

PAC1 Joan Mora Peralta

Joan Mora Peralta

2024-10-25

```
system("git clone https://github.com/nutrimetabolomics/metaboData.git")
```

Aquí hem descarregat l'estat actual del repositori Git.

```
assayData <- read.csv("metaboData/Datasets/2018-MetabotypingPaper/DataValues_S013.csv", header=TRUE, sep = ";")
phenoData <- read.csv("metaboData/Datasets/2018-MetabotypingPaper/DataInfo_S013.csv", header=TRUE, sep = ";")
experiment <- SummarizedExperiment(assays = list(assayData = assayData), colData = phenoData)
```

Aquí hem accedit a la carpeta descarregada i transformat els arxius .csv en dues variables que passen a formar part del nou objecte 'experiment' de la classe SummarizedExperiment.

```
experiment
```

```
## class: SummarizedExperiment
## dim: 39 695
## metadata(0):
## assays(1): assayData
## rownames(39): 1 2 ... 38 39
## rowData names(0):
## colnames(695): SUBJECTS SURGERY ... SM.C24.0_T5 SM.C24.1_T5
## colData names(3): VarName varTpe Description
```

```
colData(experiment)
```

```
## DataFrame with 695 rows and 3 columns
##           VarName    varTpe Description
##           <factor> <factor>  <factor>
## SUBJECTS    SUBJECTS integer   dataDesc
## SURGERY      SURGERY  character dataDesc
## AGE          AGE      integer   dataDesc
## GENDER       GENDER   character dataDesc
## Group        Group    integer   dataDesc
## ...          ...      ...       ...
## SM.C18.0_T5  SM.C18.0_T5 numeric   dataDesc
## SM.C18.1_T5  SM.C18.1_T5 numeric   dataDesc
## SM.C20.2_T5  SM.C20.2_T5 numeric   dataDesc
## SM.C24.0_T5  SM.C24.0_T5 numeric   dataDesc
## SM.C24.1_T5  SM.C24.1_T5 numeric   dataDesc
```

```
rowData(experiment)
```

```
## DataFrame with 39 rows and 0 columns
```

En aquest primer anàlisi veiem que es tracta d'un dataset d'e l'experiment d'un experiment amb dimensions: 39x695 (observacions i variables). Els subjectes estudiats estan numerats i les variables son de diferent naturalesa ("SURGERY", "SM.C24.0_T5" ...). Per tant, les primeres variables s'espera que siguin més categòriques i més generals i després signifiquin valors en diferents registres metabolòmics. Això concorda amb la descripció de l'experiment escrit en el catàleg del repositori, que diu : Data used in the paper "Metabotypes of response to bariatric surgery independent of the magnitude of weight loss". Les descripcions no estan ben anotades, ja que en totes hi tenim el mateix valor "dataDesc".

També veiem com rowData no presenta més columnes d'informació, ja que no hem carregat cap anotació al instanciar l'objecte SummarizedExperiment.

```
rowData(experiment) <- data.frame(anotacions_ = c(rep("info de fila", 39)))
```

Ja que no tenim anotacions de les files, les posem de manera artificial a mode d'exemple.

```
assay(experiment)[,1:2] <- apply(assay(experiment)[,1:2], 2, as.factor)
assay(experiment)[,4:9] <- apply(assay(experiment)[,4:9], 2, as.factor)
```

Transformem el tipus de certes columnes a factors, ja que d'altra manera R les interpretarà com a numèriques.

```
na_total <- sum(is.na(assay(experiment)))
na_prop <- na_total / prod(dim(experiment))
cat("Proporció de NA en el dataset: ",round(na_prop*100, 0),"%")
```

```
## Proporció de NA en el dataset: 13 %
```

```
na_cols <- as.matrix(apply(assay(experiment), 2, function(x) sum(is.na(x))>0))
na_cols_total <- sum(na_cols)
na_cols_prop <- na_cols_total / dim(assay(experiment))[2]
cat("Proporció columnes amb NA en el dataset: ",round(na_cols_prop*100, 0),"%")
```

```
## Proporció columnes amb NA en el dataset: 78 %
```

```
na_fila <- as.matrix(apply(assay(experiment), 1, function(x) sum(is.na(x))>0))
na_fila_total <- sum(na_fila)
na_fila_prop <- na_fila_total / dim(assay(experiment))[1]
cat("Proporció files amb NA en el dataset: ",round(na_fila_prop*100, 0),"%")
```

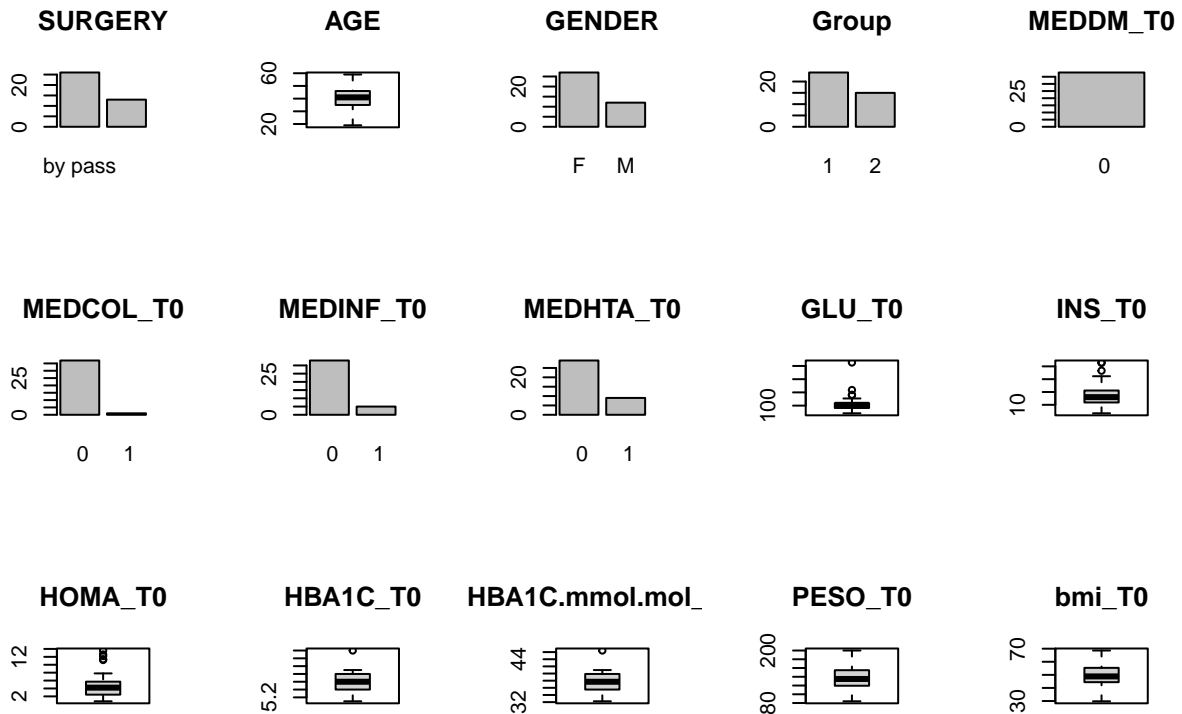
```
## Proporció files amb NA en el dataset: 100 %
```

Veiem que hi ha molta presència de valors nuls, i que aquests no estan concentrats en cap regió, sino que totes les observacions i gran part de les columnes en tenen. En un cas pràctic crec que no es pot eliminar cap columna s'hauria d'imputar valors que afectin mínimament a la resta de valors com ara la mitjana en columnes quantitatives o medianes en factors. Però primer s'hauria d'interpretar molt bé què significa cada una.

```
tipus <- sapply(assay(experiment), class)

getPlot <- function(i) {
  if (tipus[i] == "integer" || tipus[i] == "numeric") {
    boxplot(as.numeric(assay(experiment)[, i]), main = colnames(assay(experiment))[i])
  } else {
    barplot(table(assay(experiment)[, i]), main = colnames(assay(experiment))[i])
  }
}

par(mfrow = c(3,5))
invisible(lapply(2:16, getPlot))
```



Aquí es pot comprovar que hi ha una important disparitat entre els dos tipus de cirurgies i els dos sexes, que les edats son bastant properes al voltant dels 40 anys. Els nivells de glucosa, insulina i altres metabolits presenten diferents distribucions, en especial TAS_T0 que clarament no segueix una distribució normal, el què significa que no podem, de moment, fer cap estudi paramètric. MEDDM_T0 només presenta un sol valor, el què significa que la taula tampoc és, de moment, de rang màxim i una regressió tindrà infinites solucions possibles.