PEC1 - Análisis de datos ómicos

Anna Sommer

2025-03-27

ENLACE AL REPOSITORIO GITHUB: (https://github.com/gitafps/PEC1 (https://github.com/gitafps/PEC1))

Tabla de contenidos

Resumen

Objetivos

1. Integrar el uso de GitHub y Bioconductor en el contexto del análisis de datos ómicos.
2. Hacer un análisis explorativo de un dataset de metabolómica.

Metodos

Para la creación de la clase SummarizedExperiment instalé y utilicé el paquete del mismo nombre (“SummarizedExperiment”) de Bioconductor.

Para el análisis explorativo de los datos utilicé los paquetes “reshape2” y “ggplot2” para graficar los datos.

Finalmente, para crear el repositorio en GitHub utilicé las funciones del paquete “usethis” (vea resultados para información más detallada).

Resultados

Ejercicio 1

Elegí el dataset 2024-Cachexia del repositorio de Github proporcionado en el ejercicio.

```
cachexia = read.csv("/Users/Lenovo/Downloads/human_cachexia.csv")

class(cachexia)
```

```
## [1] "data.frame"
```

```
dim(cachexia)
```

```
## [1] 77 65
```

```
sum(is.na(cachexia))
```

```
## [1] 0
```

Elegí este dataset, porque el estudio me pareció muy interesante, dado que la caquexia es una enfermedad que a menudo viene asociada a la presencia de un tumor y hasta puede causar la muerte. Aparte, el diseño del estudio me pareció intuitivo, habiendo solo dos grupos de pacientes: caquéctico y control. De todos los pacientes se tomaron una variedad de valores metabólicos que se podrían comparar entre los dos grupos para identificar qué valores están particularmente afectados por la condición de la caquexia. Un breve análisis nos demuestra que los datos vienen organizados en un objeto de clase data.frame, habiendo 77 registros de pacientes y 65 variables (incluídas la ID de paciente y la condición (grupo)). Una ventaja del dataset es también la ausencia de valores faltantes (NA).

Ejercicio 2

Ya que la clase SummarizedExperiment es una clase de Bioconductor, primero instalé el paquete ("SummarizedExperiment") correspondiente usando el BiocManager (ejemplo abajo).

```
BiocManager::install("SummarizedExperiment")
```

```
## Bioconductor version 3.20 (BiocManager 1.30.25), R 4.4.3 (2025-02-28 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-4.4.3/library
##   packages:
##     cluster, foreign, lattice, MASS, Matrix, mgcv, nlme
```

```
## Old packages: 'curl', 'gert', 'httr2', 'jsonlite', 'stringi'
```

```
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, saveRDS, setdiff,
##     table, tapply, union, unique, unsplit, which.max, which.min
```

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

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

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

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

```
## Loading required package: IRanges
```

```
##
## 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
```

```
sumex <- SummarizedExperiment(assay = list(counts = cachexia))
```

Creé el objeto SummarizedExperiment proporcionando como “assay data” los datos de caquexia anteriormente descargados. Ya que el dataset de GitHub no proporciona informaciones adicionales sobre las muestras ni tampoco sobre los diferentes metabolitos, no fue posible agregar la “colData” ni la “rowData” al SummarizedExperiment. Sin embargo, igual que para la clase ExpressionSet, la información adicional no es obligatoria para la creación de la clase. El único elemento indispensable para crear un SummarizedExperiment es la matrix de datos experimentales.

Podemos comprobar que se trata de un objeto de clase SummarizedExperiment:

```
class(sumex)
```

```
## [1] "SummarizedExperiment"
## attr(,"package")
## [1] "SummarizedExperiment"
```

Por último, solo falta guardar el objeto SummarizedExperiment en formato “.Rda”:

```
save(sumex, file = "summarizedexperiment.Rda")
```

La diferencia principal entre la clase SummarizedExperiment en comparación con la clase ExpressionSet es que la primera es más flexible con respecto a la información contenida en las filas de la tabla, ya que permite tanto el almacenamiento de objetos GRanges como de objetos DataFrame.

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

Ejercicio 3

La primera parte del análisis explorativo de los datos realizado, como se puede ver en el documento del código correspondiente (“pec1 code”, en el repositorio de GitHub), es un resumen estadístico de las variables. Sin embargo, dado que el dataset contiene muchos variables diferentes, la interpretación de este resumen no resultó muy intuitiva. Al graficar los datos, sin embargo, se mostraron algunas tendencias interesantes. El primer histograma nos muestra, que los sujetos tienen un nivel de metabolitos relativamente bajo para la gran mayoría de los metabolitos medidos. Esto, para mi, levantó la pregunta de si este hecho aplica tanto para pacientes sanos como`para pacientes caquexicos. Por eso, en el siguiente paso, dividé el SummarizedExperiment en dos subsets, uno conteniendo solo los registros de pacientes sanos y el otro conteniendo solo los registros de pacientes enfermos, y volví a analizar la distribución de los niveles de metabolitos. Este análisis demostró que en el grupo de los pacientes caquéxicos, algunos de los niveles de metabolitos llegan a ser mucho más altos que en el grupo control. La desventaja de estos gráficos fue, que aunque me permitió contrastar los niveles por grupo, no proporcionó una comparación por variables específicas. Por esto, decidí incluír en el análisis explorativo también algunos violin plots que compararan los niveles de metabolitos por grupo y por metabolito específico. Considerando que el dataset contiene datos de 63 diferentes metabolitos, y hacer 63 gráficos podría ser algo abrumador en la interpretación, decidí ejemplificar los gráficos solo para los primeros 10 metabolitos. De esta manera esperaba poder confirmar o desechar lo observado en los anteriores gráficos: la tendencia de que los niveles de metabolitos suelen ser más altos en pacientes caquexicos. Por supuesto, el análisis se podría extender según el interés que uno tiene en metabolitos específicos del dataset. Los violin plots confirmaron claramente que los metabolitos observados llegan a niveles elevados en los individuos que sufren de caquexia. En resumen, el análisis explorativo mostró que la caquexia parece provocar niveles elevados de metabolitos.

Ejercicio 4

La elaboración de este informe la realizé conforme iba solucionando los ejercicios anteriores, para así anotar simultáneamente el código, los comentarios al código, mi razonamiento y otras explicaciones.

Ejercicio 5

Para crear el repositorio github usé la función use_github() del paquete “usethis”. A mi nuevo repositorio de github le di el nombre de “PEC1”. Se puede encontrar siguiendo el enlace proporcionado al inicio de este enforme (https://github.com/gitafps/PEC1 (https://github.com/gitafps/PEC1)).

Discusión

El dataset elegido tiene una estructura intuitiva y por eso resultó relativamente conveniente trabajar con él. Sin embargo, faltaron los valores de ColData y RowData (es decir, datos adicionales sobre los registros y el experimento en general). Esto no fue un obstáculo al crear el objeto de clase SummarizedExperiment, aunque en general sería mejor contar también con esta información. El análisis explorativo de los datos mostró una tendencia interesante de los niveles de metabolitos, contrastando el grupo de pacientes caquexicos con el grupo control.

Conclusiones

En el futuro sería interesante trabajar con un dataset que cuente con más información, para aprovechar mejor las funciones de la clase SummarizedExperiment. Sin embargo, considero que los objetivos se han cumplido, ya que esta tarea sirvió de practicar el uso integrado de BioConductor y GitHub y además introdujo la clase SummarizedExperiment.

Referencias

https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html (https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html)
https://github.com/nutrimetabolomics/metaboData/tree/main/Datasets/2024-Cachexia (https://github.com/nutrimetabolomics/metaboData/tree/main/Datasets/2024-Cachexia)