PEC1

Edurne Solabarrieta Larrañaga

2024-10-27

Contents

Introducción

La clase SummarizedExperiment, incluida en el paquete SummarizedExperiment de Bioconductor, sirve para almacenar datos experimentales, comúnmente generados en experimentos de secuenciación y microarrays, en forma de matrices rectangulares. Esta clase permite gestionar varios conjuntos de datos experimentales al mismo tiempo, siempre y cuando compartan las mismas dimensiones.

Cada objeto guarda observaciones de una o más muestras, junto con metadatos adicionales que describe tanto las características observadas como los fenotipos de las muestras.

Aunque SummarizedExperiment es similar al ExpressionSet, ofrece mayor flexibilidad en el manejo de filas, lo que permite el uso de estructuras basadas en GRanges o DataFrames arbitrarios. Por eso es adecuado para una amplia variedad de experimentos, como los realizados con técnicas de secuenciación (RNA-Seq, ChIP-Seq, entre otros).

Abstract

Este informe presenta un análisis de datos metabolómicos obtenido del repositorio de GitHub proporcionado para esta PEC, pero originalmente obtenido del Metabolomics Workbench. Estos datos proporcionan muestras intestinales humanas antes y después de un transplante. El conjunto de datos utilizado es después de realizar una limpieza de datos. Lo que se hizo es crear un objeto SummarizedExperiment en R, lo que nos permite organizar de forma más eficiente los metadatos.

Objetivos del estudio

El objetivo de este estudio es ejecutar una versión simplificada del proceso de análisis los datos metabolómicos de muestras intestinales humanas antes y después de un trasplante. Se busca identificar metabolitos que muestren diferencias significativas entre los grupos, proporcionando información relevante sobre el impacto del trasplante en el metabolismo intestinal.

Materiales y métodos

Origen de los datos

El dataset fue obtenido del repositorio de GitHub proporcionado para esta PEC, pero originalmente obtenido del Metabolomics Workbench Se puede encontrar más información en Metabolomics Workbench, con ID de proyecto PR000002 y y DOI doi: 10.21228/M8WC7D.

URL links al dataset:

https://github.com/nutrimetabolomics/metaboData/tree/main/Datasets/2023-UGrX-4MetaboAnalystTutorial

https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Study&StudyID=ST000002&StudyType=MS&ResultType=1

Tipo de datos

Este conjunto de datos incluye medidas de metabolitos en muestras intestinales de humanos antes y después de un trasplante ("Intestinal Samples II pre/post transplantation"). El dataset fue previamente procesado para su análisis, donde se realizaron algunos pasos de limpieza para eliminar las filas de metadatos, reemplazar etiquetas de factor por "Before" y "After," y añadir un prefijo ("B" o "A") a cada muestra según el grupo al que pertenece. Después, se guardó en formato CSV. Para el análisis, se utilizaros estos datos preprocesados, compuesto por 142 metabolitos medidos en 12 muestras de tejido intestinal humano, clasificadas en dos grupos: antes y después del trasplante.

Herramientas utilizadas

Se utilizaron R y Bioconductor, específicamente la biblioteca SummarizedExperiment para gestionar los datos.

Procedimiento de análisis

Los pasos seguidos para el procedimiento de análisis fueron las siguientes:

1. Cargar los paquetes necesarios

Para el análisis y visualización de datos, se utilizaron las siguientes librerías de R:

```
library(readr)
library(SummarizedExperiment)
```

```
## Loading required package: MatrixGenerics

## Loading required package: matrixStats

## Warning: package 'matrixStats' was built under R version 4.4.1

##
## 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

## Warning: package 'S4Vectors' was built under R version 4.4.1

##
## 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

## Warning: package 'IRanges' was built under R version 4.4.1

##
## Attaching package: 'IRanges'

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

## 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
```

2. Cargar el dataset

Se cargaron los datos preprocesados guardados en un archivo CSV:

```
data_path <- "C:/Users/Edurne/OneDrive - Sanquin/Edurne/UOC/5 Análisis de datos ómicos/PEC1/ST000002_AN
data <- read.csv(data_path, sep = "\t", header = TRUE)
```

3. Limpieza y transformación de datos

La primera fila contenía información de los grupos (Before y After), por lo que se omitió. Los datos estaban en formato de caracteres, por lo que se transformaron a numérico. Luego, se asignaron las matrices de datos necesarios para la creación del objeto SummarizedExperiment:

```
# Omitimos la fila de grupos
data <- data[-1, ]

# Extraemos los nombres de los metabolitos (primera columna)
metabolites <- data[, 1]

# Extraemos los datos de metabolitos (el resto de las columnas)
samples <- as.matrix(data[, -c(1)])

# Convertimos los datos a formato numérico
samples <- apply(samples, 2, as.numeric)

# Creamos los metadatos
## Las filas serán los metabolitos y las columnas las muestras
row_data <- DataFrame(metabolite_id = metabolites)
rownames(row_data) <- metabolites  # Asignar nombres de metabolitos como filas
col_data <- DataFrame(sample_id = colnames(samples))
```

4. Creación de un objeto SummarizedExperiment

Se utilizaron la matriz de datos (samples) y los metadatos (row_data y col_data) para crear el objeto SummarizedExperiment, de forma que podamos organizar la información de manera más eficiente:

```
# Creamos el objeto SummarizedExperiment
se <- SummarizedExperiment(assays = SimpleList(counts = samples),
                           rowData = row_data,
                           colData = col_data)
```

5. Análisis de resultados

Se relizó un análisis de los datos utilizando las siguientes funciones de R:

dim(): para ver las dimensiones del objeto (filas y columnas). colnames(): para obtener los metadatos de las columnas (muestras). *rownames(): para obtener los metadatos de las filas (metabolitos).* summary(): para obtener un resumen estadístico general.

Resultados

Los resultados del resumen de los datos comfirma que el objeto SummarizedExperiment contiene 142 metabolitos y 12 muestras. A continuación, podemos ver los nombres de las muestras y los metabolitos.

```r
# Vemos la estructura del objeto
se
```

```
## class: SummarizedExperiment
## dim: 142 12
## metadata(0):
## assays(1): counts
## rownames(142): 1-monoolein 1-monostearin ... xanthine xylose
## rowData names(1): metabolite_id
## colnames(12): A_684508 A_684512 ... B_684499 B_684503
## colData names(1): sample_id
```

```r
# Vemos las dimensiones del objeto SummarizedExperiment
dim(se)
```

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

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

```
##  [1] "A_684508" "A_684512" "A_684516" "A_684520" "A_684524" "A_684528"
##  [7] "B_684483" "B_684487" "B_684491" "B_684495" "B_684499" "B_684503"
```

```r
# Nombres de los metabolitos
rownames(se)
```

```
##   [1] "1-monoolein"                "1-monostearin"
##   [3] "2-hydroxybutanoic acid"     "2-hydroxyglutaric acid"
##   [5] "2-ketoisocaproic acid"      "2-monopalmitin"
##   [7] "2-monostearin NIST"         "3-aminoisobutyric acid"
##   [9] "3-hydroxybutanoic acid"     "3-hydroxypropionic acid"
##  [11] "3-phenyllactic acid"        "5-aminovaleric acid"
##  [13] "acetohydroxamic acid"       "acetophenone NIST"
##  [15] "adipic acid"                "alanine"
##  [17] "alpha ketoglutaric acid"    "aminomalonate"
##  [19] "arabinose"                  "arabitol"
##  [21] "arachidic acid"             "arachidonic acid"
##  [23] "arginine + ornithine"       "asparagine"
##  [25] "aspartic acid"              "behenic acid"
##  [27] "benzoic acid"               "benzylalcohol"
##  [29] "beta-alanine"               "beta-gentiobiose"
##  [31] "beta-sitosterol"            "caffeic acid"
##  [33] "cerotic acid"               "cholesterol"
##  [35] "citric acid"                "citrulline"
##  [37] "creatinine"                 "cyano-L-alanine"
##  [39] "cysteine"                   "cystine"
##  [41] "dihydroabietic acid"        "elaidic acid"
##  [43] "erythritol"                 "ethanolamine"
##  [45] "fructose"                   "fucose 1 + rhamnose 2"
##  [47] "fucose 2"                   "fumaric acid"
##  [49] "GABA"                       "galactose"
```

```
##  [51] "galacturonic acid"           "gamma-tocopherol"
##  [53] "gluconic acid"               "glucose"
##  [55] "glucuronic acid"             "glutamic acid"
##  [57] "glutamine"                   "glutaric acid"
##  [59] "glyceric acid"               "glycerol"
##  [61] "glycerol-3-galactoside NIST" "glycerol-alpha-phosphate"
##  [63] "glycine"                     "glycolic acid"
##  [65] "glycyl proline"              "guanine"
##  [67] "guanosine"                   "heptadecanoic acid"
##  [69] "homoserine"                  "hydrocinnamic acid"
##  [71] "hydroxycarbamate NIST"       "hydroxylamine"
##  [73] "icosenoic acid"              "indole-3-acetate"
##  [75] "indole-3-lactate"            "inosine"
##  [77] "inositol-4-monophosphate"    "inositol myo-"
##  [79] "isoleucine"                  "isolinoleic acid NIST"
##  [81] "isothreonic acid"            "lactic acid"
##  [83] "lanosterol"                  "lauric acid"
##  [85] "leucine"                     "levanbiose"
##  [87] "levoglucosan"                "lignoceric acid"
##  [89] "linoleic acid"               "lysine"
##  [91] "lyxitol"                     "malate"
##  [93] "maleimide"                   "methanolphosphate"
##  [95] "methionine"                  "methionine sulfoxide"
##  [97] "monomyristin NIST"           "monopalmitin-1-glyceride"
##  [99] "myristic acid"               "naphtalene"
## [101] "N-methylalanine"             "nonadecanoic acid"
## [103] "octadecanol"                 "oleic acid"
## [105] "ornithine"                   "orotic acid"
## [107] "oxoproline"                  "palatinose"
## [109] "palmitic acid"               "palmitoleic acid"
## [111] "pelargonic acid"             "pentadecanoic acid"
## [113] "phenylalanine"               "phosphoric acid"
## [115] "proline"                     "putrescine"
## [117] "pyruvate"                    "ribitol"
## [119] "ribose"                      "serine"
## [121] "shikimic acid"               "sorbitol"
## [123] "spermidine"                  "stearic acid"
## [125] "succinic acid"               "sucrose"
## [127] "taurine"                     "threitol"
## [129] "threonic acid"               "threonine"
## [131] "tocopherol"                  "trehalose"
## [133] "tryptophan"                  "tyramine"
## [135] "tyrosine"                    "uracil"
## [137] "urea"                        "uric acid"
## [139] "uridine"                     "valine"
## [141] "xanthine"                    "xylose"
```

A continuación, se presenta el resumen estadístico de los metabolitos:

```
# Resumen estadístico general
summary(assay(se))
```

```
##     A_684508          A_684512          A_684516          A_684520
```

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

Con este resumen estadístico, se puede observar que los valores mínimos y máximos varían ampliamente entre las muestras, lo que indica que existe mucha variabilidad entre los metabolitos.

Además, se observa que en muchos casos la media es elevada en comparación con la mediana. Esto sugiere que existen outliers en los datos, lo que podrían influir en los resultados estadísticos.

Discusión, limitaciones y conclusiones

En este estudio, se llevó a cabo una versión un proceso de análisis los datos metabolómicos de muestras intestinales humanas antes y después de un trasplante. Para ello, se utilizó la herramienta de R y el paquete SummarizedExperiment para gestionar el dataset, que consistía en 142 metabolitos y 12 muestras diferentes.

A partir de este análisis, se observó que los valores de los metabolitos varian considerablemente entre las muestras, lo que sugiere diferencias metabólicas entre los grupos de estudio (antes y después del transplante). Además, las diferencia entre las medias y las medianas de los datos sugiere la posible presencia de valores atípicos (outliers), que podrían influir en los resultados estadísticos. También se observo una variabilidad en los valores mínimos y máximos de los metabolitos, lo que indica que hay metabolitos que pueden responder de manera diferente antes y después del transplante.

Sin embargo, hay que tener en cuenta que este estudio presenta algunas limitaciones. En primer lugar, las muestras obtenidas antes y después del trasplante provienen de diferentes individuos. Esta diferencia en los grupos de comparación podría dar lugar a una interpretación de los datos no adecuada, ya que se introduce una fuente adicional de variabilidad que puede no estar relacionada con el efecto del trasplante en sí. Idealmente, se debería realizar un estudio en donde se evalua los cambios en los metabolitos en los mismos pacientes antes y después del trasplante

Además, como ya se ha mencionado, la presencia de outliers en los datos podría influir en los resultados. Por lo tanto, sería importante realizar una análisis adicional de estos valores extremos.

En conclusión, aunque este estudio ha proporcionado una visión general sobre los varios metabolitos en relación con el trasplante, sería necesario realizar investigaciones adicionales. Estos incluírian un diseño más robusto y un análisis estadístico más detallado que permitirán una mejor comprensión de cómo los cambios metabólicos se relacionan con el transplante.

Referencias

https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

https://www.bioconductor.org/packages/release/bioc/vignettes/structToolbox/inst/doc/data_analysis_omics_using_the_structtoolbox.html

https://www.bioconductor.org/help/course-materials/2019/BSS2019/04_Practical_CoreApproachesInBioconductor.html

https://swcarpentry.github.io/git-novice-es/

Información adicional

```
sessionInfo()
```

```
## R version 4.4.0 (2024-04-24 ucrt)
## Platform: x86_64-w64-mingw32/x64
## Running under: Windows 10 x64 (build 19045)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.utf8
## [2] LC_CTYPE=English_United Kingdom.utf8
## [3] LC_MONETARY=English_United Kingdom.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.utf8
##
## time zone: Europe/Berlin
## tzcode source: internal
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] SummarizedExperiment_1.34.0 Biobase_2.64.0
##  [3] GenomicRanges_1.56.1        GenomeInfoDb_1.40.1
##  [5] IRanges_2.38.1              S4Vectors_0.42.1
##  [7] BiocGenerics_0.50.0         MatrixGenerics_1.16.0
##  [9] matrixStats_1.4.1           readr_2.1.5
##
## loaded via a namespace (and not attached):
##  [1] Matrix_1.7-0            jsonlite_1.8.9          compiler_4.4.0
##  [4] crayon_1.5.3            yaml_2.3.10             fastmap_1.2.0
##  [7] lattice_0.22-6          R6_2.5.1                XVector_0.44.0
## [10] S4Arrays_1.4.1          knitr_1.48              DelayedArray_0.30.1
## [13] tibble_3.2.1            GenomeInfoDbData_1.2.12 pillar_1.9.0
## [16] tzdb_0.4.0              rlang_1.1.4             utf8_1.2.4
```

```
## [19] xfun_0.47              SparseArray_1.4.8      cli_3.6.3
## [22] magrittr_2.0.3         zlibbioc_1.50.0        grid_4.4.0
## [25] digest_0.6.37          rstudioapi_0.16.0      hms_1.1.3
## [28] lifecycle_1.0.4        vctrs_0.6.5            evaluate_1.0.0
## [31] glue_1.7.0             abind_1.4-8            fansi_1.0.6
## [34] rmarkdown_2.28         httr_1.4.7             tools_4.4.0
## [37] pkgconfig_2.0.3        htmltools_0.5.8.1      UCSC.utils_1.0.0
```

Reposición de los datos en GitHub

Se ha creado un repositorio de GitHub que contiene: * este 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 csv, * los metadatos acerca del dataset en un archivo markdown.

La dirección (url) del repositorio es la siguiente:

https://github.com/Eeeeee100/PEC1-Analisis-de-datos-omicos