PEC1 - Análisis de Datos Ómicos

Eva Montoliu Silvestre

2024-11-06

Contents

En el caso de que no tengamos instalado bioconductor, realizamos la instalación necesaria. Esto incluye tanto `BiocManager` como el paquete especifico para trabajar con contenedores de tipo `SummarizedExperiment`:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install(version = "3.20")
BiocManager::install("SummarizedExperiment")
```

Además, antes de empezar, cargamos todos los paquetes que vamos a necesitar a lo largo de la tarea:

```
# Carga de paquetes necesarios
library(readr)  # Para leer archivos csv
library(tidyverse)  # Esto incluye dplyr y otras librerías útiles
library(SummarizedExperiment)
```

Selección de los datos

Tras revisar algunas de las opciones presentes en las bases de datos indicadas, he seleccionado los datos que corresponden a un estudio llamado 'El metagenoma del rumen y su metaboloma en los yaks asociados con el régimen de alimentación'. Podemos encontrar los datos en el siguiente enlace:

https://www.ebi.ac.uk/metabolights/editor/MTBLS10856/

He descargado los 4 archivos que se encuentran en el apartado 'Files/ISA_METADATA'. A continuación incluyo una breve descripción de cada uno de los archivos:

- **m_MTBLS10856_LC-MS_positive_hilic_metabolite_profiling_v2_maf.tsv**

Contiene la tabla de abundancias de cada metabolito en las muestras, así como otra información sobre los metabolitos como 'mass-to-charge' o 'retention time'.

- **s_MTBLS10856.txt**

Contiene información sobre las muestras como: nombre, origen, descripción, grupo al que pertenece, etc.

- **a_MTBLS10856_LC-MS_positive_hilic_metabolite_profiling.txt**

Contiene información adicional sobre las muestras. En este caso los datos estan relacionados con el procesado de las muestras: protocolos, equipos utizados, etc.

- **i_Investigation.txt**

Contiene información sobre el laboratorio y el estudio, incluyendo una descripción detallada del estudio y del protocolo utilizado.

Ahora leemos todos estos datos y los guardamos para poder trabajar con ellos. El primer archivo es de tipo tsv, por lo que utilizamos `read_tsv`, mientras que los dos siguientes son de tipo txt aunque contienen tablas en formato tsv, por lo que los podemos leer con `read.table`. En cambio, el último archivo es un archivo de texto de tipo txt que contiene textos descriptivos (no en formato tabla), por lo que utilizamos readLines.

```
abundance_data <- read_tsv("data/m_MTBLS10856_LC-MS_positive_hilic_metabolite_profiling_v2_maf.tsv", sh
sample_info <- read.table("data/s_MTBLS10856.txt", header = TRUE, sep = "\t")
additional_sample_info <- read.table("data/a_MTBLS10856_LC-MS_positive_hilic_metabolite_profiling.txt",
metadata <- readLines("data/i_Investigation.txt")
```

Vemos ahora las dimensiones y la estructura de cada elemento para comprobar que los archivos se han leído correctamente. Ahora se van a mostrar mensajes muy largos sobre cada uno de los archivos, pero es importante comprobar que se han leído correctamente:

```
# Dimensiones y estructura de cada elemento
dim(abundance_data)
```

```
## [1] 4292   40
```

```
str(abundance_data)
```

```
## spc_tbl_ [4,292 x 40] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ database_identifier                  : logi [1:4292] NA NA NA NA NA NA ...
##  $ chemical_formula                     : logi [1:4292] NA NA NA NA NA NA ...
##  $ smiles                               : logi [1:4292] NA NA NA NA NA NA ...
##  $ inchi                                : logi [1:4292] NA NA NA NA NA NA ...
##  $ metabolite_identification            : chr [1:4292] "Adenine" "Piperidine" "Hypoxanthine" "Bakken
##  $ mass_to_charge                       : num [1:4292] 136.1 86.1 137 349.2 351.2 ...
##  $ fragmentation                        : logi [1:4292] NA NA NA NA NA NA ...
##  $ modifications                        : logi [1:4292] NA NA NA NA NA NA ...
##  $ charge                               : logi [1:4292] NA NA NA NA NA NA ...
##  $ retention_time                       : num [1:4292] 141.8 283.2 190.9 81.8 134.6 ...
##  $ taxid                                : logi [1:4292] NA NA NA NA NA NA ...
##  $ species                              : logi [1:4292] NA NA NA NA NA NA ...
##  $ database                             : chr [1:4292] "BiotreeDB" "BiotreeDB" "BiotreeDB" "BiotreeD
##  $ database_version                     : logi [1:4292] NA NA NA NA NA NA ...
##  $ reliability                          : logi [1:4292] NA NA NA NA NA NA ...
##  $ uri                                  : logi [1:4292] NA NA NA NA NA NA ...
##  $ search_engine                        : logi [1:4292] NA NA NA NA NA NA ...
##  $ search_engine_score                  : logi [1:4292] NA NA NA NA NA NA ...
##  $ smallmolecule_abundance_sub          : logi [1:4292] NA NA NA NA NA NA ...
##  $ smallmolecule_abundance_stdev_sub    : logi [1:4292] NA NA NA NA NA NA ...
##  $ smallmolecule_abundance_std_error_sub: logi [1:4292] NA NA NA NA NA NA ...
##  $ YK.RM.1                              : num [1:4292] 8.19e+06 4.39e+08 1.55e+08 2.75e+07 1.65e+07
##  $ YK.RM.2                              : num [1:4292] 1.14e+07 3.11e+08 2.03e+08 2.45e+07 2.31e+07
##  $ YK.RM.3                              : num [1:4292] 2.84e+07 3.48e+08 2.35e+08 2.25e+07 2.24e+07
##  $ YK.RM.4                              : num [1:4292] 2.42e+07 3.03e+08 7.48e+07 2.16e+07 2.77e+07
##  $ YK.RM.5                              : num [1:4292] 2.36e+07 3.60e+08 1.46e+08 2.18e+07 3.03e+07
##  $ YK.RM.6                              : num [1:4292] 6.99e+06 6.11e+08 9.92e+07 3.28e+07 2.14e+07
##  $ YK.RM.7                              : num [1:4292] 1.25e+07 1.85e+08 1.28e+08 1.06e+07 4.60e+06
##  $ YK.RM.15                             : num [1:4292] 1.47e+07 1.55e+08 6.44e+07 1.94e+07 2.17e+07
##  $ YK.RM.16                             : num [1:4292] 9.42e+06 1.14e+08 1.01e+08 1.99e+07 1.96e+07
##  $ YK.RM.17                             : num [1:4292] 1.93e+07 1.80e+08 6.94e+07 1.39e+07 1.77e+07
##  $ YK.RM.18                             : num [1:4292] 2.12e+07 1.40e+08 1.27e+08 2.00e+07 1.45e+07
##  $ YK.RM.19                             : num [1:4292] 2907108 40382514 59585407 6274234 2810152 ...
##  $ YK.RM.20                             : num [1:4292] 2.08e+07 1.54e+08 1.12e+08 2.37e+07 2.32e+07
##  $ YK.RM.21                             : num [1:4292] 5.03e+07 1.32e+08 1.52e+08 2.09e+07 2.04e+07
##  $ QC1                                  : num [1:4292] 1.77e+07 4.13e+08 1.40e+08 3.06e+07 1.89e+07
##  $ QC2                                  : num [1:4292] 2.04e+07 3.09e+08 1.39e+08 2.28e+07 1.47e+07
##  $ QC3                                  : num [1:4292] 1.99e+07 2.79e+08 1.15e+08 2.29e+07 1.21e+07
##  $ QC4                                  : num [1:4292] 1.94e+07 2.78e+08 1.14e+08 2.17e+07 1.76e+07
##  $ QC5                                  : num [1:4292] 1.98e+07 2.83e+08 1.37e+08 2.21e+07 1.37e+07
##  - attr(*, "spec")=
##   .. cols(
##   ..   database_identifier = col_logical(),
##   ..   chemical_formula = col_logical(),
##   ..   smiles = col_logical(),
##   ..   inchi = col_logical(),
##   ..   metabolite_identification = col_character(),
##   ..   mass_to_charge = col_double(),
##   ..   fragmentation = col_logical(),
##   ..   modifications = col_logical(),
```

```
##   ..   charge = col_logical(),
##   ..   retention_time = col_double(),
##   ..   taxid = col_logical(),
##   ..   species = col_logical(),
##   ..   database = col_character(),
##   ..   database_version = col_logical(),
##   ..   reliability = col_logical(),
##   ..   uri = col_logical(),
##   ..   search_engine = col_logical(),
##   ..   search_engine_score = col_logical(),
##   ..   smallmolecule_abundance_sub = col_logical(),
##   ..   smallmolecule_abundance_stdev_sub = col_logical(),
##   ..   smallmolecule_abundance_std_error_sub = col_logical(),
##   ..   YK.RM.1 = col_double(),
##   ..   YK.RM.2 = col_double(),
##   ..   YK.RM.3 = col_double(),
##   ..   YK.RM.4 = col_double(),
##   ..   YK.RM.5 = col_double(),
##   ..   YK.RM.6 = col_double(),
##   ..   YK.RM.7 = col_double(),
##   ..   YK.RM.15 = col_double(),
##   ..   YK.RM.16 = col_double(),
##   ..   YK.RM.17 = col_double(),
##   ..   YK.RM.18 = col_double(),
##   ..   YK.RM.19 = col_double(),
##   ..   YK.RM.20 = col_double(),
##   ..   YK.RM.21 = col_double(),
##   ..   QC1 = col_double(),
##   ..   QC2 = col_double(),
##   ..   QC3 = col_double(),
##   ..   QC4 = col_double(),
##   ..   QC5 = col_double()
##   .. )
##   - attr(*, "problems")=<externalptr>
```

```
dim(sample_info)
```

```
## [1] 19 18
```

```
str(sample_info)
```

```
## 'data.frame':    19 obs. of  18 variables:
##  $ Source.Name                   : chr  "QC01.mzXML" "QC02.mzXML" "QC03.mzXML" "QC04.mzXML" ...
##  $ Characteristics.Organism.     : chr  "Bos grunniens" "Bos grunniens" "Bos grunniens" "Bos grunnie
##  $ Term.Source.REF               : chr  "NCBITaxon" "NCBITaxon" "NCBITaxon" "NCBITaxon" ...
##  $ Term.Accession.Number         : chr  "http://purl.obolibrary.org/obo/NCBITaxon_30521" "http://pur
##  $ Characteristics.Organism.part.: chr  "ruminal fluid" "ruminal fluid" "ruminal fluid" "ruminal flu
##  $ Term.Source.REF.1             : chr  "BTO" "BTO" "BTO" "BTO" ...
##  $ Term.Accession.Number.1       : chr  "http://purl.obolibrary.org/obo/BTO_0004789" "http://purl.ob
##  $ Characteristics.Variant.      : chr  "Zhongdian" "Zhongdian" "Zhongdian" "Zhongdian" ...
##  $ Term.Source.REF.2             : logi  NA NA NA NA NA NA ...
##  $ Term.Accession.Number.2       : logi  NA NA NA NA NA NA ...
##  $ Characteristics.Sample.type.  : chr  "pooled quality control sample" "pooled quality control samp
##  $ Term.Source.REF.3             : chr  "" "" "" "" ...
##  $ Term.Accession.Number.3       : chr  "" "" "" "" ...
```

```
##  $ Protocol.REF                  : chr  "Sample collection" "Sample collection" "Sample collection" "
##  $ Sample.Name                   : chr  "QC1" "QC2" "QC3" "QC4" ...
##  $ Factor.Value.Diet.            : chr  "QC" "QC" "QC" "QC" ...
##  $ Term.Source.REF.4             : logi  NA NA NA NA NA NA ...
##  $ Term.Accession.Number.4       : logi  NA NA NA NA NA NA ...
```

```
dim(additional_sample_info)
```

```
## [1] 19 37
```

```
str(additional_sample_info)
```

```
## 'data.frame':    19 obs. of  37 variables:
##  $ Sample.Name                                : chr  "QC1" "QC2" "QC3" "QC4" ...
##  $ Protocol.REF                               : chr  "Extraction" "Extraction" "Extraction" "Extracti
##  $ Parameter.Value.Post.Extraction.           : chr  "acetonitrile:methanol (1:1, v/v)" "acetonitrile
##  $ Parameter.Value.Derivatization.            : logi  NA NA NA NA NA NA ...
##  $ Extract.Name                               : chr  "QC1" "QC2" "QC3" "QC4" ...
##  $ Protocol.REF.1                             : chr  "Chromatography" "Chromatography" "Chromatograph
##  $ Parameter.Value.Chromatography.Instrument.: chr  "Thermo Scientific Vanquish UHPLC System" "Therm
##  $ Term.Source.REF                            : chr  "MTBLS" "MTBLS" "MTBLS" "MTBLS" ...
##  $ Term.Accession.Number                      : chr  "http://www.ebi.ac.uk/metabolights/ontology/MTBL
##  $ Parameter.Value.Autosampler.model.         : logi  NA NA NA NA NA NA ...
##  $ Parameter.Value.Column.model.              : chr  "XBridge BEH Amide (1.7 µm, 2.1 mm x 100 mm; Wat
##  $ Parameter.Value.Column.type.               : chr  "HILIC" "HILIC" "HILIC" "HILIC" ...
##  $ Parameter.Value.Guard.column.              : logi  NA NA NA NA NA NA ...
##  $ Labeled.Extract.Name                       : logi  NA NA NA NA NA NA ...
##  $ Label                                      : logi  NA NA NA NA NA NA ...
##  $ Term.Source.REF.1                          : logi  NA NA NA NA NA NA ...
##  $ Term.Accession.Number.1                    : logi  NA NA NA NA NA NA ...
##  $ Protocol.REF.2                             : chr  "Mass spectrometry" "Mass spectrometry" "Mass sp
##  $ Parameter.Value.Scan.polarity.             : chr  "positive" "positive" "positive" "positive" ...
##  $ Parameter.Value.Scan.m.z.range.            : chr  "100-1000" "100-1000" "100-1000" "100-1000" ...
##  $ Parameter.Value.Instrument.                : chr  "Thermo Q Exactive HFX" "Thermo Q Exactive HFX"
##  $ Term.Source.REF.2                          : logi  NA NA NA NA NA NA ...
##  $ Term.Accession.Number.2                    : logi  NA NA NA NA NA NA ...
##  $ Parameter.Value.Ion.source.                : chr  "electrospray ionization" "electrospray ionizati
##  $ Term.Source.REF.3                          : logi  NA NA NA NA NA NA ...
##  $ Term.Accession.Number.3                    : logi  NA NA NA NA NA NA ...
##  $ Parameter.Value.Mass.analyzer.             : chr  "orbitrap" "orbitrap" "orbitrap" "orbitrap" ...
##  $ Term.Source.REF.4                          : logi  NA NA NA NA NA NA ...
##  $ Term.Accession.Number.4                    : logi  NA NA NA NA NA NA ...
##  $ MS.Assay.Name                              : chr  "QC1" "QC2" "QC3" "QC4" ...
##  $ Raw.Spectral.Data.File                     : chr  "FILES/QC01.raw" "FILES/QC02.raw" "FILES/QC03.ra
##  $ Protocol.REF.3                             : chr  "Data transformation" "Data transformation" "Data
##  $ Normalization.Name                         : logi  NA NA NA NA NA NA ...
##  $ Derived.Spectral.Data.File                 : logi  NA NA NA NA NA NA ...
##  $ Protocol.REF.4                             : chr  "Metabolite identification" "Metabolite identifi
##  $ Data.Transformation.Name                   : logi  NA NA NA NA NA NA ...
##  $ Metabolite.Assignment.File                 : chr  "m_MTBLS10856_LC-MS_positive_hilic_metabolite_pr
```

Creación del contenedor `SummarizedExperiment`

Preparación

Ahora vamoos a crear el contenedor para los datos con los que estamos trabajando, para ello, en primer lugar tenemos que preparar cada uno de los diferentes elementos por separado:

Assay

Empezamos por el elemento `Assay`, que es una matriz que tiene como filas los metabolitos y como columnas las diferentes muestras. Para crearlo, partimos de los datos que hemos guardado como `abundance_data`. A partir de este archivo, creamos una matriz para la cual seleccionamos las columnas que contienen los datos de abundancia de los metabolitos para cada muestra. Seguimos los siguientes pasos:

```
## Assay
# Guardar los nombres de las muestras
sample_names <- sample_info$Sample.Name
# Crea la matriz con todas las filas y solo las columnas correspondientes a las muestras
abundance_matrix <- as.matrix(abundance_data[, sample_names])
# Nombra las filas con el nombre del metabolito correspondiente
rownames(abundance_matrix) <- abundance_data$metabolite_identification
```

rowData

Este elemento contiene información sobre las filas, que en este caso son metabolitos, por lo tanto vamos a utilizar el resto de la información que encontramos en `abundance_data`. Para ello, eliminamos las columnas que no contienen información, es decir, que el 100% de sus casillas contienen y también las columnas que contienen los datos de abundancia que hemos tratado en el apartado anterior:

```
## rowData
# Eliminamos las columnas que no tienen información (100% de NA)
cols_borrar <- which(colMeans(is.na(abundance_data)) == 1)
rowData <- data.frame(abundance_data[, -cols_borrar])
# Eliminamos las columnas que tienen la información de abundancias (guardada en Assay)
rowData <- rowData[, 1:(ncol(rowData)-length(sample_names))]
```

En este caso, podríamos eliminar también la columna que corresponde a `abundance_data$metabolite_identification`, ya que hemos incluido esta información como nombres de las filas, por lo que estará almacenada de otra manera, pero puede que tenerla también en una de estas columnas sea útil para algún análisis futuro en el caso de que se descarguen los datos y, además, el elemento tiene pocas columnas, por lo que tampoco supone una molestia para la lectura de la información.

colData

En este apartado encontramos información sobre las muestras (que se encuentran en las columnas, de ahí el nombre). Tenemos 2 archivos que contienen información sobre las muestras y podríamos incluirlos tal cual estan, ya que los hemos cargado en formato data frame, pero también podemos 'limpiarlos', ya que hay muchas columnas que no contienen ninguna información. Procesamos estos datos a continuación:

```
## colData
# Eliminamos las columnas que no tienen información (100% de NA)
cols_borrar <- which(colMeans(is.na(sample_info)) == 1)
sample_info <- sample_info[, -cols_borrar]

cols_borrar <- which(colMeans(is.na(additional_sample_info)) == 1)
additional_sample_info <- additional_sample_info[, -cols_borrar]
```

Metadata

Este último apartado contiene información relativa a los métodos experimentales y referencias a la publicación de los datos. Hemos cargado estos datos en `metadata`. Se trata de un archivo de texto, por lo que hemos leído el archivo por líneas, que se han almacenado en forma de lista. Esta lista es lo que a continuación utilizaremos en la creación del contenedor.

Contenedor `SummarizedExperiment`

Ahora que ya hemos preparado todos los elementos, podemos crear con ellos el contenedor:

```
# CREA EL CONTENEDOR SummarizedExperiment
se <- SummarizedExperiment(assays = list(counts = abundance_matrix),
                           rowData = rowData,
                           colData = list(sample_info = sample_info, additional_sample_info = additional
                           metadata = list(metadata = metadata)
                           )
```

Exploración del dataset

En primer lugar, comprobamos que se ha creado correctamente:

```
# Revisamos el objeto SummarizedExperiment
se
```

```
## class: SummarizedExperiment
## dim: 4292 19
## metadata(1): metadata
## assays(1): counts
## rownames(4292): Adenine Piperidine ... Unknown Unknown
## rowData names(4): metabolite_identification mass_to_charge
##   retention_time database
## colnames(19): QC1 QC2 ... YK.RM.20 YK.RM.21
## colData names(35): sample_info.Source.Name
##   sample_info.Characteristics.Organism. ...
##   additional_sample_info.Protocol.REF.4
##   additional_sample_info.Metabolite.Assignment.File
```

Y ahora podemos revisar los diferentes elementos que contiene:

En primer lugar, podemos ver los datos que contiene `assay()`, pero se trata de una matriz de grandes dimensiones, ya que contiene `dim(assay(se))[1]` filas y `dim(assay(se))[2]` columnas, por lo que vamos a ver las primeras y las últimas filas:

```
head(assay(se))
```

```
##                         QC1       QC2       QC3       QC4       QC5   YK.RM.1
## Adenine              17704093  20425604  19916131  19406499  19768411   8192877
## Piperidine          412733126 309444090 279134741 277655754 283405481 439238374
## Hypoxanthine        139739042 138815058 114573540 113772679 136798975 154511975
## Bakkenolide C        30633540  22797625  22892299  21736227  22096390  27505652
## 8-iso-15-keto-PGE2   18930898  14662216  12083264  17608652  13655171  16504050
## Histamine           241389854 210273821 207618191 202990998 204645416 217166617
##                       YK.RM.2   YK.RM.3   YK.RM.4   YK.RM.5   YK.RM.6   YK.RM.7
## Adenine              11430196  28434202  24154076  23581498   6988687  12510584
## Piperidine          311177079 348327001 303167817 360319673 610548254 185183396
## Hypoxanthine        202648977 234858796  74784826 146002114  99171564 127701781
```

```
## Bakkenolide C       24467816  22524075  21631996  21789401  32767883  10597186
## 8-iso-15-keto-PGE2  23052916  22405922  27685791  30331392  21352452   4597619
## Histamine          192810642 204055997 111249109 220885898  77738235 158769939
##                     YK.RM.15  YK.RM.16  YK.RM.17  YK.RM.18 YK.RM.19  YK.RM.20
## Adenine             14720360   9416370  19291767  21187703  2907108  20788579
## Piperidine         154968182 114425681 180484179 139793775 40382514 153716655
## Hypoxanthine        64411386 101060711  69380944 126567057 59585407 111971920
## Bakkenolide C       19435529  19922187  13851027  19950255  6274234  23726021
## 8-iso-15-keto-PGE2  21655881  19595766  17748350  14526095  2810152  23192672
## Histamine           58694413 148839099  82330908 164881888 33200796 119286878
##                     YK.RM.21
## Adenine             50332023
## Piperidine         132475527
## Hypoxanthine       152146073
## Bakkenolide C       20888830
## 8-iso-15-keto-PGE2  20388048
## Histamine          239923264
```

```
tail(assay(se))
```

```
##                QC1         QC2         QC3          QC4         QC5
## Unknown 121586505.9 104392650.0 133378505.8 116367264.50 109053027.1
## Unknown    571661.2    225633.8    103651.6     88013.09    308704.5
## Unknown   3870544.2   5004404.5   2810223.8   1989312.46   2274358.6
## Unknown  34409276.4   6660163.7   1877470.8   1355630.92    220377.0
## Unknown   1522314.8    224822.5   2461815.9   2191634.36   2436016.2
## Unknown   5094533.2   6991958.8   6720314.5   4389050.85   6738737.2
##              YK.RM.1   YK.RM.2     YK.RM.3    YK.RM.4    YK.RM.5     YK.RM.6
## Unknown 102354598.5 118278059 101903640.4 95932121.8 90939437.0 102018445.3
## Unknown    377022.2         0    265771.0   213286.2   335143.0    196928.2
## Unknown   2565899.4   3160985   3411453.2  7605131.7  3707944.6   5435504.5
## Unknown   2925423.5   3796948    775806.9  1049759.3   575080.2   1535166.3
## Unknown   3142021.1   3527557   2799799.4  1439786.2  1600298.6   2629916.1
## Unknown   6403831.7   4863711   7261671.7  4300590.8  6653774.9   9231763.9
##            YK.RM.7    YK.RM.15    YK.RM.16  YK.RM.17    YK.RM.18    YK.RM.19
## Unknown 166921268 107944654.5 169206641.5 128261927 155252639.8 174945630.4
## Unknown         0    112589.4         0.0         0    562720.7         0.0
## Unknown   1290471   1680363.8    833068.1   1077769   1024608.8    182743.6
## Unknown   1175738   3613532.0    405695.5   2309073   1312692.3    776143.9
## Unknown   1609736   3177557.4   2518599.1   2555069   2318779.2   2850691.6
## Unknown  12287575   5002980.1   7731036.9   7845086   6165698.7   8166188.2
##           YK.RM.20   YK.RM.21
## Unknown 114818939 93152846.6
## Unknown         0    76890.9
## Unknown   2523052  1179120.0
## Unknown   1830417  1156443.9
## Unknown   2280116  1557842.5
## Unknown   8790728  5542942.4
```

Vemos que al principio del documento encontramos los nombres de los metabolitos como nombres de las filas, aunque en los últimos los metabolitos son desconocidos ('Unknown').

También podemos ver las primeras filas de la información que tenemos sobre los metabolitos:

```
head(rowData(se))
```

```
## DataFrame with 6 rows and 4 columns
##                     metabolite_identification mass_to_charge retention_time
##                                   <character>      <numeric>      <numeric>
## Adenine                               Adenine        136.062       141.8480
## Piperidine                         Piperidine         86.097       283.2475
## Hypoxanthine                     Hypoxanthine        137.046       190.8525
## Bakkenolide C                  Bakkenolide C        349.199        81.8224
## 8-iso-15-keto-PGE2        8-iso-15-keto-PGE2        351.214       134.5720
## Histamine                           Histamine        112.087       401.7700
##                       database
##                    <character>
## Adenine              BiotreeDB
## Piperidine           BiotreeDB
## Hypoxanthine         BiotreeDB
## Bakkenolide C        BiotreeDB
## 8-iso-15-keto-PGE2   BiotreeDB
## Histamine            BiotreeDB
```

Y lo mismo sobre las muestras, aunque selecciono solo algunas columnas porque hay mucha información y ocupa mucho en el documento:

```
head(colData(se)[1:5])
```

```
## DataFrame with 6 rows and 5 columns
##         sample_info.Source.Name sample_info.Characteristics.Organism.
##                     <character>                            <character>
## QC1                  QC01.mzXML                          Bos grunniens
## QC2                  QC02.mzXML                          Bos grunniens
## QC3                  QC03.mzXML                          Bos grunniens
## QC4                  QC04.mzXML                          Bos grunniens
## QC5                  QC05.mzXML                          Bos grunniens
## YK.RM.1            YK-R-1.mzXML                          Bos grunniens
##         sample_info.Term.Source.REF sample_info.Term.Accession.Number
##                         <character>                          <character>
## QC1                       NCBITaxon             http://purl.obolibra..
## QC2                       NCBITaxon             http://purl.obolibra..
## QC3                       NCBITaxon             http://purl.obolibra..
## QC4                       NCBITaxon             http://purl.obolibra..
## QC5                       NCBITaxon             http://purl.obolibra..
## YK.RM.1                   NCBITaxon             http://purl.obolibra..
##         sample_info.Characteristics.Organism.part.
##                                       <character>
## QC1                                 ruminal fluid
## QC2                                 ruminal fluid
## QC3                                 ruminal fluid
## QC4                                 ruminal fluid
## QC5                                 ruminal fluid
## YK.RM.1                             ruminal fluid
```

También podemos ver el metadata:

```
head(metadata(se))
```

```
## $metadata
##  [1] "ONTOLOGY SOURCE REFERENCE"
##  [2] "Term Source Name\tEDAM\tNCIT\tOBI\tEFO\tVBO\tNCBITaxon"
##  [3] "Term Source File\t\t\thttp://data.bioontology.org/ontologies/OBI\t\thttps://www.ebi.ac.uk/ols4/
```

```
##  [4] "Term Source Version\t\t\t29\t\t\t"
##  [5] "Term Source Description\t\t\tOntology for Biomedical Investigations\t\t\t"
##  [6] "INVESTIGATION"
##  [7] "Investigation Identifier\tMTBLS10856"
##  [8] "Investigation Title\tInvestigation"
##  [9] "Investigation Description\tCreated using the MetaboLights Online Editor (MOE)"
## [10] "Investigation Submission Date\t2024-08-08"
## [11] "Investigation Public Release Date\t2024-10-01"
## [12] "Comment[Created With Configuration]\tMetaboLightsConfig20150707"
## [13] "Comment[Last Opened With Configuration]\tMetaboLightsConfig20150707"
## [14] "INVESTIGATION PUBLICATIONS"
## [15] "Investigation PubMed ID"
## [16] "Investigation Publication DOI"
## [17] "Investigation Publication Author List"
## [18] "Investigation Publication Title"
## [19] "Investigation Publication Status"
## [20] "Investigation Publication Status Term Accession Number"
## [21] "Investigation Publication Status Term Source REF"
## [22] "INVESTIGATION CONTACTS"
## [23] "Investigation Person Last Name"
## [24] "Investigation Person First Name"
## [25] "Investigation Person Mid Initials"
## [26] "Investigation Person Email"
## [27] "Investigation Person Phone"
## [28] "Investigation Person Fax"
## [29] "Investigation Person Address"
## [30] "Investigation Person Affiliation"
## [31] "Investigation Person Roles"
## [32] "Investigation Person Roles Term Accession Number"
## [33] "Investigation Person Roles Term Source REF"
## [34] "STUDY"
## [35] "Study Identifier\tMTBLS10856"
## [36] "Study Title\tThe rumen metagenome and its metabolome in yaks associated with feeding regime"
## [37] "Study Description\t<p>Background: Grazing yearly on pasture is a traditional practice for yaks
## [38] "Study Submission Date\t2024-08-08"
## [39] "Study Public Release Date\t2024-10-01"
## [40] "Study File Name\ts_MTBLS10856.txt"
## [41] "STUDY DESIGN DESCRIPTORS"
## [42] "Study Design Type\tMetabolomics\tRumen\tBos"
## [43] "Study Design Type Term Accession Number\thttp://edamontology.org/topic_3172\thttp://purl.oboli
## [44] "Study Design Type Term Source REF\tEDAM\tNCIT\tNCBITaxon"
## [45] "STUDY PUBLICATIONS"
## [46] "Study PubMed ID\t"
## [47] "Study Publication DOI\t"
## [48] "Study Publication Author List\tShuli Yang, Jieyi Zheng, Huaming Mao, Dongwang Wu, Jianmin Chai
## [49] "Study Publication Title\tMulti-omic dataset of the rumen metagenome and its metabolome and the
## [50] "Study Publication Status\tIn preparation"
## [51] "Study Publication Status Term Accession Number\thttp://www.ebi.ac.uk/efo/EFO_0001795"
## [52] "Study Publication Status Term Source REF\tEFO"
## [53] "STUDY FACTORS"
## [54] "Study Factor Name\tDiet"
## [55] "Study Factor Type\tDiet"
## [56] "Study Factor Type Term Accession Number\thttp://purl.obolibrary.org/obo/NCIT_C15222"
## [57] "Study Factor Type Term Source REF\tNCIT"
```

```
## [58] "STUDY ASSAYS"
## [59] "Study Assay File Name\ta_MTBLS10856_LC-MS_positive_hilic_metabolite_profiling.txt"
## [60] "Study Assay Measurement Type\tmetabolite profiling"
## [61] "Study Assay Measurement Type Term Accession Number\thttp://purl.obolibrary.org/obo/OBI_0000366
## [62] "Study Assay Measurement Type Term Source REF\tOBI"
## [63] "Study Assay Technology Type\tmass spectrometry"
## [64] "Study Assay Technology Type Term Accession Number\thttp://purl.obolibrary.org/obo/OBI_0000470"
## [65] "Study Assay Technology Type Term Source REF\tOBI"
## [66] "Study Assay Technology Platform\tLiquid Chromatography MS - positive - hilic"
## [67] "STUDY PROTOCOLS"
## [68] "Study Protocol Name\tSample collection\tExtraction\tChromatography\tMass spectrometry\tData tra
## [69] "Study Protocol Type\tSample collection\tExtraction\tChromatography\tMass spectrometry\tData tra
## [70] "Study Protocol Type Term Accession Number\t\t\t\t\t\t"
## [71] "Study Protocol Type Term Source REF\t\t\t\t\t\t"
## [72] "Study Protocol Description\t<p>A total of 50 healthy Zhongdian yaks aged 3 to 4 years old were
## [73] "Study Protocol URI\t\t\t\t\t\t"
## [74] "Study Protocol Version\t\t\t\t\t\t"
## [75] "Study Protocol Parameters Name\t\tPost Extraction;Derivatization\tChromatography Instrument;Au
## [76] "Study Protocol Parameters Name Term Accession Number\t\t;\t;;;;\t;;;;\t\t"
## [77] "Study Protocol Parameters Name Term Source REF\t\t;\t;;;;\t;;;;\t\t"
## [78] "Study Protocol Components Name\t\t\t\t\t\t"
## [79] "Study Protocol Components Type\t\t\t\t\t\t"
## [80] "Study Protocol Components Type Term Accession Number\t\t\t\t\t\t"
## [81] "Study Protocol Components Type Term Source REF\t\t\t\t\t\t"
## [82] "STUDY CONTACTS"
## [83] "Study Person Last Name\tchai"
## [84] "Study Person First Name\tjianmin"
## [85] "Study Person Mid Initials\t"
## [86] "Study Person Email\tjchai@uark.edu"
## [87] "Study Person Phone\t4793011636"
## [88] "Study Person Fax\t"
## [89] "Study Person Address\t#33 guangyun road, foshan"
## [90] "Study Person Affiliation\t"
## [91] "Study Person Roles\tPrincipal Investigator"
## [92] "Study Person Roles Term Accession Number\thttp://purl.obolibrary.org/obo/NCIT_C19924"
## [93] "Study Person Roles Term Source REF\tNCIT"
```

Viendo todos estos datos, realizamos un análisis del estudio similar al realizado en tareas anteriores. Vemos diferentes aspectos:

Pregunta biológica

La pregunta clave fue cómo el sistema de alimentación impacta sobre el microbioma ruminal, los metabolitos ruminales y el metaboloma del hospedador en yaks, y cómo esto influye en el crecimiento de los animales. La exploración de esto podría tener implicaciones para mejorar la eficiencia de producción en la industria de los yaks, optimizando su nutrición y salud a través de prácticas de alimentación más controladas.

Diseño experimental

El diseño experimental muestra dos grupos de tratamiento: uno que pastorea libremente (grupo de pastoreo, identificado como QC) y otro alimentado bajo un sistema intensivo (grupo intensivo, identificado como YK.RM). Las diferencias en la dieta y las condiciones de recolección de muestras podrían influir en la interpretación de los resultados, como la variabilidad en el microbioma ruminal y los metabolitos.

Obtención de datos crudos

La elección de técnicas tanto de recolección como de análisis, por ejemplo, el uso de LC-MS/MS para metabolómica, puede influir en la precisión y tipo de datos obtenidos, como la detección de metabolitos específicos. La técnica influye en la capacidad de detectar metabolitos volátiles o componentes microbiológicos que varían según el tratamiento.

Control de Calidad, Preprocesado y Normalización

Un desafío en la calidad de los datos crudos puede ser la variabilidad en las muestras, especialmente en metabolómica, debido a la heterogeneidad biológica de los animales. El preprocesamiento podría implicar la normalización de los datos de metabolitos y la corrección de cualquier sesgo derivado de la técnica o del manejo de las muestras.

Respuesta a la Pregunta Biológica

Los hallazgos clave incluyen el aumento en la concentración de ácidos grasos volátiles (VFA) y el crecimiento mejorado en los yaks bajo el sistema intensivo, así como cambios en el microbioma ruminal. Estos resultados responden a la pregunta biológica mostrando que el sistema de alimentación intensiva mejora la producción y la eficiencia en los yaks. Las implicaciones más amplias incluyen la potencial aplicación de este sistema de alimentación para mejorar la productividad en la industria de los yaks.

Repositorio GitHub

Preparación de los archivos

Para el repositorio necesito los siguientes elementos:

- El informe (en formato pdf)

Lo obtengo a partir de la compilación de este archivo. Llamado 'Montoliu-Silvestre-Eva-PEC1.pdf'.

- El objeto contenedor con los datos y los metadatos en formato binario

Guardamos el archivo en formato `.Rda`:

```
# Guardar el objeto 'SummarizedExperiment' en formato .Rda
save(se, file = "data/se_dataset.Rda")
```

- El código R para la exploración de los datos

Lo encontramos en el archivo 'Montoliu-Silvestre-Eva-PEC1.R'

- Los datos en formato texto

4 archivos en formato tsv o txt.

- Los metadatos acerca del dataset en un archivo markdown.

A continuación guardo los metadatos en formato `.md`.

```
# Guardar metadata en formato .md
writeLines(metadata, "data/metadata.md")
```

Creación del repositorio

Ahora que tenemos todos los archivos necesarios, podemos crear el repositorio, que encontramos en el siguiente enlace:

https://github.com/evamontolius/Montoliu-Silvestre-Eva-PEC1-ADO.git