PAC 1

Pau Pérez Rebassa

Preparació de l'objecte *SummarizedExperiment*

El *dataset* escollit és el *2023-CIMCBTutorial* del github que l'enunciat de la PAC ens ha facilitat. Com es pot veure, tenim un fitxer *.xlsx* anomenat *GastricCancer_NMR* que a dins té dos fulls de càlcul, el primer d'ells conté les dades i el segon les metadades. El primer que s'ha de fer és llegir cada una dels fulls i emmagatzemar-los dins dues variables separades. Gràcies al paquet *readxl* es pot llegir el fitxer.

```
library(readxl)
dades = read_excel("metaboData/Datasets/2023-CIMCBTutorial/GastricCancer_NMR.xlsx",
                   sheet = "Data")
meta = read_excel("metaboData/Datasets/2023-CIMCBTutorial/GastricCancer_NMR.xlsx",
                   sheet = "Peak")
```

Com es pot veure a la informació sobre les dades carregades, el *dataframe dades* té 140 observacions i 153 variables i el *meta* té 149 observacions i 5 variables. S'haurà de modificar les variables de *dades* perquè coincideixi amb les observacions de *meta*. També, el que veim, és que les 4 primeres variables són informació sobre les mostres i la resta de variables són pròpies dades en brut, per tant, el que hem de fer és separar aquest tipus d'informació en dues variable més, de la següent manera:

```
informacio_mostra = dades[, 1:4]
dades = dades[, -c(1:4)]
```

Després d'haver executat les dues intruccions anteriors, tenim separada la informació de la mostra i las dades en brut, d'aquesta manera es veu que ja coincideixen les variables de *dades* i les observacions de *meta*.

Per poder contruir l'objecte *SummarizedExperiment* hem de modificar la informació de la mostra. El primer que feim es convertir en factor els camps *SampleType* i *Class*.

```
informacio_mostra$SampleType = as.factor(informacio_mostra$SampleType)
informacio_mostra$Class = as.factor(informacio_mostra$Class)
```

Per poder seguir endavant en la creació de l'objecte, hem de relacionar d'alguna manera la informació de les mostres amb les dades. Per això cream una nova columna que contendrà un identificador que nosaltres crearem depenent de cada valor de *Idx*, *SampleType*, *Class* de la variable *informació_mostra*, per exmple, en el primer valor quedaria així: ***1_QC_QC***.

```
suppressMessages(suppressPackageStartupMessages(library(dplyr)))
library(dplyr)
informacio_mostra = informacio_mostra %>%
  mutate(new_Id = paste(informacio_mostra$Idx, substr(informacio_mostra$SampleType, 1, 1),
                        informacio_mostra$Class, sep = "_"))
```

```
informacio_mostra = as.data.frame(informacio_mostra)
rownames(informacio_mostra) = informacio_mostra$new_Id
informacio_mostra$new_Id = NULL
```

Eliminam la variable *Idx* i *SampleType* perquè una vegada creat el nou identificador per la mostra, les dues variables contenen informació redundant.

```
informacio_mostra$Idx = NULL
informacio_mostra$SampleID = NULL
head(informacio_mostra)
```

```
##        SampleType Class
## 1_Q_QC         QC    QC
## 2_S_GC     Sample    GC
## 3_S_BN     Sample    BN
## 4_S_HE     Sample    HE
## 5_S_GC     Sample    GC
## 6_S_BN     Sample    BN
```

Generam una matriu amb les dades en cru, per poder crear l'objecte final i afegim el nom de les files que hem creat a *informacio_mostra*, d'aquesta manera ja tenim tot relacionat entre si.

```
matriu_dades = as.matrix(dades)
rownames(matriu_dades) = rownames(informacio_mostra)
```

Ja podem crear l'objecte *SummarizedExperiment* i ho feim de la següent manera:

```
suppressMessages(suppressPackageStartupMessages(library(SummarizedExperiment)))
library(SummarizedExperiment)
experiment = SummarizedExperiment(assays = list(rawValues = matriu_dades),
                                  rowData = informacio_mostra, colData = meta)
```