Ferrando-León-Miguel-PEC1

Miguel Ferrando León

2024-11-06

En primer lugar, descargamos unos datos desde el repositorio de github proporcionado. Para ello, accedemos a la carpeta "datasets" y nos dirigimos a la subcarpeta que nos interesa "2018-MetabotypingPaper". De ahí descargamos tanto los datos (DataValues_S013.csv), como los metadatos (DataInfo_S013.csv).

A continuación, insertamos y definimos los dos archivos de datos en mi script de R:

```r
metabodata <-
read.csv("/Users/miguelferrando/Downloads/DataValues_S013.csv")
metaboinfo <-
read.csv("/Users/miguelferrando/Downloads/DataInfo_S013.csv")
```

Es preferible ajustar la matriz de metabolitos para quedarme solo con las columnas que contienen los valores numéricos. También la trasponemos para cambiar filas por columnas y así ponemos los valores metabolómicos y sus tiempos como filas y los diferentes pacientes como columnas:

```r
datavalues <- t(metabodata[, 7:dim(metabodata)[2]])
```

Con head, imprimimos la tabla para ver de manera sintetizada cómo ha quedado:

```r
head(datavalues)
```

```
##           [,1] [,2]  [,3]  [,4]  [,5]  [,6] [,7] [,8]  [,9] [,10] [,11]
[,12]
## MEDDM_T0   0.0  0.0  0.00   0.0  0.00  0.00  0.0  0.0  0.00  0.00   0.0
0.0
## MEDCOL_T0  0.0  0.0  0.00   0.0  0.00  0.00  0.0  0.0  0.00  0.00   0.0
0.0
## MEDINF_T0  0.0  0.0  0.00   0.0  0.00  0.00  0.0  1.0  0.00  0.00   1.0
0.0
## MEDHTA_T0  1.0  0.0  0.00   0.0  0.00  0.00  0.0  0.0  0.00  0.00   0.0
0.0
## GLU_T0    85.0 78.0 75.00  71.0 82.00 71.00 80.0 90.0 92.00 84.00  75.0
108.0
## INS_T0    11.4 12.1  8.41  12.8  6.01  9.88  9.2  3.4  5.43  6.98  13.3
16.8
##          [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22]
[,23]
## MEDDM_T0    0.0   0.0   0.0     0   0.0   0.0   0.0     0     0   0.0
0.0
## MEDCOL_T0   0.0   0.0   0.0     0   0.0   0.0   0.0     0     0   0.0
0.0
## MEDINF_T0   0.0   0.0   0.0     0   0.0   0.0   0.0     1     0   0.0
```

```
0.0
## MEDHTA_T0    1.0   1.0   1.0     1   0.0   0.0   0.0     0     0   0.0
0.0
## GLU_T0     101.0 105.0 139.0   106 159.0 103.0 106.0   107   127 111.0
141.0
## INS_T0      17.1  21.3  36.6    20  17.6  29.5  13.3    15    15  12.2
32.3
##            [,24] [,25] [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33]
[,34]
## MEDDM_T0      0   0.0    NA   0.0   0.0   0.0     0     0   0.0   0.0
0.0
## MEDCOL_T0     0   0.0    NA   0.0   1.0   0.0     0     0   0.0   0.0
0.0
## MEDINF_T0     1   0.0    NA   0.0   0.0   0.0     0     0   0.0   0.0
0.0
## MEDHTA_T0     0   0.0    NA   0.0   1.0   0.0     0     0   0.0   1.0
1.0
## GLU_T0      100 100.0 100.0 100.0 117.0 100.0   263   115 108.0 114.0
101.0
## INS_T0       16  12.8  11.1  19.6  11.6  13.7    21    19  23.1  27.8
23.7
##            [,35] [,36] [,37] [,38] [,39]
## MEDDM_T0    0.0   0.0     0   0.0   0.0
## MEDCOL_T0   0.0   0.0     0   0.0   0.0
## MEDINF_T0   0.0   0.0     1   0.0   0.0
## MEDHTA_T0   1.0   0.0     0   0.0   0.0
## GLU_T0    108.0 106.0   115 102.0 108.0
## INS_T0     17.7  16.1    43  21.9  42.7
```

Luego creamos un nuevo objeto desde el mismo dataset metabodata, para separar y aglutinar las columnas donde tenemos información complementaria como el género, el tipo de cirugía, etc. En paralelo, reducimos la matriz de metadatos metaboinfo para borrar duplicados. Gracias a estOs ajustes, podemos insertar todo de manera ordenada en el contenedor. Por tanto ajustamos y comprobamos las tablas:

```
coldata <- metabodata [,2:6]
rowdata <- metaboinfo [6:dim(metaboinfo)[1], 2:dim(metaboinfo)[2]]

head(coldata)

##   SUBJECTS SURGERY AGE GENDER Group
## 1        1 by pass  27      F     1
## 2        2 by pass  19      F     2
## 3        3 by pass  42      F     1
## 4        4 by pass  37      F     2
## 5        5 tubular  42      F     1
## 6        6 by pass  24      F     2

head(rowdata)
```

```
##        VarName  varTpe Description
## 6    MEDDM_T0 integer    dataDesc
## 7   MEDCOL_T0 integer    dataDesc
## 8   MEDINF_T0 integer    dataDesc
## 9   MEDHTA_T0 integer    dataDesc
## 10    GLU_T0 integer    dataDesc
## 11    INS_T0 numeric    dataDesc
```

Antes de crear el contenedor, definimos el objeto SummarizedExperiment cargando la librería en cuestión:

```
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, 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

## 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
```

Ahora creamos el contenedor SummarizedExperiment con los datos y metadatos como se pide en el ejercicio:

```
se <- SummarizedExperiment(assays = list(datavalues), colData = coldata,
rowData = rowdata)
```

Llamamos a class para comprobar nuestro contenedor:

```
class(se)
```

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

Ahora procedemos con una exploración general de los datos para hacernos una idea superficial de lo que tenemos y así poder realizar posteriormente un análisis con más detalle.

Comenzamos con una medida de la dimensionalidad de la matriz de datos con la que vamos a trabajar:

```
dim(datavalues)
```

```
## [1] 690  39
```

Repetimos head como parte de nuestro trabajo exploratorio para ver las 6 primeras filas:

```
head(datavalues)
```

```
##           [,1] [,2]  [,3] [,4]  [,5]  [,6] [,7] [,8]  [,9] [,10] [,11]
[,12]
## MEDDM_T0    0.0  0.0  0.00  0.0  0.00  0.00  0.0  0.0  0.00  0.00   0.0
0.0
## MEDCOL_T0   0.0  0.0  0.00  0.0  0.00  0.00  0.0  0.0  0.00  0.00   0.0
0.0
## MEDINF_T0   0.0  0.0  0.00  0.0  0.00  0.00  0.0  1.0  0.00  0.00   1.0
0.0
## MEDHTA_T0   1.0  0.0  0.00  0.0  0.00  0.00  0.0  0.0  0.00  0.00   0.0
0.0
## GLU_T0     85.0 78.0 75.00 71.0 82.00 71.00 80.0 90.0 92.00 84.00  75.0
108.0
## INS_T0     11.4 12.1  8.41 12.8  6.01  9.88  9.2  3.4  5.43  6.98  13.3
16.8
##           [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22]
[,23]
## MEDDM_T0    0.0   0.0   0.0     0   0.0   0.0   0.0     0     0   0.0
0.0
## MEDCOL_T0   0.0   0.0   0.0     0   0.0   0.0   0.0     0     0   0.0
0.0
## MEDINF_T0   0.0   0.0   0.0     0   0.0   0.0   0.0     1     0   0.0
0.0
## MEDHTA_T0   1.0   1.0   1.0     1   0.0   0.0   0.0     0     0   0.0
0.0
## GLU_T0    101.0 105.0 139.0   106 159.0 103.0 106.0   107   127 111.0
141.0
## INS_T0     17.1  21.3  36.6    20  17.6  29.5  13.3    15    15  12.2
32.3
##           [,24] [,25] [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33]
```

```
[,34]
## MEDDM_T0      0    0.0    NA    0.0    0.0    0.0      0      0    0.0    0.0
0.0
## MEDCOL_T0     0    0.0    NA    0.0    1.0    0.0      0      0    0.0    0.0
0.0
## MEDINF_T0     1    0.0    NA    0.0    0.0    0.0      0      0    0.0    0.0
0.0
## MEDHTA_T0     0    0.0    NA    0.0    1.0    0.0      0      0    0.0    1.0
1.0
## GLU_T0      100  100.0  100.0  100.0  117.0  100.0    263    115  108.0  114.0
101.0
## INS_T0       16   12.8   11.1   19.6   11.6   13.7     21     19   23.1   27.8
23.7
##          [,35] [,36] [,37] [,38] [,39]
## MEDDM_T0    0.0    0.0      0    0.0    0.0
## MEDCOL_T0   0.0    0.0      0    0.0    0.0
## MEDINF_T0   0.0    0.0      1    0.0    0.0
## MEDHTA_T0   1.0    0.0      0    0.0    0.0
## GLU_T0    108.0  106.0    115  102.0  108.0
## INS_T0     17.7   16.1     43   21.9   42.7
```

Continuamos con la exploración de los metadatos también, para ver información adicional sobre nuestros datos metabolómicos. Primero vemos los nombres de las columnas de coldata:

```
colnames(coldata)
```

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

Y de rowdata:

```
colnames(rowdata)
```

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

Seguimos investigando y ahora cuantificamos metabolitos con la función length, y para obtener el número total real dividimos la cantidad total por los 5 tiempos de muestreo:

```
length(datavalues)
```

```
## [1] 26910
```

Finalmente terminamos con un summary para ver los principales estadísticos descriptivos como el mínimo, máximo, cuartiles y promedio para cada paciente:

```
summary(datavalues)
```

```
##       V1                  V2                 V3                 V4
## Min.   :-99.0000   Min.   : -9.0000   Min.   :-99.00   Min.   :-99.00
## 1st Qu.:  0.5225   1st Qu.:  0.5125   1st Qu.:  0.51   1st Qu.:  0.50
## Median :  3.6400   Median :  3.9500   Median :  3.70   Median :  4.47
```

```
##  Mean   : 37.4385   Mean   : 42.2386   Mean   : 34.43   Mean   : 41.05
##  3rd Qu.: 33.5500   3rd Qu.: 42.1750   3rd Qu.: 33.32   3rd Qu.: 41.00
##  Max.   :876.0000   Max.   :951.0000   Max.   :723.00   Max.   :987.00
##  NA's   :16         NA's   :158        NA's   :1        NA's   :1
##        V5                 V6                 V7                  V8
##  Min.   : -99.0000  Min.   : -99.000   Min.   :  -9.0000  Min.   :-
99.000
##  1st Qu.:   0.4862  1st Qu.:   0.540   1st Qu.:   0.6275  1st Qu.:
0.535
##  Median :   2.9050  Median :   3.945   Median :   5.1350  Median :
4.260
##  Mean   :  39.8087  Mean   :  42.687   Mean   :  41.3502  Mean   :
40.040
##  3rd Qu.:  44.5000  3rd Qu.:  38.150   3rd Qu.:  42.6250  3rd Qu.:
45.900
##  Max.   :1070.0000  Max.   :1320.000   Max.   :1250.0000  Max.
:926.000
##  NA's   :22         NA's   :2          NA's   :174        NA's   :15
##        V9                V10                V11                V12
##  Min.   : -99.000   Min.   :-99.00     Min.   :-99.00     Min.   :-99.00
##  1st Qu.:   0.545   1st Qu.: 0.53      1st Qu.: 0.46      1st Qu.: 1.00
##  Median :   4.370   Median : 3.57      Median : 3.98      Median : 4.24
##  Mean   :  43.993   Mean   : 43.53     Mean   : 40.57     Mean   : 41.45
##  3rd Qu.:  43.775   3rd Qu.: 47.95     3rd Qu.: 43.50     3rd Qu.: 40.20
##  Max.   :1150.000   Max.   :954.00     Max.   :892.00     Max.   :746.00
##  NA's   :176        NA's   :174        NA's   :181        NA's   :17
##        V13               V14                V15                V16
##  Min.   :-99.00     Min.   : -99.00    Min.   :  -9.0000  Min.   : -
9.000
##  1st Qu.: 0.91      1st Qu.:   0.64    1st Qu.:   0.5978  1st Qu.:
0.540
##  Median : 4.26      Median :   4.40    Median :   4.8750  Median :
4.025
##  Mean   : 39.31     Mean   :  46.65    Mean   :  50.6099  Mean   :
42.813
##  3rd Qu.: 42.05     3rd Qu.:  42.10    3rd Qu.:  56.4750  3rd Qu.:
43.750
##  Max.   :654.00     Max.   :1050.00    Max.   :1350.0000  Max.
:833.000
##  NA's   :15         NA's   :13         NA's   :10         NA's   :8
##        V17               V18                V19                V20
##  Min.   : -99.00    Min.   :  -9.00    Min.   : -9.0000   Min.   : -
99.000
##  1st Qu.:   0.55    1st Qu.:   0.50    1st Qu.: 0.5582    1st Qu.:
0.600
##  Median :   4.51    Median :   4.18    Median : 4.3450    Median :
5.075
##  Mean   :  46.11    Mean   :  49.27    Mean   : 42.5893   Mean   :
45.525
##  3rd Qu.:  43.95    3rd Qu.:  47.50    3rd Qu.: 41.0500   3rd Qu.:
```

```
48.075
##  Max.   :1040.00   Max.   :1560.00   Max.   :863.0000   Max.
:1240.000
##  NA's   :7          NA's   :5          NA's   :10         NA's   :8
##       V21                V22                V23                V24
##  Min.   : -9.0000   Min.   : -99.000   Min.   :-99.000    Min.   : -
99.000
##  1st Qu.:  0.4625   1st Qu.:   0.545   1st Qu.:  0.370    1st Qu.:
0.675
##  Median :  3.2500   Median :   4.570   Median :  3.045    Median :
4.875
##  Mean   : 35.3213   Mean   :  48.874   Mean   : 36.021    Mean   :
51.309
##  3rd Qu.: 34.9750   3rd Qu.:  47.550   3rd Qu.: 33.450    3rd Qu.:
50.375
##  Max.   :677.0000   Max.   :1230.000   Max.   :965.000    Max.
:1280.000
##  NA's   :4          NA's   :183        NA's   :22         NA's   :8
##       V25                V26                V27                V28
##  Min.   :-99.0000   Min.   :-99.000    Min.   :-99.0000   Min.   :-
99.000
##  1st Qu.:  0.4725   1st Qu.:  0.535    1st Qu.:  0.5025   1st Qu.:
0.520
##  Median :  3.4300   Median :  3.950    Median :  3.7950   Median :
4.155
##  Mean   : 41.3190   Mean   : 35.082    Mean   : 40.4768   Mean   :
38.734
##  3rd Qu.: 43.0000   3rd Qu.: 34.850    3rd Qu.: 40.8000   3rd Qu.:
37.000
##  Max.   :935.0000   Max.   :871.000    Max.   :907.0000   Max.
:713.000
##  NA's   :4          NA's   :31         NA's   :24         NA's   :334
##       V29                V30                V31                V32
##  Min.   :-99.0000   Min.   : -99.000   Min.   : -99.00    Min.   : -
9.0000
##  1st Qu.:  0.4925   1st Qu.:   0.767   1st Qu.:   0.70    1st Qu.:
0.4485
##  Median :  4.1150   Median :   5.705   Median :   4.68    Median :
3.2900
##  Mean   : 40.4764   Mean   :  57.432   Mean   :  52.67    Mean   :
39.4644
##  3rd Qu.: 42.8500   3rd Qu.:  59.200   3rd Qu.:  59.00    3rd Qu.:
39.7500
##  Max.   :846.0000   Max.   :1240.000   Max.   :1140.00    Max.
:921.0000
##  NA's   :156        NA's   :10         NA's   :181        NA's   :27
##       V33                V34                V35                V36
##  Min.   : -9.0000   Min.   : -99.00    Min.   : -9.0000   Min.   : -
9.0000
##  1st Qu.:  0.4375   1st Qu.:   0.56    1st Qu.:  0.4875   1st Qu.:
```

```
0.3818
## Median :  2.6500   Median :   4.18   Median :  3.3950   Median :
3.7600
## Mean    : 37.1846   Mean    :  46.91   Mean    : 42.2323   Mean    :
38.3678
## 3rd Qu.: 38.3500   3rd Qu.:  46.50   3rd Qu.: 41.2500   3rd Qu.:
49.2750
## Max.    :903.0000   Max.    :1220.00   Max.    :990.0000   Max.
:562.0000
## NA's    :22          NA's    :149       NA's    :22        NA's    :328
##      V37              V38              V39
## Min.    :-99.00   Min.    : -9.00   Min.    :-99.0000
## 1st Qu.:  0.88   1st Qu.:  0.41   1st Qu.:  0.9808
## Median :  4.54   Median :  3.56   Median :  4.4150
## Mean    : 37.63   Mean    : 36.79   Mean    : 44.7166
## 3rd Qu.: 38.00   3rd Qu.: 40.30   3rd Qu.: 50.3000
## Max.    :593.00   Max.    :680.00   Max.    :552.0000
## NA's    :357       NA's    :187       NA's    :328
```

Toda esta información que hemos obtenido nos sirve como paso inicial para ver rápidamente qué tenemos entre manos y resumir las características principales de nuestros datos. Todo ello nos ayudará a planificar los próximos pasos donde tendremos que indagar mucho más para poder sacar conclusiones válidas y poder responder a nuestra pregunta biológica: Determinar la influencia de la cirugía bariátrica en el metabolismo de los pacientes.

Para terminar, creamos un repositorio de github donde almacenar todos los datos y el contenedor, así como un informe con todo el procedimiento realizado anteriormente. Los pasos a seguir para constituir un repositorio y almacenar nuestro proyecto son los siguientes:

1. Registro con cuenta propia (Username: mikeferrando).
2. Creación de repositorio: Home > New > New Repository.
3. Asignación de nombre: ferrando-leon-miguel-pec1.
4. Carga de archivos: Add File.

He empezado subiendo los datos crudos en formato de texto .csv, además de los metadatos en Rmd, el informe y el códdigo en R markdown.

Para guardar el objeto contenedor de los datos y los metadatos en formato binario .Rda, utilizamos la función save para cada objeto:

```
save(se, file = "se.rda")

save(metaboinfo, file = "metaboinfo.rda")
```

Después subiremos directamente estos archivos restantes a nuestro respositorio de github, que se puede acceder mediante el siguiente link:

https://github.com/mikeferrando/ferrando-leon-miguel-pec1