PEC 1

MARTINEZ REQUENA ADRIAN

2024-11-04

Contents

Abstract

Este trabajo se centra en la exploración y análisis de datos de metabolitos en muestras de individuos con y sin cachexia. Para ello se utiliza la herramienta SummarizedExperiment, con la que se organizan los datos en una estructura propia del paquete Bioconductor para que el alumno se familiarice con ello. Por último, los resultados se exponen y comparten a traves de un repositorio de GitHub.

Objetivos del estudio

Mediante este trabajo se realiza un ejercicio de repaso y ampliación para trabajar con Bioconductor y la exploración de datos. Se utilizará para ello datos de muestras de individuos con cachexia y sin ella.

Materiales y Métodos

Se ha utilizado un dataframe obtenido de specmine.datasets que reune un conjunto de datos de cachexia, el cual contiene datos de metabolitos medidos en diversas muestras.

Por otro lado, las herramientas principales utilizadas incluyen SummarizedExperiment y specmine.datasets de R, así como el paquete de Bioconductor.

Resultados

Selección del dataset

Para ello hemos creado un nuevo proyecto clonando el repositorio de GitHub que se ha entregado junto a la práctica. Se muestran a continuación los archivos que contiene:

```
## [1] "2024-metaboData.Rproj" "Data_Catalog.xlsx"     "Datasets"
## [4] "LICENSE"               "README.html"           "README.md"
```

Ahora mostramos las opciones de datasets a escoger

```
## [1] "2018-MetabotypingPaper"           "2018-Phosphoproteomics"
## [3] "2023-CIMCBTutorial"               "2023-UGrX-4MetaboAnalystTutorial"
## [5] "2024-Cachexia"                    "2024-fobitools-UseCase_1"
```

Vamos a seleccionar el dataset de 2024-Cachexia. El archivo Data_Catalog.xlsx nos indica que este dataset puede encontrarse en el paquete de R "specmine.datasets". Para acceder a él se ha tenido que realizar este paso indicado en RDocumentation

```
install.packages("devtools")
devtools::install_github("BioSystemsUM/specmine.datasets")
```

Y por lo tanto, una vez instalados estos paquetes podemos acceder al conjunto de datos de cachexia de esta forma:

```
library(specmine.datasets)
data("cachexia")
```

Observemos la estructura de los datos y los metadatos

```
##  num [1:63, 1:77] 40.9 65.4 18.7 26.1 71.5 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:63] "1.6-Anhydro-beta-D-glucose" "1-Methylnicotinamide" "2-Aminobutyrate" "2-Hydroxyi
##   ..$ : chr [1:77] "PIF_178" "PIF_087" "PIF_090" "NETL_005_V1" ...
```

Podemos apreciar que 'cachexia' es una matriz con 63 filas y 77 columnas de valores numéricos. Las filas corresponden los metabolitos que se han medido y las columnas a las diferentes muestras.

```
## 'data.frame':    77 obs. of  1 variable:
##  $ Muscle.loss: Factor w/ 2 levels "cachexic","control": 1 1 1 1 1 1 1 1 1 1 ...
```

La estructura de los metadatos nos indica que hay 77 observaciones y una sola variable llamada Muscle.loss que es un factor con dos niveles: "cachexic" y "control".

SummarizedExperiment del dataset

Siguiendo las indicaciones expuestas en la guía de Bioconductor podemos crear nuestro propio Summarized-Experiment.

```r
library(SummarizedExperiment)
library(specmine.datasets)

# Creamos el contenedor SummarizedExperiment
se_cachexia <- SummarizedExperiment(assays = list(counts = as.matrix(cachexia$data)),
    rowData = DataFrame(Compound = rownames(cachexia$data)),
    colData = cachexia$metadata)
```

Para ello colocamos nuestros datos en forma de matriz, por motivo de compatibilidad, utilizando as.matrix. Esto será el 'assay' de nuestro SummarizedExperiment. Después daremos nombre a las filas de la matriz (que son los metabolitos) y las columnas les daremos la información de metadata sobre la perdida de músculo y peso.

Tras realizar este proceso, podemos ver el resultado:

```
## class: SummarizedExperiment
## dim: 63 77
## metadata(0):
## assays(1): counts
## rownames(63): 1.6-Anhydro-beta-D-glucose 1-Methylnicotinamide ...
##   pi-Methylhistidine tau-Methylhistidine
## rowData names(1): Compound
## colnames(77): PIF_178 PIF_087 ... NETL_003_V1 NETL_003_V2
## colData names(1): Muscle.loss
```

Exploración del dataset

Ahora proseguimos explorando los datos que tenemos.

```
## [1] "SummarizedExperiment object of length 63 with 1 metadata column"
```

```
## Dimensiones (filas, columnas): 63 77
```

Vemos que los datos tienen una longitud de 63 filas, con 77 columnas, y tan solo una columna de metadata que nos indica si catalogamos la muestra como 'normal' o 'cachexia'.

```
## Nombres de las filas (primeras 5): 1.6-Anhydro-beta-D-glucose 1-Methylnicotinamide 2-Aminobutyrate 2-
```

```
## Nombres de las columnas (primeras 5): PIF_178 PIF_087 PIF_090 NETL_005_V1 PIF_115
```

De esta manera comprobamos que las filas incluyen los diferentes metabolitos y las columnas los ID de las muestras. Debido a la cantidad extensa de datos, se ha decidido tan solo mostrar las 5 primeras.

Veamos por ejemplo ahora el primer y último metabolito, junto a los resultados obtenidos en cada muestra.

```
## Primer metabolito
```

```
##                            PIF_178 PIF_087 PIF_090 NETL_005_V1 PIF_115 PIF_110
## 1.6-Anhydro-beta-D-glucose   40.85   62.18  270.43      154.47    22.2  212.72
##                            NETL_019_V1 NETCR_014_V1 NETCR_014_V2 PIF_154
## 1.6-Anhydro-beta-D-glucose      151.41         31.5        51.42  117.92
```

```
##                            NETL_022_V1 NETL_022_V2 NETL_008_V1 PIF_146 PIF_119
## 1.6-Anhydro-beta-D-glucose        20.7      127.74       59.74   89.12   23.57
##                            PIF_099 PIF_162 PIF_160 PIF_113 PIF_143 NETCR_007_V1
## 1.6-Anhydro-beta-D-glucose   41.26  589.93  112.17  167.34  183.09       208.51
##                            NETCR_007_V2 PIF_137 PIF_100 NETL_004_V1 PIF_094
## 1.6-Anhydro-beta-D-glucose        34.81  333.62   32.46        4.71   68.72
##                            PIF_132 PIF_163 NETCR_003_V1 NETL_028_V1 NETL_028_V2
## 1.6-Anhydro-beta-D-glucose  214.86   304.9        37.71        45.6       34.12
##                            NETCR_013_V1 NETL_020_V1 NETL_020_V2 PIF_192
## 1.6-Anhydro-beta-D-glucose       107.77       13.33       27.94  141.17
##                            NETCR_012_V1 NETCR_012_V2 PIF_089 NETCR_002_V1
## 1.6-Anhydro-beta-D-glucose        14.01       244.69  123.97       141.17
##                            PIF_179 PIF_114 NETCR_006_V1 PIF_141 NETCR_025_V1
## 1.6-Anhydro-beta-D-glucose   35.16   685.4       278.66    15.8        29.96
##                            NETCR_025_V2 NETCR_016_V1 PIF_116 PIF_191 PIF_164
## 1.6-Anhydro-beta-D-glucose        16.95       292.95   29.67   18.92  127.74
##                            NETL_013_V1 PIF_188 PIF_195 NETCR_015_V1 PIF_102
## 1.6-Anhydro-beta-D-glucose       34.81   65.37   15.18        70.81   25.28
##                            NETL_010_V1 NETL_010_V2 NETL_001_V1 NETCR_015_V2
## 1.6-Anhydro-beta-D-glucose       34.47       18.54       37.34        33.78
##                            NETCR_005_V1 PIF_111 PIF_171 NETCR_008_V1
## 1.6-Anhydro-beta-D-glucose        22.42  146.94   64.07        32.46
##                            NETCR_008_V2 NETL_017_V1 NETL_017_V2 NETL_002_V1
## 1.6-Anhydro-beta-D-glucose        113.3        22.2       46.53      192.48
##                            NETL_002_V2 PIF_190 NETCR_009_V1 NETCR_009_V2
## 1.6-Anhydro-beta-D-glucose      528.48   28.79       181.27        47.47
##                            NETL_007_V1 PIF_112 NETCR_019_V2 NETL_012_V1
## 1.6-Anhydro-beta-D-glucose       15.96   22.87        35.16       16.95
##                            NETL_012_V2 NETL_003_V1 NETL_003_V2
## 1.6-Anhydro-beta-D-glucose        9.39       37.71       38.47

## Último metabolito

##                      PIF_178 PIF_087 PIF_090 NETL_005_V1 PIF_115 PIF_110
## tau-Methylhistidine  160.77  130.32   83.93      254.68   79.84   68.72
##                      NETL_019_V1 NETCR_014_V1 NETCR_014_V2 PIF_154 NETL_022_V1
## tau-Methylhistidine       21.98        17.29       101.49   81.45       47.94
##                      NETL_022_V2 NETL_008_V1 PIF_146 PIF_119 PIF_099 PIF_162
## tau-Methylhistidine       95.58       60.95  159.17       8    36.6   75.94
##                      PIF_160 PIF_113 PIF_143 NETCR_007_V1 NETCR_007_V2 PIF_137
## tau-Methylhistidine   43.82   41.26   78.26       151.41       172.43   55.15
##                      PIF_100 NETL_004_V1 PIF_094 PIF_132 PIF_163 NETCR_003_V1
## tau-Methylhistidine   18.73       64.72  170.72  130.32  265.07        15.18
##                      NETL_028_V1 NETL_028_V2 NETCR_013_V1 NETL_020_V1
## tau-Methylhistidine       119.1       84.77       287.15       46.06
##                      NETL_020_V2 PIF_192 NETCR_012_V1 NETCR_012_V2 PIF_089
## tau-Methylhistidine       26.31    36.6        62.18       317.35    62.8
##                      NETCR_002_V1 PIF_179 PIF_114 NETCR_006_V1 PIF_141
## tau-Methylhistidine          137   28.22  127.74        76.71  210.61
##                      NETCR_025_V1 NETCR_025_V2 NETCR_016_V1 PIF_116 PIF_191
## tau-Methylhistidine       239.85       249.64       144.03   18.54   12.55
##                      PIF_164 NETL_013_V1 PIF_188 PIF_195 NETCR_015_V1 PIF_102
## tau-Methylhistidine  125.21       16.44   11.13    8.58       156.02  170.72
##                      NETL_010_V1 NETL_010_V2 NETL_001_V1 NETCR_015_V2
```

```
## tau-Methylhistidine       18.54       16.78       20.09        113.3
##                      NETCR_005_V1 PIF_111 PIF_171 NETCR_008_V1 NETCR_008_V2
## tau-Methylhistidine       184.93   26.31  100.48        16.12        79.84
##                      NETL_017_V1 NETL_017_V2 NETL_002_V1 NETL_002_V2 PIF_190
## tau-Methylhistidine        55.7       15.96       71.52      287.15   52.46
##                      NETCR_009_V1 NETCR_009_V2 NETL_007_V1 PIF_112 NETCR_019_V2
## tau-Methylhistidine        48.42         9.03       29.67   17.46        84.77
##                      NETL_012_V1 NETL_012_V2 NETL_003_V1 NETL_003_V2
## tau-Methylhistidine        44.7       28.22       90.02       27.39
```

Ahora podemos explorar los metadatos que nos indican si la muestra pertenece al grupo control o ha desarrollado cachexia:

```
## Metadatos:
```

```
## DataFrame with 77 rows and 1 column
##              Muscle.loss
##                 <factor>
## PIF_178         cachexic
## PIF_087         cachexic
## PIF_090         cachexic
## NETL_005_V1     cachexic
## PIF_115         cachexic
## ...                  ...
## NETCR_019_V2     control
## NETL_012_V1      control
## NETL_012_V2      control
## NETL_003_V1      control
## NETL_003_V2      control
```

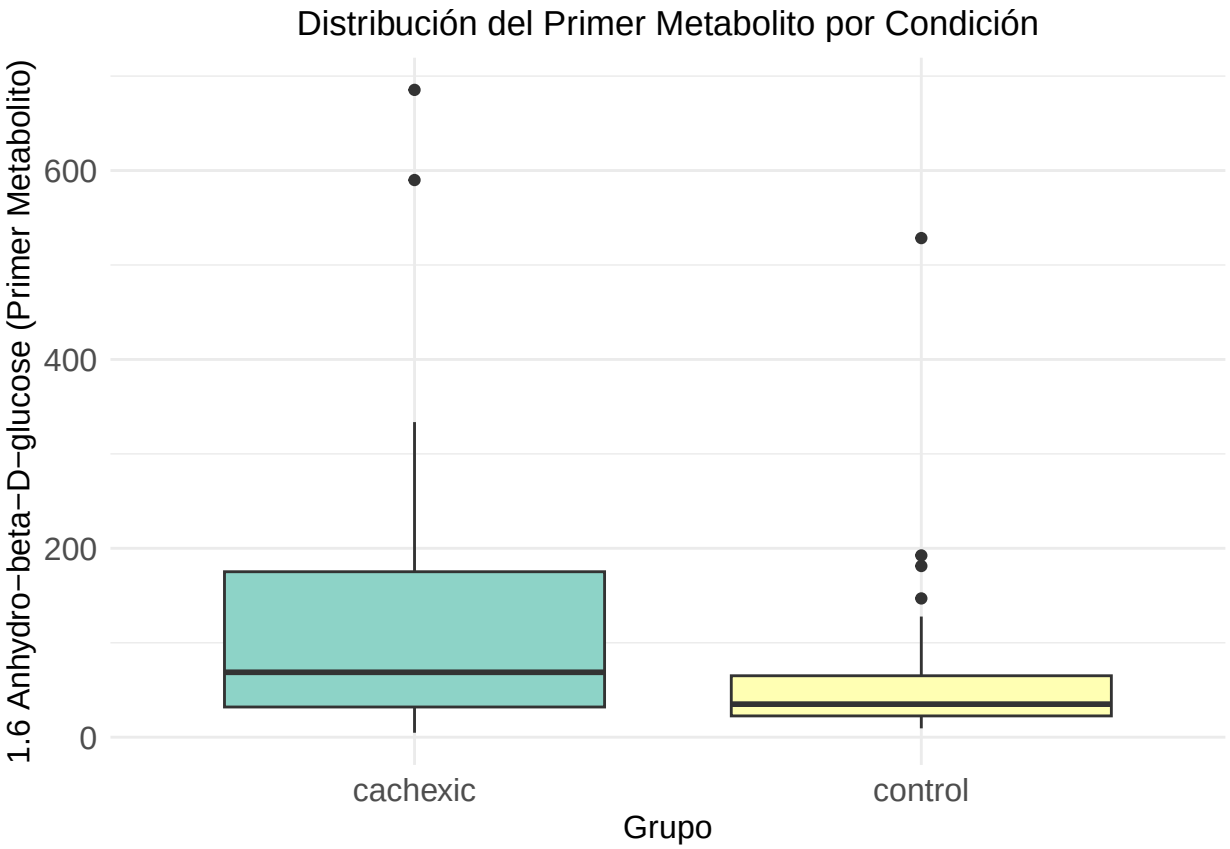
```
##
## cachexic  control
##       47       30
```

Vemos que hay 30 muestras normales y 47 con cachexia.

```
## Valores faltantes en el dataset: 0
```

Además, no hay ningún missing value en nuestra tabla

Por último, vamos a ver los valores del primer metabolito en un Boxplot comparativo entre 'control' y 'cachexia'



Se puede apreciar que los niveles de 1.6 Anhydro-beta-D-glucose son mayores en las muestras con cachexia.

Discusión y limitaciones y conclusiones del estudio

Algunas de las limitaciones a las que puede enfrentarse el estudio y datos recogidos serían:

-La muestra, con únicamente 77 personas, podría no ser lo bastante representativa.

- Podrían hacer falta más metabolitos que no se estuvieran teniendo en cuento.
- No se tienen en cuenta otros factores que puedan influir en los niveles de metabolitos, como la dieta o la actividad física de los participantes.

En conclusión, este trabajo se ha enfocado en investigar y examinar datos de metabolitos en muestras de personas con y sin cachexia, empleando herramientas del paquete Bioconductor. La aplicación de SummarizedExperiment facilitó la organización y estructuración de los datos, y se pudieron aplicar algunas de las ideas impartidas en la asignatura durante las primeras actividades.

Enlace a repositorio

El repositorio puede encontrarse en este enlace: https://github.com/RhysMoonbeam/MARTINEZ-REQUENA-ADRIAN-PEC1.git