# PAC1

Núria Farran Centelles

2024-11-05

## Contents



## 1 Abstract

En aquesta PAC hem treballat els diferents aspectes que hem après en la primera part del curs. En concret hem utilitzat Bioconductor per introduir el nostre dataset d'interés en un contenidor de dades -òmiques com és SummarizedExpression (SE) per tal d'obtenir la matriu amb els valors del nostre dataset i les característiques que l'acompanyen, tot veient una versió una mica diferent al ExpressionSet que havíem tocat en la primera activitat. D'altra banda hem aplicat alguns dels coneixements obtinguts de les -òmiques, per tal de saber com es troben col · locades les dades en una taula (files x columnes) i com hem de manipular aquesta taula original per tal de poder aplicar el SE. Tot i així, hi ha alguns aspectes ja treballats que no els hem integrat en aquesta PAC, com és el control de qualitat (que ja ho vam fer amb els arxius en format FASTA), o els anàlisis diferencials i gràfics derivats, que això ho portarem a la pràctica en les següents unitats.

# 2 Objectius de l'estudi

L'objectiu principal d'aquest estudi és aplicar els coneixements adquirits en l'activitat 1.1-1.3 sobre manipulació de dades -òmiques, mitjançant Bioconductor, i saber penjar-ho tot al repositori de Github.

# 3 Materials i mètodes

Les dades sobre les quals treballem provenen d'un estudi de metabolòmica, on s'estudia un seguit de metabòlits en pacients amb caquèxia i pacients control. No sabem exactament d'on provenen les dades, ja que no hi havia cap informació addicional com l'enllaç a l'article, o l'enllaç a un Github amb les dades crues. No obstant, hem importat les dades a un contenedor, i les hem analitzat des de diferents perspectives pel que pugui ser.

# 4 Resultats

A continuació, detallarem els diferents passos que hem anat seguint amb el nostre dataset, tot mostrant el codi utilitzat en cada cas i quina informació concreta ens proporciona. Tot i ser molt escàs, és el primer pas necessari abans de començar a fer un anàlisis diferencial i extreure conclusions sobre com els diferents metabòlits podrien ser un possible biomarcador de caquèxia, o quines vies són les que estan més repercutides en aquesta patologia.

Com hem mencionat, partim d'unes dades que no sabem exctament d'on provenen, no sabem del cert quina és la mostra origen (teixit, plasma, orina, etc), i tampoc quina tècnica s'ha utilitzat per obtenir aquestes dades, tot i que podriem intuir que, segurament, és espectometria de masses, ja que és una tècnica ampliment utilitzada en metabolòmica. En concret, tenim 77 pacients dels quals se'ls hi ha analitzat 63 metabòlits. Anem a carregar aquestes dades i seguim amb l'estudi

## 4.1 Carregar dades

En primer lloc carreguem el paquet que volem per tal de crear el contenedor de SummarizedExperiment.

```
library(SummarizedExperiment)
```

```
## Loading required package: MatrixGenerics
```

```
## Loading required package: matrixStats
```

```
## Warning: package 'matrixStats' was built under R version 4.2.3
```

```
##
## 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, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

##
## Attaching package: 'IRanges'

## The following object is masked from 'package:grDevices':
##
##     windows
```

```
## 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
```

D'altra banda hem de llegir el nostre dataset

```
raw <- read.csv("human_cachexia.csv")
```

## 4.2 Adaptació del dataset

A continuació hem d'adaptar el dataset per tal d'obtenir les dades numèriques i les metadades. Per això dividim la taula que ens donen en dos. En les metadades (meta) tindrem la informació corresponen a quin grup pertany cada subjecte (columnes 1 i 2). D'altra banda tenim les dades (data), on hi ha els valors de cada metabòlit per cada pacient (columna 1 i 3-65). Finalment transformem el data en matriu (pas necessari perquè es puguin llegir correctament les dades en el contenedor).

```
meta <- raw[ , 1:2]
predata <- raw[ ,c(1,3:65)]
data_matrix <- as.matrix(predata)
data <- t(data_matrix)
colnames(data) <- data[1, ]
data <- data[-1, ]
```

## 4.3 Creació objecte SummarizedExperiment (SE) i anàlisi estructura

Per tal de crear un contenedor com SummarizedExperiment (SE), necessitem el paquet SummarizedExperiment (ja carregat anteriorment), una matriu on hi haurà els counts (data), és a dir, els valors de cada metabòlit per cada pacient, i les metadades (meta) que, com ja hem dit, és la informació sobre aquest conjunt de dades, tot descrivint les característiques, el context o les propietats de les dades (en aquest cas únicament contindrà la informació de en quin grup pertany cada pacient: control o caquèxia).

```
se <- SummarizedExperiment(assays = list(counts = data), colData = meta)
```

Una vegada tenim el SE creat, anem a analitzar quina estructura té, mirant la classe, les dimensions i observant les primeres línies del fitxer obtingut.

```
class(se)
```

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

```
dim(se)
```

```
## [1] 63 77
```

```
head(assay(se), n=2)
```

```
##                             PIF_178   PIF_087   PIF_090   NETL_005_V1 PIF_115
## X1.6.Anhydro.beta.D.glucose " 40.85"  " 62.18"  "270.43"  "154.47"    " 22.20"
## X1.Methylnicotinamide       "  65.37" " 340.36" "  64.72" "  52.98"   "  73.70"
##                             PIF_110   NETL_019_V1 NETCR_014_V1 NETCR_014_V2
## X1.6.Anhydro.beta.D.glucose "212.72"  "151.41"    " 31.50"     " 51.42"
## X1.Methylnicotinamide       "  31.82" "  36.60"   "   6.82"    "  30.27"
##                             PIF_154   NETL_022_V1 NETL_022_V2 NETL_008_V1
## X1.6.Anhydro.beta.D.glucose "117.92"  " 20.70"    "127.74"    " 59.74"
## X1.Methylnicotinamide       "  52.46" " 221.41"   " 177.68"   "  50.91"
##                             PIF_146   PIF_119   PIF_099   PIF_162   PIF_160
## X1.6.Anhydro.beta.D.glucose " 89.12"  " 23.57"  " 41.26"  "589.93"  "112.17"
## X1.Methylnicotinamide       "  32.79" "   6.89" "   8.67" "  21.98" "  25.28"
##                             PIF_113   PIF_143   NETCR_007_V1 NETCR_007_V2
## X1.6.Anhydro.beta.D.glucose "167.34"  "183.09"  "208.51"     " 34.81"
## X1.Methylnicotinamide       "  19.89" "  90.92" "  53.52"    "  95.58"
##                             PIF_137   PIF_100   NETL_004_V1 PIF_094   PIF_132
## X1.6.Anhydro.beta.D.glucose "333.62"  " 32.46"  "  4.71"    " 68.72"  "214.86"
## X1.Methylnicotinamide       "  35.87" "   9.68" "  11.13"   "  13.87" " 127.74"
##                             PIF_163   NETCR_003_V1 NETL_028_V1 NETL_028_V2
## X1.6.Anhydro.beta.D.glucose "304.90"  " 37.71"     " 45.60"    " 34.12"
## X1.Methylnicotinamide       "  25.79" "  10.80"    " 473.43"   "  92.76"
##                             NETCR_013_V1 NETL_020_V1 NETL_020_V2 PIF_192
## X1.6.Anhydro.beta.D.glucose "107.77"     " 13.33"    " 27.94"    "141.17"
## X1.Methylnicotinamide       "  16.61"    "  50.91"   "  80.64"   "  68.03"
##                             NETCR_012_V1 NETCR_012_V2 PIF_089   NETCR_002_V1
## X1.6.Anhydro.beta.D.glucose " 14.01"     "244.69"     "123.97"  "141.17"
## X1.Methylnicotinamide       "  46.06"    " 116.75"    "  81.45" "  28.50"
##                             PIF_179   PIF_114   NETCR_006_V1 PIF_141
## X1.6.Anhydro.beta.D.glucose " 35.16"  "685.40"  "278.66"     " 15.80"
## X1.Methylnicotinamide       "  26.58" "  36.23" "  40.45"    "  23.57"
##                             NETCR_025_V1 NETCR_025_V2 NETCR_016_V1 PIF_116
## X1.6.Anhydro.beta.D.glucose " 29.96"     " 16.95"     "292.95"     " 29.67"
## X1.Methylnicotinamide       "  96.54"    " 114.43"    "  57.97"    "  70.11"
##                             PIF_191   PIF_164   NETL_013_V1 PIF_188   PIF_195
```

```
## X1.6.Anhydro.beta.D.glucose " 18.92"  "127.74"  " 34.81"    " 65.37"  " 15.18"
## X1.Methylnicotinamide       "  24.53" "1032.77" "  12.30"   "  24.05" "  94.63"
##                             NETCR_015_V1 PIF_102   NETL_010_V1 NETL_010_V2
## X1.6.Anhydro.beta.D.glucose " 70.81"     " 25.28"  " 34.47"    " 18.54"
## X1.Methylnicotinamide       "  75.94"    " 101.49" "  12.81"   "   8.41"
##                             NETL_001_V1 NETCR_015_V2 NETCR_005_V1 PIF_111
## X1.6.Anhydro.beta.D.glucose " 37.34"    " 33.78"     " 22.42"     "146.94"
## X1.Methylnicotinamide       "  55.15"   "  53.52"    "  55.15"    "  10.07"
##                             PIF_171   NETCR_008_V1 NETCR_008_V2 NETL_017_V1
## X1.6.Anhydro.beta.D.glucose " 64.07"  " 32.46"     "113.30"     " 22.20"
## X1.Methylnicotinamide       "   6.42" "  14.01"    "  43.38"    "  20.70"
##                             NETL_017_V2 NETL_002_V1 NETL_002_V2 PIF_190
## X1.6.Anhydro.beta.D.glucose " 46.53"    "192.48"    "528.48"    " 28.79"
## X1.Methylnicotinamide       "   9.78"   " 108.85"   " 225.88"   "   9.21"
##                             NETCR_009_V1 NETCR_009_V2 NETL_007_V1 PIF_112
## X1.6.Anhydro.beta.D.glucose "181.27"     " 47.47"     " 15.96"    " 22.87"
## X1.Methylnicotinamide       "  48.42"    "   7.69"    "  16.12"   "  10.38"
##                             NETCR_019_V2 NETL_012_V1 NETL_012_V2 NETL_003_V1
## X1.6.Anhydro.beta.D.glucose " 35.16"     " 16.95"    "  9.39"    " 37.71"
## X1.Methylnicotinamide       "  52.46"    "  15.80"   "  14.01"   "  18.17"
##                             NETL_003_V2
## X1.6.Anhydro.beta.D.glucose " 38.47"
## X1.Methylnicotinamide       "  12.55"
```

Veiem que ens trobem amb una matriu de 63x77, on 63 son els diferents metabòlits que s'analitzen i 77 són els diferents pacients sobre els quals es fa l'anàlisi. Quan mostrem la taula veiem tan sols les dos primeres files (els dos primers metabòlits) pels 77 pacients.

D'altra banda també podem accedir a la informació relacionada a aquestes dades. Així doncs per tal d'accedir al meta, ho cridem amb la funció colData(), que ens mostrarà de quina informació disposem.

```
colData(se)
```

```
## DataFrame with 77 rows and 2 columns
##                Patient.ID Muscle.loss
##               <character> <character>
## PIF_178           PIF_178    cachexic
## PIF_087           PIF_087    cachexic
## PIF_090           PIF_090    cachexic
## NETL_005_V1   NETL_005_V1    cachexic
## PIF_115           PIF_115    cachexic
## ...                   ...         ...
## NETCR_019_V2 NETCR_019_V2     control
## NETL_012_V1   NETL_012_V1     control
## NETL_012_V2   NETL_012_V2     control
## NETL_003_V1   NETL_003_V1     control
## NETL_003_V2   NETL_003_V2     control
```

En aquest cas únicament tenim la categoria de "muscle loss", és a dir, si el pacient té caquèxia o és un control. Tot i així, podria haver més informació en aquest dataframe, com per exemple, l'edat del pacient, el sexe, l'estadi de càncer en que es troba el pacient (sempre i quan la caquèxia estigui relacionada amb un càncer, el que desconeixem), etc.

## 4.4 Seleccionar subconjunts de dades

Podem fer subsets de les nostres dades per agafar únicament:

a) Aquells pacients que tenen caquèxia

```
se[ ,se$Muscle.loss == "cachexic"]
```

```
## class: SummarizedExperiment
## dim: 63 47
## metadata(0):
## assays(1): counts
## rownames(63): X1.6.Anhydro.beta.D.glucose X1.Methylnicotinamide ...
##   pi.Methylhistidine tau.Methylhistidine
## rowData names(0):
## colnames(47): PIF_178 PIF_087 ... NETCR_016_V1 PIF_116
## colData names(2): Patient.ID Muscle.loss
```

En aquest cas veiem uqe hi ha 47 pacients dels 77 que tenen caquèxia.

b) Pacients control de l'estudi

```
se[ ,se$Muscle.loss == "control"]
```

```
## class: SummarizedExperiment
## dim: 63 30
## metadata(0):
## assays(1): counts
## rownames(63): X1.6.Anhydro.beta.D.glucose X1.Methylnicotinamide ...
##   pi.Methylhistidine tau.Methylhistidine
## rowData names(0):
## colnames(30): PIF_191 PIF_164 ... NETL_003_V1 NETL_003_V2
## colData names(2): Patient.ID Muscle.loss
```

D'altra banda, com podíem esperar, tenim 30 pacients que tenen el paper de controls en l'estudi.

A partir d'aquí podriem fer dos grups i fer un anàlisis diferencial entre pacients amb caquèxia i controls, mitjançant un anàlisi de regressió lineal (típic en microarrays), mitjançant el paquet limma. No obstant també podriem aplicar un test estadístic per dades que no segueixen una distribució normal (el que sol passar en -òmiques) com el test de Wilcoxon, i per últim corregint les dades amb el mètode de Bonferroni per a comparacions múltiples, o més utilitzat encara, el mètode de Benjamini i Hochberg, per tal que el p-value resultant sigui el màxim rigorós possible. Finalment podriem generar un document amb aquest anàlsisi diferencial entre les dues condicions.

## 4.5 Entrega de l'anàlisi en els diferents formats demanats

En el nostre cas, el següent pas a realitzar és la generació del informe derivat d'aquest micro anàlisi, i la generació de un repositori Github, per tal de pujar tant aquest informe, com tots els arxius derivats com son:

### 4.5.1 Objecte contenedor creat amb les dades i metadades (format .Rda)

```r
save(se, file = "objecte_se.Rda")
```

#### 4.5.2 Script de R utilitzat amb els apunts corresponents i metadades en un arxiu markdown

Això no ho farem aquí sinó que obrirem un script a part, ja sigui estàndard de R o markdown

#### 4.5.3 Dades i metadades en dos arxius .txt

```r
write.table(data, file = "data.txt", sep = "\t", quote = FALSE, row.names = TRUE)
write.table(data, file = "meta.txt", sep = "\t", quote = FALSE, row.names = TRUE)
```

#### 4.5.4 URL Github amb tots els documents generats amb aquest informe inclós

Per tal de penjar-ho a Github hem creat un repositori amb el nom "Farran-Centelles-Nuria-PEC1" i mitjançant el GitHub Desktop, hem enllaçat els documents des del meu portàtil al repositori online.A continuació deixo l'enllaç al repositori per poder accedir a tota la informació addicional de les dades:

https://github.com/nuriiia/Farran-Centelles-Nuria-PEC1

## 5 Discussió, limitacions i conclusions de l'estudi

En aquest estudi hem pogut aprendre com s'analitzaria un dataset derivat d'un experiment -omic, on és necessari un anàlisi bioninformàtic per tal d'identificar possibles metabòlits o vies de senyalització afectades per la caquèxia.Tot i així, ens hem quedat en la primera part, que es la creació del contenedor SE, per tal que, a posteriori, es pugui fer un anàlisi diferencial, cosa que durem a terme en les següents unitats. No obstant, abans d'aquest pas també caldria normalitzar les dades obtingudes per tal que l'expressió d'un metabòlit no tingui una major importància que un altre, sols per un tema d'escala de valors. Una vegada haguéssim fet aquest anàlisis diferencial, podríem senyalar algunes vies o metabòlits en concret, que poden estar afectats en el procés de caquèxia, tot fent un anàlisi d'enriquiment. En la mateixa línea, també podríem trobar biomarcadors de diagnòstic o de pronòstic, tot comparant el valor dels metabòlits de cada pacient amb la seva història i perfil clínic.

Finalment, com a observació personal, voldria destacar que el pas més complex o almenys on jo m'he entrebancat, ha sigut el fet de saber en un inici, com s'havia de dividir la taula i presentar-la davant del contenedor SE, per tal que les dades es puguessin carregar correctament i a partir d'aquí, es puguessin generar les diferents cerques i anàlisis. Així doncs, tinc ganes de veure tot el procés posterior per ser capaç de dur a terme un anàlisi global a partir de dades -òmiques, ja provinguin d'un anàlisi de transcriptòmica, metabolòmica o proteòmica.