Montero-Coedo-Samuel-PEC1

Samuel Montero

2024-10-29

PEC 1 - ANÁLISIS DE DATOS ÓMICOS

DESCARGA Y LECTURA DE DATOS Y CARGA DE PAQUETES Y LIBRERÍAS

En primer lugar he procedido a la descarga de mis datos a partir del siguiente repositorio de Github: https://github.com/nutrimetabolomics/metaboData/. Dentro del cual accediendo a la carpeta de datasets he seleccionado el titulado 2018-MetabotypingPaper. Luego descargué tanto el archivo DataInfo el cual contiene los Metadatos asociados a las variables y el DataValues que contiene nuestros datos que son medidas de datos de 39 pacientes en 5 puntos temporales.

```
# Leer y cargar los datos
data_values <- read.csv("C:\\Users\\Samuel Montero\\Desktop\\MASTER\\ANÁLISIS DE DATOS ÓMICOS\\PEC1\\Da
data_info <- read.csv("C:\\Users\\Samuel Montero\\Desktop\\MASTER\\ANÁLISIS DE DATOS ÓMICOS\\PEC1\\Data
```

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("SummarizedExperiment")
```

```
## Bioconductor version 3.18 (BiocManager 1.30.25), R 4.3.2 (2023-10-31 ucrt)
```

```
## Warning: package(s) not installed when version(s) same as or greater than current; use
##   'force = TRUE' to re-install: 'SummarizedExperiment'
```

```
## Installation paths not writeable, unable to update packages
##   path: C:/Program Files/R/R-4.3.2/library
##   packages:
##     boot, cluster, codetools, foreign, KernSmooth, lattice, mgcv, nlme, rpart,
##     survival
```

```
## Old packages: 'cli', 'digest', 'rlang', 'xfun', 'yaml'
```

```
library(S4Vectors)
```

```
## 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

##
## Attaching package: 'S4Vectors'

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

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

```
library(SummarizedExperiment)
```

```
## Loading required package: MatrixGenerics

## Loading required package: matrixStats

## Warning: package 'matrixStats' was built under R version 4.3.3

##
## 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: IRanges

##
## Attaching package: 'IRanges'

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

## Loading required package: GenomeInfoDb

## Warning: package 'GenomeInfoDb' was built under R version 4.3.3

## 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
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.3

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:Biobase':
##
##     combine
```

```
## The following objects are masked from 'package:GenomicRanges':
##
##     intersect, setdiff, union

## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect

## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union

## The following object is masked from 'package:matrixStats':
##
##     count

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

PREPARACIÓN Y MANIPULACIÓN DE LOS DATOS

A continuación me dispongo a crear el Contenedor del tipo SummarizedExperiment. En primer lugar revisé documentación para conocer más acerca de esta expresión, en lugares como https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html , https://www.rdocumentation.org/packages/SummarizedExperiment/versions/1.2.3/topics/SummarizedExperiment-class o https://uclouvain-cbio.github.io/bioinfo-training-02-rnaseq/sec-se.html. Lo que tenemos al final es una matriz de datos, en nuestro caso los DataValues que contienen valores clínicos y metabolómicos para 39 pacientes y luego los metadatos asociados a nuestras variables en este caso y no a las muestras. Por lo que habrá que manipular los datos. En primer lugar voy a extraer ciertas columnas de data_values para crear un dataframe de metadatos de las muestras y así tendremos información sobre las muestras y no sobre las variables. En este caso extraigo age, gender, group y surgery.De este modo tendremos un objeto SummarizedExperiment con todos sus campos, sin obviar ninguno. También debemos hacer que el número de columnas en assays coincida con el número de filas en colData por lo que transpongo la matriz data values y de este modo las muestras están en filas y las variables en columnas.

```r
# Transponer data_values para que las muestras estén en filas
data_values_t <- t(data_values)

# Extraer columnas de metadatos de muestras de data_values
metadata <- data_values[, c("SURGERY","AGE", "GENDER", "Group")]

# Asegurar que los nombres de fila de metadata coincidan con los nombres de columna de data_values_t
rownames(metadata) <- colnames(data_values_t)

# Convertir metadata a DataFrame de Bioconductor
metadata <- DataFrame(metadata)
```

Debemos hacer que el número de columnas en assays coincida con el número de filas en colData por lo que transpongo la matriz data values y de este modo las muestras están en filas y las variables en columnas.

COMPROBACIONES PREVIAS

```r
# Verificar que los nombres de fila en data_values_t coinciden con los nombres de fila en data_info
stopifnot(rownames(data_values_t) == rownames(data_info))

# Verificar que los nombres de columna en data_values_t coinciden con los nombres de fila en metadata
stopifnot(colnames(data_values_t) == rownames(metadata))
```

SUMMARIZEDEXPERIMENT

Tras estas comprobaciones es cuando podemos crear tranquilamente el objeto.

```r
# Crear el objeto SummarizedExperiment
se <- SummarizedExperiment(
    assays = list(counts = as.matrix(data_values_t)),
    colData = metadata,
    rowData = data_info
)
```

EXPLORACIÓN GENERAL

Ahora realizaremos una exploración general del dataset.

```r
se
```

```
## class: SummarizedExperiment
## dim: 695 39
## metadata(0):
## assays(1): counts
## rownames(695): SUBJECTS SURGERY ... SM.C24.0_T5 SM.C24.1_T5
## rowData names(3): VarName varTpe Description
## colnames(39): 1 2 ... 38 39
## colData names(4): SURGERY AGE GENDER Group
```

```r
dim(se)
```

```
## [1] 695  39
```

```r
head(assay(se))       # Las primeras filas de la matriz de datos principal
```

```
##          1         2         3         4         5         6         7
## SUBJECTS " 1"      " 2"      " 3"      " 4"      " 5"      " 6"      " 7"
## SURGERY  "by pass" "by pass" "by pass" "by pass" "tubular" "by pass" "tubular"
## AGE      "27"      "19"      "42"      "37"      "42"      "24"      "33"
## GENDER   "F"       "F"       "F"       "F"       "F"       "F"       "F"
## Group    "1"       "2"       "1"       "2"       "1"       "2"       "1"
## MEDDM_TO " 0"      " 0"      " 0"      " 0"      " 0"      " 0"      " 0"
##          8         9         10        11        12        13        14
## SUBJECTS " 8"      " 9"      "10"      "11"      "12"      "13"      "14"
## SURGERY  "tubular" "tubular" "tubular" "tubular" "by pass" "by pass" "by pass"
## AGE      "55"      "40"      "47"      "33"      "57"      "56"      "45"
## GENDER   "F"       "F"       "M"       "M"       "F"       "F"       "F"
## Group    "1"       "1"       "1"       "1"       "2"       "1"       "1"
## MEDDM_TO " 0"      " 0"      " 0"      " 0"      " 0"      " 0"      " 0"
##          15        16        17        18        19        20        21
## SUBJECTS "15"      "16"      "17"      "18"      "19"      "20"      "21"
## SURGERY  "by pass" "by pass" "by pass" "by pass" "by pass" "by pass" "by pass"
## AGE      "55"      "39"      "29"      "27"      "41"      "46"      "41"
## GENDER   "F"       "F"       "F"       "F"       "F"       "F"       "F"
## Group    "1"       "1"       "1"       "1"       "2"       "1"       "2"
## MEDDM_TO " 0"      " 0"      " 0"      " 0"      " 0"      " 0"      " 0"
##          22        23        24        25        26        27        28
## SUBJECTS "22"      "23"      "24"      "25"      "26"      "27"      "28"
## SURGERY  "by pass" "by pass" "by pass" "by pass" "by pass" "tubular" "tubular"
## AGE      "44"      "35"      "58"      "37"      "46"      "42"      "59"
## GENDER   "F"       "F"       "F"       "F"       "F"       "F"       "F"
## Group    "2"       "2"       "2"       "2"       "1"       "1"       "1"
## MEDDM_TO " 0"      " 0"      " 0"      " 0"      NA        " 0"      " 0"
##          29        30        31        32        33        34        35
## SUBJECTS "29"      "30"      "31"      "32"      "33"      "34"      "35"
## SURGERY  "tubular" "by pass" "by pass" "by pass" "by pass" "tubular" "tubular"
## AGE      "55"      "46"      "39"      "33"      "38"      "37"      "39"
## GENDER   "F"       "M"       "M"       "M"       "M"       "M"       "M"
## Group    "2"       "1"       "2"       "1"       "2"       "1"       "2"
## MEDDM_TO " 0"      " 0"      " 0"      " 0"      " 0"      " 0"      " 0"
##          36        37        38        39
## SUBJECTS "36"      "37"      "38"      "39"
## SURGERY  "tubular" "by pass" "tubular" "by pass"
## AGE      "35"      "46"      "41"      "26"
## GENDER   "M"       "M"       "M"       "M"
## Group    "1"       "2"       "1"       "1"
## MEDDM_TO " 0"      " 0"      " 0"      " 0"
```

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

```
## [1] "SURGERY" "AGE"     "GENDER"  "Group"
```

```r
head(colData(se))     # Muestra las primeras filas de colData para un vistazo general
```

```
## DataFrame with 6 rows and 4 columns
##       SURGERY       AGE      GENDER     Group
##   <character> <integer> <character> <integer>
## 1     by pass        27           F         1
## 2     by pass        19           F         2
## 3     by pass        42           F         1
## 4     by pass        37           F         2
## 5     tubular        42           F         1
## 6     by pass        24           F         2
```

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

```
## [1] "VarName"     "varTpe"      "Description"
```

```r
head(rowData(se))     # Muestra las primeras filas de rowData
```

```
## DataFrame with 6 rows and 3 columns
##              VarName      varTpe Description
##          <character> <character> <character>
## SUBJECTS    SUBJECTS     integer    dataDesc
## SURGERY      SURGERY   character    dataDesc
## AGE              AGE     integer    dataDesc
## GENDER        GENDER   character    dataDesc
## Group           Group     integer    dataDesc
## MEDDM_TO    MEDDM_TO     integer    dataDesc
```

También podemos agrupar las muestras dependiendo del tipo de cirgía que se les haya realizado a los diversos sujetos. También podemos visualizarlo, además de hacer un análisis descriptivo.

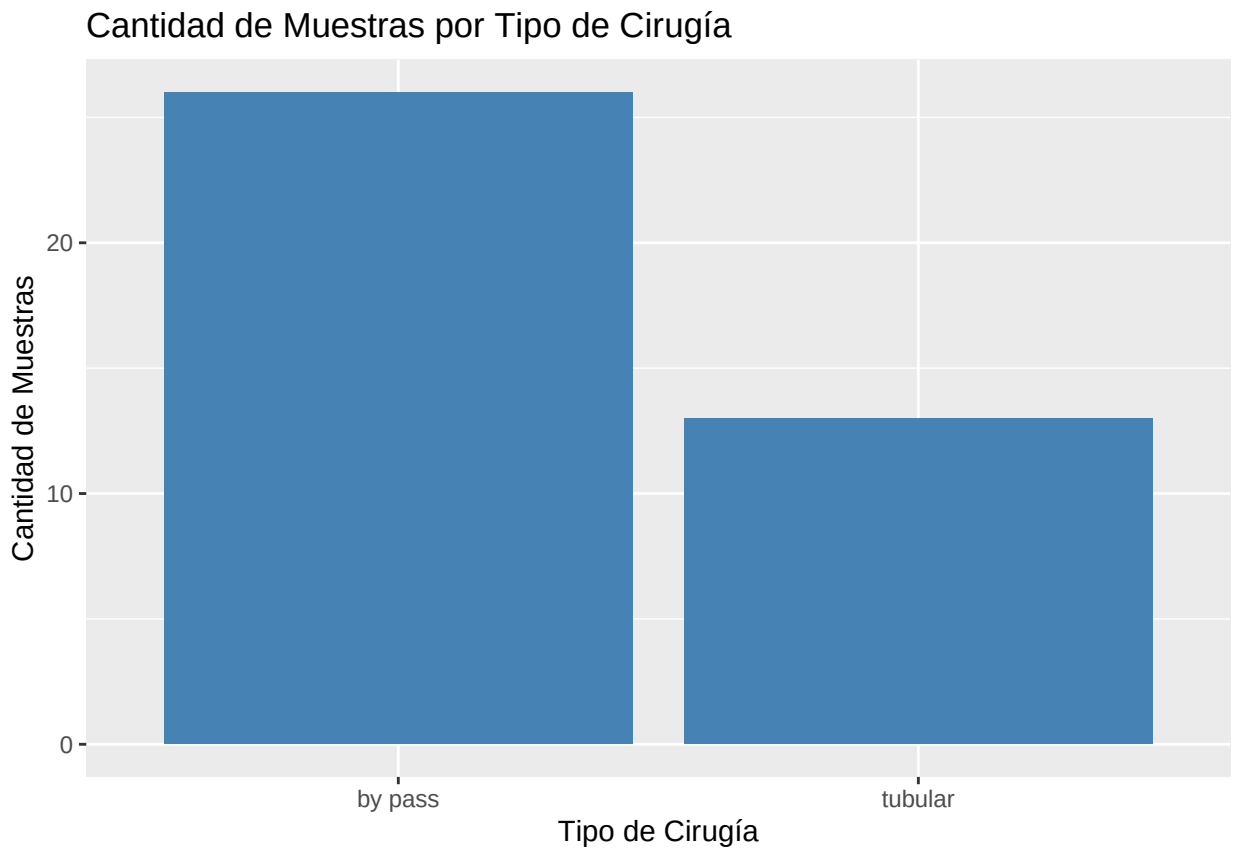
```r
colData(se)
```

```
## DataFrame with 39 rows and 4 columns
##         SURGERY       AGE      GENDER     Group
##     <character> <integer> <character> <integer>
## 1       by pass        27           F         1
## 2       by pass        19           F         2
## 3       by pass        42           F         1
## 4       by pass        37           F         2
## 5       tubular        42           F         1
## ...         ...       ...         ...       ...
## 35      tubular        39           M         2
## 36      tubular        35           M         1
## 37      by pass        46           M         2
## 38      tubular        41           M         1
## 39      by pass        26           M         1
```

```r
# Extraer los datos para "by pass"
se_bypass <- se[, colData(se)$SURGERY == "by pass"]

# Extraer los datos para "tubular"
se_tubular <- se[, colData(se)$SURGERY == "tubular"]


# Convertir colData a data frame
metadata_df <- as.data.frame(colData(se))

# Resumen estadístico agrupado por tipo de cirugía
metadata_df %>%
  group_by(SURGERY) %>%
  summarise(
    avg_age = mean(AGE, na.rm = TRUE),
    count = n()
  )
```

```
## # A tibble: 2 x 3
##   SURGERY avg_age count
##   <chr>     <dbl> <int>
## 1 by pass    39.7    26
## 2 tubular    42.9    13
```

```r
# Gráfico para la visualización
ggplot(metadata_df, aes(x = SURGERY)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Cantidad de Muestras por Tipo de Cirugía", x = "Tipo de Cirugía", y = "Cantidad de Mues
```



OBJETO CONTENEDOR EN FORMATO .RDA

```
save(se, file = "SummarizedExperiment_Objeto.Rda")
rm(list = ls()) # Limpia todos los objetos del entorno
load("C:/Users/Samuel Montero/Desktop/SummarizedExperiment_Objeto.Rda")
ls() # Muestra solo los objetos cargados desde el archivo .rda
```

```
## [1] "se"
```

REPOSITORIO GITHUB

En esta parte del informe se adjunta el link que lleva al repositorio GitHub https://github.com/smonteroco/MONTERO-COEDO-SAMUEL-PEC1.git