample_metadata
)

# Verificamos el objeto creado
print(se)
```

```
## class: SummarizedExperiment
## dim: 190053 66
## metadata(0):
## assays(1): counts
## rownames(190053): ENST00000456328 ENST00000515242 ... ENST00000435741
##   ENST00000431853
## rowData names(0):
## colnames(66): T020 T077 ... T134 T149
## colData names(2): Sample_ID NASH_IDENTIFIER
```

```
summary(se)
```

```
## [1] "SummarizedExperiment object of length 190053 with 0 metadata colu
mns"
```

El objeto SummarizedExperiment se ha creado y presenta las siguientes características:

- Dimensiones: El objeto contiene 190,053 filas (correspondientes a los transcritos) y 66 columnas (correspondientes a las muestras).
- Assays: Incluye una matriz de datos (counts) que contiene los valores de expresión de cada transcrito en cada muestra.
- Rownames y Colnames: Los nombres de las filas corresponden a los identificadores de los transcritos (por ejemplo, ENST00000456328), mientras que los nombres de las columnas representan los identificadores únicos de las muestras (por ejemplo, T020, T077, etc.).
- colData: Contiene dos columnas de metadatos: Sample_ID y NASH_IDENTIFIER, que describen el ID de cada muestra y el estado de la enfermedad asociado a cada una, respectivamente.
- Metadata: Actualmente, el objeto no contiene columnas de metadatos adicionales.

3) Análisis exploratorio del dataset

Lo siguiente que nos pide la PEC es un análisis exploratorio del dataset, lo que incluye observar la distribución de los datos, identificar posibles valores atípicos y evaluar patrones generales en las muestras y transcritos.

Vamos a empezar con el resumen de los datos. Para ello, implementamos:

```
# Resumen estadístico básico de los datos de conteo
counts <- assay(se)  # Extraer la matriz de conteos
```

```r
# Calculamos estadísticas descriptivas por transcrito (filas)
row_stats <- data.frame(
    mean = rowMeans(counts),
    median = apply(counts, 1, median),
    sd = apply(counts, 1, sd),
    min = apply(counts, 1, min),
    max = apply(counts, 1, max)
)

# Mostramos las primeras filas del resumen estadístico
head(row_stats)
```

```
##                         mean   median          sd min      max
## ENST00000456328  0.01250303  0.00000  0.07337163   0   0.5120
## ENST00000515242  0.02027424  0.00000  0.11897417   0   0.8302
## ENST00000518655  0.25737576  0.00000  0.89101554   0   4.0083
## ENST00000450305  0.00000000  0.00000  0.00000000   0   0.0000
## ENST00000438504  7.17116818  0.07965 20.07952601   0 107.6980
## ENST00000541675 56.80951364 40.42880 60.42276142   0 252.1055
```

El resumen estadístico generado muestra valores de expresión para los primeros transcritos en el dataset. En este caso, para cada transcrito se calculan la media, la mediana, la desviación estándar (sd), el valor mínimo (min) y el valor máximo (max) de sus niveles de expresión a través de las diferentes muestras.

- **Media**: Indica el nivel promedio de expresión de cada transcrito a lo largo de todas las muestras. Observamos que algunos transcritos tienen niveles de expresión promedio muy bajos (por ejemplo, el transcrito ENST00000456328 con una media de 0.0272), mientras que otros muestran niveles de expresión más altos, como ENST00000541675, cuya media es 52.01. Esto sugiere que ciertos transcritos tienen niveles de expresión más elevados de manera consistente en todas las muestras, mientras que otros son casi indetectables o están poco expresados.

- **Mediana**: La mediana, al igual que la media, nos da una medida de tendencia central, pero es menos sensible a valores extremos. En este resumen, observamos que muchos transcritos tienen una mediana de 0, lo cual indica que para la mayoría de las muestras estos transcritos no presentan expresión detectable. Esto es común en datos de RNA-Seq, donde muchos genes pueden estar silenciados o tener niveles bajos de expresión en determinadas condiciones.

- **Standard deviation (sd)**: La desviación estándar indica la variabilidad de la expresión de cada transcrito entre las muestras. Algunos transcritos, como ENST00000541675 con una desviación estándar de 47.38, muestran una gran variabilidad en sus niveles de expresión entre las diferentes muestras, lo cual sugiere que pueden estar regulados de manera diferencial entre las condiciones.

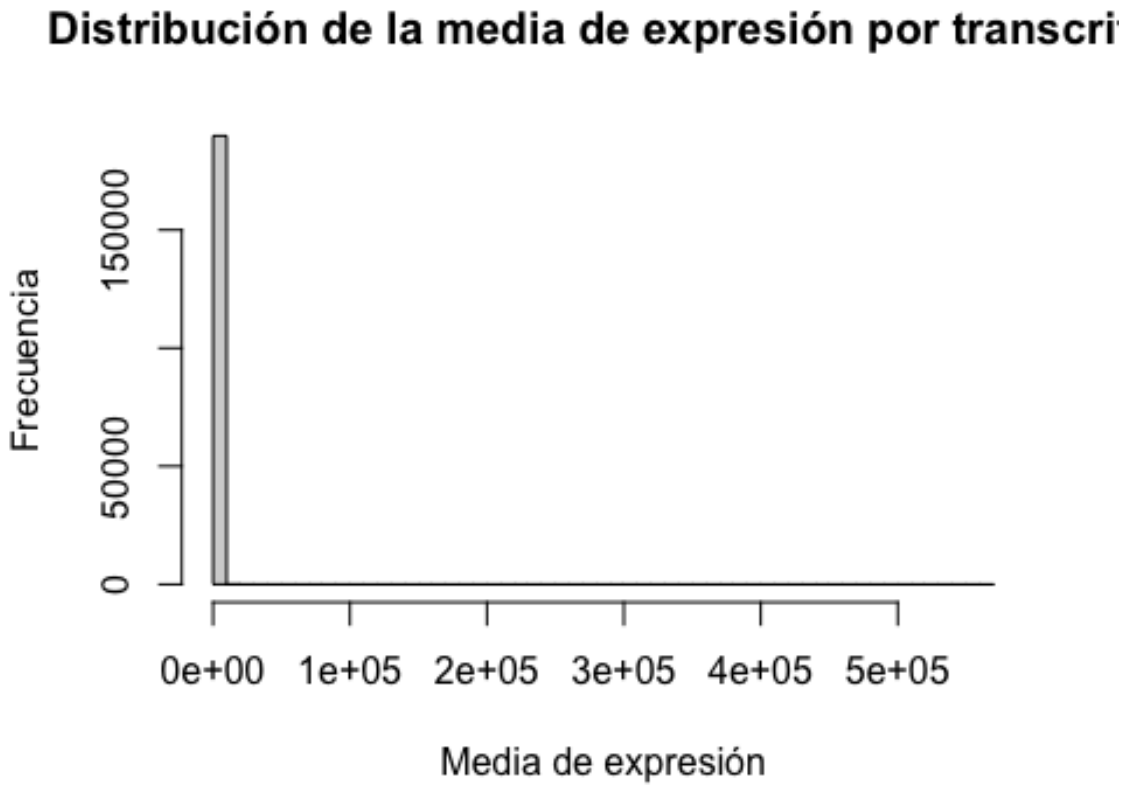
Otros transcritos tienen una desviación estándar muy baja, indicando una expresión más constante.

- **Mínimo y Máximo**: Los valores mínimo y máximo representan los rangos de expresión observados para cada transcrito. En este caso, los valores mínimos para todos los transcritos son 0, lo cual es común en RNA-Seq, ya que algunos genes pueden no estar expresados en ciertas muestras. Los valores máximos varían considerablemente entre los transcritos, alcanzando hasta 243.02 para el transcrito ENST00000541675, lo que indica un alto nivel de expresión en al menos una muestra.

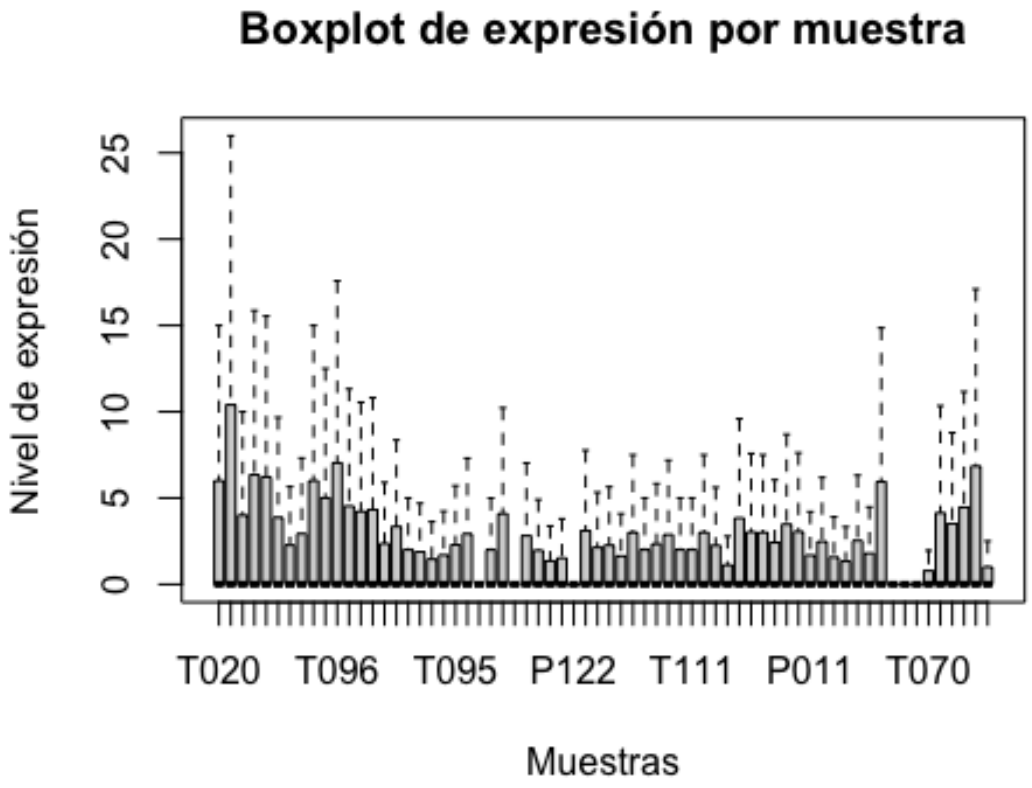
Este análisis preliminar sugiere que el dataset contiene una mezcla de transcritos con diferentes niveles de expresión y variabilidad. Algunos genes muestran una expresión uniforme y baja, mientras que otros presentan niveles elevados y alta variabilidad entre las muestras, lo cual puede ser relevante para identificar genes diferencialmente expresados en estudios posteriores.

Siguiendo con la exploración del dataset, para **visualizar la distribución de los datos**, podemos utilizar un **histograma** de los valores de expresión o un **boxplot** para ver la dispersión en las muestras.

```r
# Histograma de la media de expresión por transcrito
hist(row_stats$mean, breaks = 50, main = "Distribución de la media de exp
resión por transcrito",
     xlab = "Media de expresión", ylab = "Frecuencia")
```



```r
# Boxplot de los valores de expresión por muestra
boxplot(counts, outline = FALSE, main = "Boxplot de expresión por muestra
",
        ylab = "Nivel de expresión", xlab = "Muestras")
```



En el boxplot se muestra la distribución de los niveles de expresión en cada una de las muestras del dataset. Observamos lo siguiente:

- Variabilidad entre muestras: Cada caja representa la dispersión de los niveles de expresión en una muestra específica. Algunas muestras tienen niveles de expresión más altos y una mayor dispersión, mientras que otras muestran una dispersión más baja.
- Presencia de valores atípicos: Los puntos que aparecen por encima de las líneas de los boxplots representan valores atípicos, que son niveles de expresión significativamente más altos para ciertos transcritos en algunas muestras.
- Distribución general: La mayoría de los niveles de expresión se concentran cerca de valores bajos en muchas muestras, lo cual sugiere que solo una fracción de los

transcritos tiene niveles de expresión elevados en cada muestra. Esto es común en datos de RNA-Seq, donde algunos genes se expresan de manera abundante mientras que otros apenas se detectan.

Por otro lado, el **histograma de la media de expresión por transcrito** muestra la distribución de la media de expresión calculada por transcrito en el conjunto de datos:

- Alta frecuencia de bajos niveles de expresión: Observamos una gran cantidad de transcritos con una media de expresión cercana a 0, lo cual significa que una gran proporción de los genes tiene niveles de expresión muy bajos o están apagados en la mayoría de las muestras. Esto es común en los datos transcriptómicos, donde muchos genes no se expresan en condiciones específicas o en ciertos tipos de células.

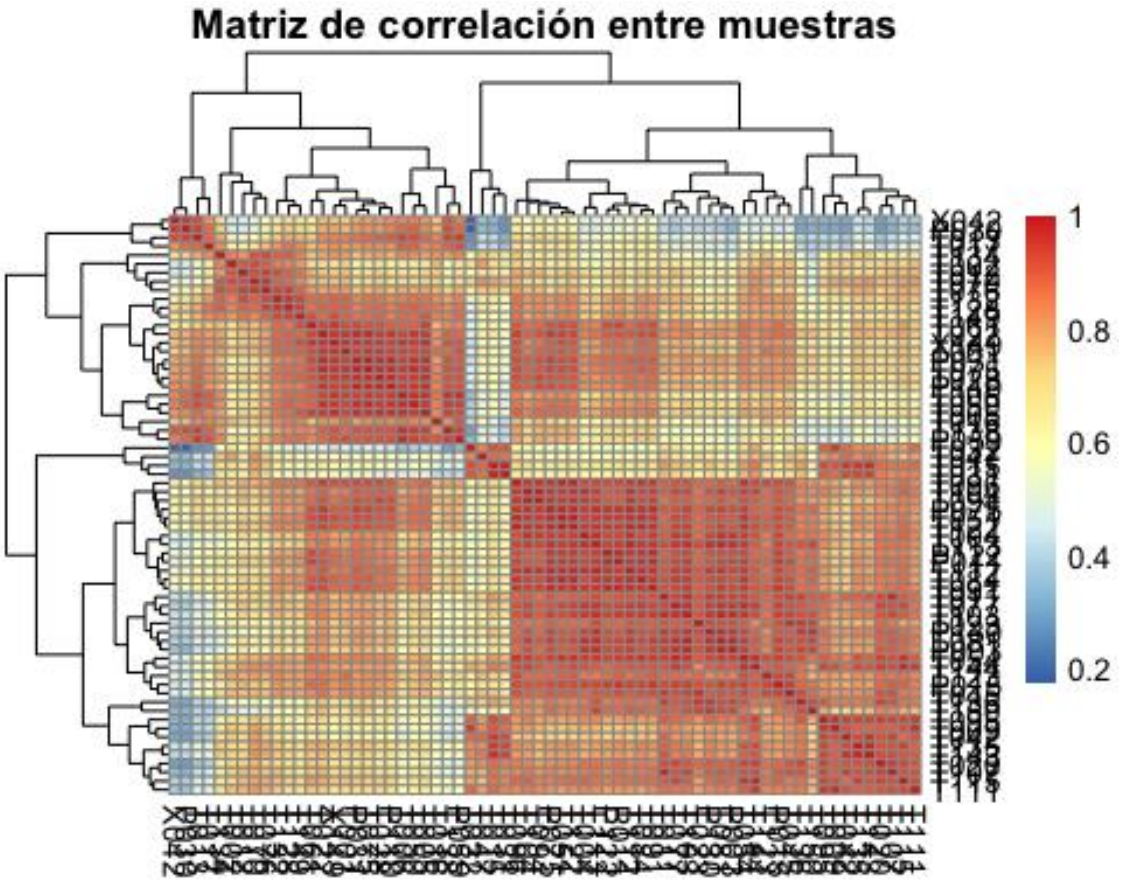
- Escala de expresión: La mayoría de los transcritos tienen niveles de expresión muy bajos, mientras que solo unos pocos alcanzan medias de expresión altas. Esta distribución es típica en datos de RNA-Seq, que tienden a seguir una distribución sesgada hacia la izquierda, con unos pocos genes altamente expresados y muchos genes con expresión baja.

Ambas visualizaciones destacan la naturaleza dispersa y sesgada de los datos de RNA-Seq, donde la mayoría de los genes tienen niveles de expresión bajos o nulos en condiciones específicas, mientras que algunos pocos muestran niveles elevados y una mayor variabilidad entre muestras. Esto es útil para identificar posibles genes de interés en análisis futuros.

Finalmente, podemos realizar un **análisis de correlación** entre las muestras para ayudarnos a entender si hay agrupamientos o similitudes en función del estado de la enfermedad.

```r
# Calculamos la matriz de correlación entre muestras
cor_matrix <- cor(counts)

# Visualizamos la matriz de correlación (se requiere el paquete pheatmap
para la visualización así que primero lo instalamos)
if (!requireNamespace("pheatmap", quietly = TRUE)) {
    install.packages("pheatmap")
}
library(pheatmap)
pheatmap(cor_matrix, main = "Matriz de correlación entre muestras")
```



Como resultado de nuestro programa,la matriz de correlación muestra la similitud en los perfiles de expresión entre las distintas muestras del estudio. En esta visualización, cada celda representa el coeficiente de correlación entre dos muestras, donde los **colores cercanos a 1 (rojo oscuro)** nos indican una alta correlación, lo que sugiere que estas muestras tienen perfiles de expresión similares. Esto puede indicar que estas muestras comparten características biológicas o condiciones experimentales similares. Por otro lado, los **colores cercanos a 0 (azul claro)** nos indican una baja correlación, lo que sugiere que estas muestras tienen perfiles de expresión diferentes y podrían pertenecer a diferentes estados de la enfermedad o a condiciones experimentales distintas.

La presencia de bloques de alta correlación agrupados en la matriz sugiere que existen subconjuntos de muestras con perfiles de expresión similares, lo cual podría reflejar agrupamientos naturales en función de factores biológicos, como diferentes estados de la enfermedad.

Para terminar la práctica y cumplir con los requisitos de entrega, vamos a guardar el objeto creado, “SummarizedExperiment” en .Rda, como piden la PEC, así como crear el archivo README.md con un editor de texto, y el repositorio en GitHub. Allí se subirán todos los archivos que pide la PEC.

Para guardar el objeto:

```
save(se, file = "SummarizedExperiment.Rda")
```