PEC 1

ARANTXA GARCIA REDON

2025-03-30

En primer lugar, vamos a cargar los paquetes que necesitamos para resolver la PEC y además vamos a cargar los datos en R para poder trabajar con ellos. En mi caso he decidido trabajar con el conjunto de datos human_cachexia. He decidido trabajar con este conjunto de datos porque es sencillo pero versatil en cuanto a los análisis que se puede hacer con él. Se trata de datos de los metabolitos de varios pacientes, un grupo control y un grupo caquéxico. En los metadatos de nos indica que las muestras no son pareadas, es decir, que no se trata de datos de una misma persona que en un momento era normal y despues desarrolla la enfermedad, por lo que nos da una guía de los tests estadísticos que podemos emplear y los que no. Nos dice además que todos los datos son numéricos y que no hay datos nulos.

```
# Se carga SummarizedExperiment. EN la cabecera se incluye:
# - warning=FALSE, message=FALSE: que permite que aunque se visualice el código, los
# warnings que aparecen al cargar la librería no ensucien el informe.
# - results='hide' porque la opción head también devuelve un resultado muy largo.
library(SummarizedExperiment)
# Se cargan los datos y los metadatos para trabajar con ellos
datos <-read.csv("human_cachexia.csv")
metadatos <-readLines("description.md")
# Se visualizan los datos.
# De nuevo he empleado
print(metadatos)
head(datos)
```

Este conjunto de datos tiene la siguiente estructura: - Las filas representan los diferentes pacientes del estudio. Algunos son caquéxicos y otros no lo son.

- En las columnas tenemos los diferentes datos que se recogen de cada paciente. La primera columna se trata de el identificador del paciente, la segunda su condicion de salud que puede ser normal o caquéxico y el resto de las columnas son diferentes metabolitos que se han recogido de cada uno de los pacientes.

1. Creación de un objeto SummarizedExperiment y estudio de las diferencias con la clase ExpressionSet

Para la creación del objeto SummarizedExperiment (de ahora en adelante, SE), he accedido a la documentación oficial de Bioconductor (https://www.bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html#subsetting).

En primer lugar, mirando el apartado de la anatomía de un objeto SE, lo que se puede observar es la necesidad de trasponer los datos daado que en un objeto SE las diferentes muestras(pacientes) están en las columnas y en las filas se deben recoger los features, que en este caso son los metabolitos. Además se admite que se tenga más de un ensayo para unas mismas samples y features y unos metadatos, que se almacenen por separado pero ligados al conjunto de datos principal.

En el código de a continuación se va a crear el objeto SE tras preprocesar los datos para poder llevar a cabo su creación.

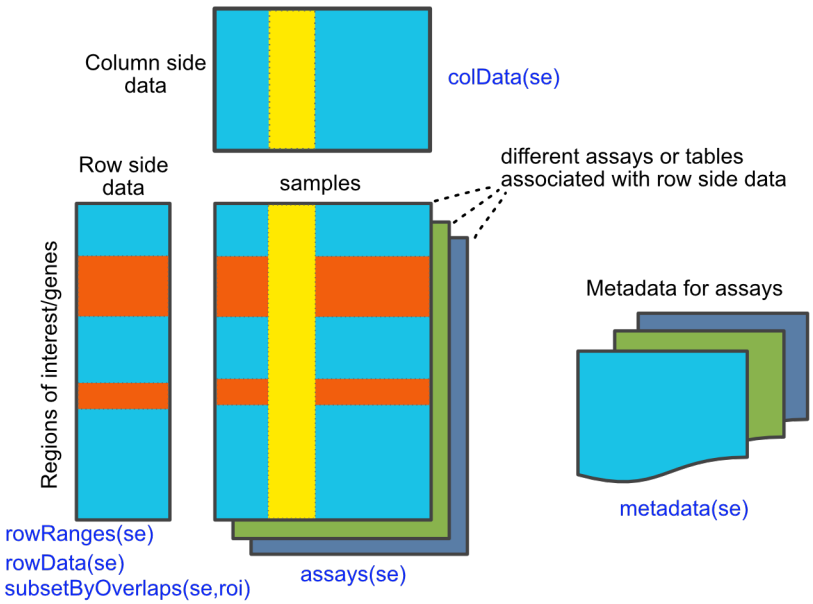


Figure 1: Estrucutra SE

```
# Creo una nueva matriz que es la traspesta de datos, en la cual elimino
# las dos primeras columnas para quedarme solo con los datos numéricos
datos_num<-t(datos[,-c(1,2)])
# El  coldatos contendrá los valores de la segunda columna de datos.
# Es decir el dato de control o enfermos guardado en un vector
coldatos<-datos[,2]
# Convertimos coldatos en un dataframe que contiene el ID del paciente y su condicion
coldatos<-data.frame(SampleName=datos[,1], Muscle.loss=datos[,2])
#Creamos el objeto SE.
se<-SummarizedExperiment(assays=list(counts=datos_num), colData=coldatos, metadata=metadatos)
```

El objeto SE contiene los siguientes componentes:

- `assays = list(counts = datos_num)`: esto contendrá los datos numéricos que hacen referencia a las cantidades de metabolitos de la matriz transpuesta datos_num.
- `colData = coldatos`: contiene los metadatos asociados a cada columna (es decir, a las muestras). En este caso, se incluye la información del dataframe coldatos, que contiene:
  - SampleName (nombre de la muestra)
  - Muscle.loss (información sobre la pérdida muscular)
- `metadata = metadatos`: Este aargumento nos permite introducir los metadatos en formato texto que hemos importado del archivo description.

En la PEC nos pide que comparemos SE con ExpressionSet. Ambas son clases, o estructuras de datos de Bioconductor que son frecuentemente utilizadas para representar y manejar datos experimentales de datos ómicos. Ambas se usan mucho en estudios de expresión génica.

Para comprender mejor la clase ExpressionSet, de nuevo se ha consultado la página oficial de bioconductor (https://www.bioconductor.org/packages/devel/bioc/vignettes/Biobase/inst/doc/ExpressionSetIntroduction.pdf). La principal diferencia entre los dos paquetes parece ser la flexibilidad en cuanto a los tipos de datos que pueden almacenarse en cada una de las estructuras de datos. Mirando diferentes repositorios en github y stackoverflow se puede ver que ExpressionSet se utiliza para estudios de expresión génica con microarrays de

forma casi exclusiva. Sin embargo, SummarizedExperiment es más flexible y permite trabajar con datos de RNAseq y otros datos ómicos.

Otra diferencia es como se estucturan los datos:

- `SummarizedExperiment` se estructura en:
  - assays: que es una matriz o dataframe con los datoss experimentales (por ejemplo los datos de expresión o de metabolómica como en el dataset que estamos usando)
  - `coldata`: un dataframe que contiene los metadatos de las muestras.
  - `rowdata`: un dataframe con los metadatos sobre las filas, como los genes
  - `metadata`: un apartado para metadatos generales del experimento, donde en este caso hemos almacenado la descripción.
- `ExpressionSet` tiene tres componentes principales:
  - `exprs`: una matriz de datos de expresión con los genes en filas y las muestras en columnas.
  - `phenodata`: un objeto que contiene metadatos sobre las muestras, como por ejemplo las condiciones del experimento.
  - `featureData`que contiene información sobre los genes del experimento.

En general, podemos decir que las diferencias en su estructura hacen que `ExpressionSet`es una estructura de datos más rígida que puede resultar muy útil en experimentos de expresión génica con microarrays, mientras que `SummarizedExperiment`tiene una estructura más flexible que nos permite almacenar más datos y es útil para otros tipos de datos ómicos, no solo de microarrays.

A continuación voy a guardar el objeto SummarizedExperiment en formato binario .Rda como se indica en las instrucciones de la PEC, para ello se usa la función save:

```
save(se, file = "cachexia_se.Rda")
```

2. Análisis exploratorio de los datos del dataset cachexia.

En primer lugar se va a explorar algunas funciones que se puede hacer con el objeto `se`.

Para poder comparar los datos entre el grupo de caquexia y el grupo control, voy a crear dos matrices distintas con los datos de cada uno de los grupos.

Después voy a comenzar ocn el análisis exploratorio de alguno de los datos. En primer lugar, utilizo la

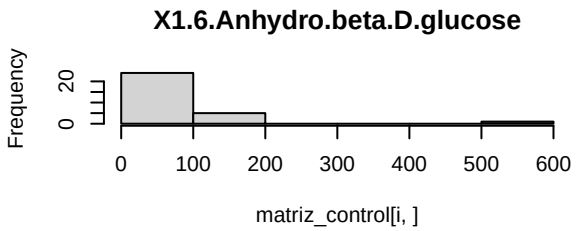
```
matriz_control<-assays(se[,se$Muscle.loss=="control"])$counts
matriz_cachexia<-assays(se[,se$Muscle.loss=="cachexic"])$counts
summary(t(matriz_control))
```

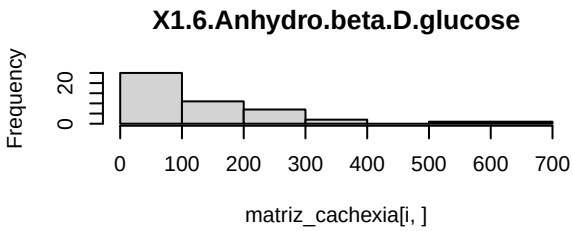
```
##  X1.6.Anhydro.beta.D.glucose X1.Methylnicotinamide X2.Aminobutyrate
##  Min.   :  9.39              Min.   :   6.42       Min.   : 1.280
##  1st Qu.: 22.53              1st Qu.:  12.36       1st Qu.: 4.210
##  Median : 34.98              Median :  19.43       Median : 7.580
##  Mean   : 69.51              Mean   :  73.16       Mean   : 9.528
##  3rd Qu.: 65.05              3rd Qu.:  54.74       3rd Qu.:12.935
##  Max.   :528.48              Max.   :1032.77       Max.   :28.790
##  X2.Hydroxyisobutyrate X2.Oxoglutarate  X3.Aminoisobutyrate X3.Hydroxybutyrate
##  Min.   : 4.85         Min.   :  5.64   Min.   :  3.13      Min.   : 2.230
##  1st Qu.:11.01         1st Qu.: 14.34   1st Qu.: 10.99      1st Qu.: 3.808
##  Median :19.30         Median : 31.68   Median : 18.41      Median : 6.655
##  Mean   :27.87         Mean   : 85.52   Mean   : 39.91      Mean   : 9.899
##  3rd Qu.:42.81         3rd Qu.: 80.07   3rd Qu.: 33.04      3rd Qu.:14.920
##  Max.   :93.69         Max.   :982.40   Max.   :208.51      Max.   :34.120
##  X3.Hydroxyisovalerate X3.Indoxylsulfate X4.Hydroxyphenylacetate
##  Min.   : 0.920        Min.   : 27.66    Min.   : 15.49
##  1st Qu.: 3.485        1st Qu.: 51.45    1st Qu.: 30.34
```

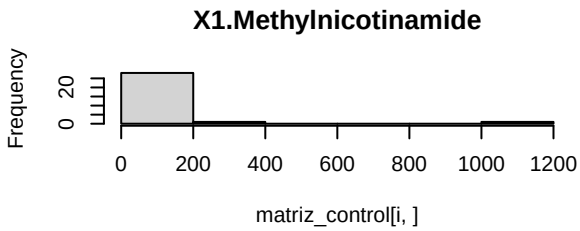
```
##  Median : 5.625        Median :105.11    Median : 50.70
##  Mean   :12.313        Mean   :146.38    Mean   : 99.80
##  3rd Qu.: 9.255        3rd Qu.:162.81    3rd Qu.: 84.35
##  Max.   :60.950        Max.   :614.00    Max.   :796.32
##     Acetate           Acetone          Adipate          Alanine
##  Min.   :  3.490   Min.   : 2.290   Min.   : 1.550   Min.   : 16.78
##  1st Qu.:  9.415   1st Qu.: 5.067   1st Qu.: 3.960   1st Qu.: 56.26
##  Median : 17.320   Median : 6.925   Median : 6.295   Median : 78.65
##  Mean   : 35.605   Mean   : 8.420   Mean   : 8.993   Mean   :157.58
##  3rd Qu.: 45.515   3rd Qu.:10.255   3rd Qu.: 8.918   3rd Qu.:212.75
##  Max.   :202.350   Max.   :23.810   Max.   :58.560   Max.   :601.85
##    Asparagine        Betaine         Carnitine         Citrate
##  Min.   :  6.69   Min.   :  2.29   Min.   :  2.72   Min.   :  59.74
##  1st Qu.: 16.95   1st Qu.: 13.68   1st Qu.: 12.19   1st Qu.: 426.25
##  Median : 29.52   Median : 32.55   Median : 19.20   Median :1043.62
##  Mean   : 41.75   Mean   : 55.97   Mean   : 32.44   Mean   :1474.72
##  3rd Qu.: 45.57   3rd Qu.: 58.15   3rd Qu.: 38.95   3rd Qu.:2332.18
##  Max.   :152.93   Max.   :311.06   Max.   :206.44   Max.   :4230.18
##     Creatine         Creatinine    Dimethylamine     Ethanolamine
##  Min.   :  2.750   Min.   : 1002   Min.   : 41.26   Min.   : 21.54
##  1st Qu.:  8.258   1st Qu.: 2322   1st Qu.:103.03   1st Qu.: 54.78
##  Median : 19.515   Median : 3697   Median :149.91   Median :123.03
##  Mean   : 51.504   Mean   : 5619   Mean   :208.68   Mean   :197.13
##  3rd Qu.: 42.962   3rd Qu.: 8146   3rd Qu.:311.87   3rd Qu.:231.50
##  Max.   :395.440   Max.   :15063   Max.   :497.70   Max.   :906.87
##     Formate           Fucose          Fumarate         Glucose
##  Min.   :  6.42   Min.   :  5.70   Min.   : 0.790   Min.   : 26.84
##  1st Qu.: 36.59   1st Qu.: 24.17   1st Qu.: 1.620   1st Qu.: 71.88
##  Median : 61.56   Median : 42.59   Median : 3.225   Median :103.59
##  Mean   : 84.48   Mean   : 57.44   Mean   : 4.552   Mean   :140.96
##  3rd Qu.:109.25   3rd Qu.: 85.44   3rd Qu.: 4.220   3rd Qu.:208.31
##  Max.   :292.95   Max.   :196.37   Max.   :36.230   Max.   :336.97
##    Glutamine         Glycine          Glycolate      Guanidoacetate
##  Min.   : 23.34   Min.   :  38.09   Min.   :  5.42   Min.   :  7.03
##  1st Qu.: 44.27   1st Qu.: 185.40   1st Qu.: 39.56   1st Qu.: 17.73
##  Median :117.41   Median : 417.05   Median : 70.61   Median : 48.00
##  Mean   :174.43   Mean   : 585.15   Mean   :138.98   Mean   : 68.74
##  3rd Qu.:261.32   3rd Qu.: 845.56   3rd Qu.:172.67   3rd Qu.: 83.72
##  Max.   :862.64   Max.   :2275.60   Max.   :720.54   Max.   :301.87
##    Hippurate         Histidine       Hypoxanthine      Isoleucine
##  Min.   : 122.7   Min.   : 16.28   Min.   :  3.78   Min.   : 1.790
##  1st Qu.: 403.8   1st Qu.: 40.28   1st Qu.: 12.30   1st Qu.: 2.902
##  Median : 602.6   Median : 88.38   Median : 33.67   Median : 4.310
##  Mean   :1364.2   Mean   :180.47   Mean   : 51.71   Mean   : 7.218
##  3rd Qu.:1195.0   3rd Qu.:267.07   3rd Qu.: 71.33   3rd Qu.: 9.485
##  Max.   :6634.2   Max.   :720.54   Max.   :175.91   Max.   :21.330
##     Lactate          Leucine           Lysine        Methylamine
##  Min.   :  7.32   Min.   : 2.510   Min.   : 10.49   Min.   : 1.51
##  1st Qu.: 24.82   1st Qu.: 7.173   1st Qu.: 21.87   1st Qu.: 4.12
##  Median : 40.88   Median : 9.120   Median : 35.34   Median : 5.34
##  Mean   : 65.75   Mean   :13.557   Mean   : 89.23   Mean   :11.36
##  3rd Qu.: 97.53   3rd Qu.:20.047   3rd Qu.: 77.56   3rd Qu.:16.01
##  Max.   :198.34   Max.   :38.090   Max.   :788.40   Max.   :44.70
##  Methylguanidine  N.N.Dimethylglycine O.Acetylcarnitine  Pantothenate
```

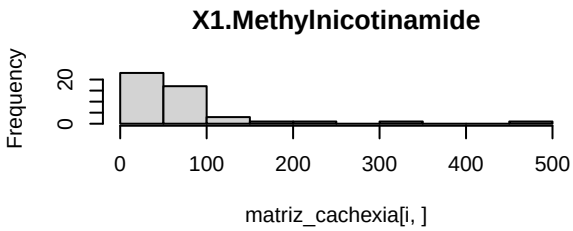
```
##  Min.   : 1.700   Min.   : 1.230      Min.   : 1.230    Min.   :  3.10
##  1st Qu.: 3.535   1st Qu.: 3.525      1st Qu.: 2.192    1st Qu.:  9.28
##  Median : 6.820   Median : 8.940      Median : 6.675    Median : 14.73
##  Mean   :12.128   Mean   :13.597      Mean   :10.598    Mean   : 52.62
##  3rd Qu.:16.543   3rd Qu.:19.497      3rd Qu.:16.360    3rd Qu.: 27.52
##  Max.   :44.260   Max.   :52.460      Max.   :43.820    Max.   :692.29
##  Pyroglutamate       Pyruvate       Quinolinate         Serine
##  Min.   : 21.33   Min.   : 0.900   Min.   :  5.21   Min.   : 16.12
##  1st Qu.: 47.04   1st Qu.: 4.372   1st Qu.: 16.88   1st Qu.: 50.03
##  Median : 83.95   Median : 6.620   Median : 27.26   Median : 98.28
##  Mean   :119.26   Mean   :12.566   Mean   : 39.32   Mean   :122.26
##  3rd Qu.:157.20   3rd Qu.:14.633   3rd Qu.: 49.29   3rd Qu.:173.45
##  Max.   :441.42   Max.   :66.690   Max.   :164.02   Max.   :383.75
##    Succinate          Sucrose          Tartrate         Taurine
##  Min.   :  1.720   Min.   :  6.49   Min.   :  2.20   Min.   :  17.81
##  1st Qu.:  4.702   1st Qu.: 16.86   1st Qu.:  6.54   1st Qu.:  70.92
##  Median : 11.805   Median : 19.39   Median : 10.75   Median : 192.21
##  Mean   : 29.836   Mean   : 55.58   Mean   : 28.68   Mean   : 320.52
##  3rd Qu.: 33.310   3rd Qu.: 39.65   3rd Qu.: 19.37   3rd Qu.: 408.09
##  Max.   :221.410   Max.   :601.85   Max.   :273.14   Max.   :1510.20
##    Threonine       Trigonelline    Trimethylamine.N.oxide   Tryptophan
##  Min.   :  9.12   Min.   : 10.07   Min.   :  55.7         Min.   : 10.49
##  1st Qu.: 18.93   1st Qu.: 38.68   1st Qu.: 131.3         1st Qu.: 15.53
##  Median : 39.45   Median : 72.28   Median : 257.7         Median : 22.69
##  Mean   : 59.52   Mean   :130.69   Mean   : 388.7         Mean   : 41.83
##  3rd Qu.: 64.08   3rd Qu.:139.45   3rd Qu.: 467.9         3rd Qu.: 43.88
##  Max.   :249.64   Max.   :566.80   Max.   :1540.7         Max.   :184.93
##     Tyrosine          Uracil           Valine           Xylose
##  Min.   :  4.22   Min.   :  3.10   Min.   : 4.100   Min.   : 10.07
##  1st Qu.: 16.36   1st Qu.:  9.58   1st Qu.: 8.678   1st Qu.: 21.59
##  Median : 27.00   Median : 21.36   Median :13.470   Median : 34.88
##  Mean   : 52.01   Mean   : 32.49   Mean   :20.133   Mean   : 56.51
##  3rd Qu.: 63.18   3rd Qu.: 48.87   3rd Qu.:30.650   3rd Qu.: 50.27
##  Max.   :179.47   Max.   :138.38   Max.   :56.830   Max.   :407.48
##  cis.Aconitate     myo.Inositol    trans.Aconitate  pi.Methylhistidine
##  Min.   : 12.94   Min.   : 11.59   Min.   :  4.90   Min.   :  11.36
##  1st Qu.: 25.99   1st Qu.: 20.71   1st Qu.: 10.02   1st Qu.:  53.65
##  Median : 56.30   Median : 30.73   Median : 14.31   Median :  73.33
##  Mean   : 91.72   Mean   : 62.64   Mean   : 27.81   Mean   : 258.64
##  3rd Qu.:157.20   3rd Qu.: 72.34   3rd Qu.: 30.05   3rd Qu.: 295.93
##  Max.   :298.87   Max.   :314.19   Max.   :181.27   Max.   :1187.97
##  tau.Methylhistidine
##  Min.   :  8.58
##  1st Qu.: 16.95
##  Median : 37.19
##  Mean   : 64.65
##  3rd Qu.: 88.71
##  Max.   :287.15
```

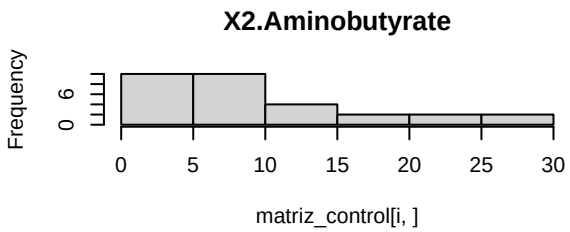
```
par(mfrow=c(3,2))
for(i in 1:3){
  hist(matriz_control[i,], main=rownames(matriz_control)[i])
  hist(matriz_cachexia[i,], main=rownames(matriz_cachexia)[i])
}
```

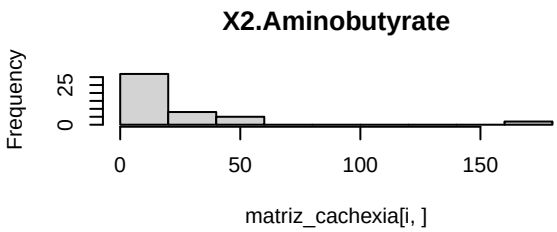

X1.6.Anhydro.beta.D.glucose
Frequency
0
20
0
100
200
300
400
500
600
matriz_control[i,]


X1.6.Anhydro.beta.D.glucose
Frequency
0
20
0
100
200
300
400
500
600
700
matriz_cachexia[i,]


X1.Methylnicotinamide
Frequency
0
20
0
200
400
600
800
1000
1200
matriz_control[i,]


X1.Methylnicotinamide
Frequency
0
20
0
100
200
300
400
500
matriz_cachexia[i,]


X2.Aminobutyrate
Frequency
0
6
0
5
10
15
20
25
30
matriz_control[i,]


X2.Aminobutyrate
Frequency
0
25
0
50
100
150
matriz_cachexia[i,]