PAC 1 - MIREIA AGUILÀ

Anàlisis de Dades Òmiques

Mireia Aguilà

Novembre 2024

Taula de continguts

Abstract

La caquèxia és una síndrome caracteritzat per una pèrdua significativa de pes, i conseqüentment una pèrdua de massa muscular, que sol trobar-se relacionada amb certes malalties cròniques greus com poden ser insuficiències renals, cardíaques i hepàtiques i, amb el càncer, entre d'altres. És un síndrome causat per canvis metabòlics complexes en l'organisme que acaben comportant una inflamació continuada en el temps el que contribueix a una degradació de proteïnes musculars importants.

Mitjançant l'anàlisi de 77 mostres d'orina provinents de pacients amb caquèxia i pacients control, s'ha pogut establir una expressió diferencial de diversos metabòlits que estan relacionats amb la degradació muscular que pateixen aquests pacients i, que comporten una possible diana terapèutica a investigar per tal de millorar el nivell de vida de la gent que pateix aquest síndrome derivat de malalties cròniques greus.

Objectiu de l'estudi

L'objectiu d'aquest estudi és el d'analitzar unes dades provinents de 77 mostres d'orina, 47 d'aquestes mostres provinents de pacients amb caquèxia i 30 mostres de pacients control, per tal d'observar si hi ha diferències notables en alguns dels 63 metabòlits estudiats entre els dos grups.

Materials i mètodes

Les dades utilitzades en aquest estudi provenen de 77 mostres d'orina en les quals s'analitza l'expressió de 63 metabòlits diferents. De les 77 mostres, n'hi ha 47 que provenen de pacients amb caquèxia (un síndrome caracteritzada per la pèrdua de massa muscular) i, n'hi ha 30 que provenen de controls sans.

Una vegada aquestes dades provinents de la Universitat de Califòrnia es van pujar a la pàgina web de "metabolomics workbench", el 22 de febrer de 2023, van ser descarregades en format csv per tal de ser analitzades en R versió 4.4.1.

Per tal de realitzar l'anàlisi es va utilitzar Bioconductor versió 3.19, concretament el paquet Summarized Experiment. Posteriorment també es va realitzar un anàlisis estadístic mitjançant R per tal d'observar si hi havia diferències estadístiques significatives en alguns dels metabòlits entre pacients malalts i pacients sants.

Resultats

Per tal de poder realitzar l'anàlisi de les dades, en primera instància necessitem instal·lar el paquet Summarized Experiment de Bioconductor i carregar les dades en format csv del nostre projecte d'interès.

```
# if (!requireNamespace("BiocManager", quietly = TRUE))
#   install.packages("BiocManager")
# BiocManager::install("SummarizedExperiment")
library(SummarizedExperiment) # Obrim el paquet necessari
```

```
## 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

library(dplyr)

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:Biobase':
##
##     combine
```

```
## The following objects are masked from 'package:GenomicRanges':
##
##     intersect, setdiff, union

## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect

## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union

## The following object is masked from 'package:matrixStats':
##
##     count

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

dades <- read.csv("/Users/mireiaaguilabargues/Downloads/human_cachexia.csv") # Carreguem les dades
```

Una vegada tenim les nostres dades necessitem modificar-les per tal de que tinguin el format adequat per poder dur a terme l'anàlisi mitjançant Summarized Experiment. Analitzant el document inicial que tenim, hem pogut observar que la primera fila els diferents metabòlits estudiats, mentre que les primeres dues columnes contenen l'identificador de diversos pacients i indiquen si és un pacient amb la malaltia (caquèxia) o un control sa.

```
# Modifiquem les dades del document csv per tal de poder analitzar-les amb el paquet SummarizedExperiment.

noms_grups <- dades[, 1] # Obtenim el nom dels grups (identificadors dels pacients)
assay_data <- as.matrix(dades[, -1]) # Matriu amb les nostres dades

dades1 <- data.frame(Sample = colnames(assay_data)) # Obtenim les dades del nom dels diferents metabòlits estudiats
```

```
informacio_files <- data.frame(Feature = noms_grups) # Obtenim les dades dels diferents pacients estudiats

# Creem el nostre objecte de summarized experiment
se <- SummarizedExperiment(
    assays = list(counts = assay_data),
    colData = DataFrame(dades1),
    rowData = DataFrame(informacio_files)
)

# Analitzem el nostre objecte de Summarized Experiment
se
```

```
## class: SummarizedExperiment
## dim: 77 64
## metadata(0):
## assays(1): counts
## rownames: NULL
## rowData names(1): Feature
## colnames(64): Muscle.loss X1.6.Anhydro.beta.D.glucose ...
##   pi.Methylhistidine tau.Methylhistidine
## colData names(1): Sample
```

Per tal de veure les dades experimentals inicials podem utilitzar la funció assay(). Si l'utilitzem en aquestes dades podem veure el que ja havíem observat en el document csv, on en la primera fila tenim una sèrie de de metabòlits dels quals s'ha estudiat la seva expressió, i en la primera columna tenim la classificació de si les mostres provenen de pacients caquèxics o de pacients control sans.

```
 # assays(se)$counts # Per tal de no ocupar molt espai, no es reproduiran aquets resultats en el present informe.
```

Una vegada tenim les dades carregades podem fer una visualització per tal de veure quins són els identificadors dels diferents pacients que estan en l'estudi i, quins són els diferents metabòlits que s'han estudiat.

```
rowData(se)
```

```
## DataFrame with 77 rows and 1 column
##          Feature
##      <character>
## 1        PIF_178
## 2        PIF_087
## 3        PIF_090
## 4    NETL_005_V1
## 5        PIF_115
## ...          ...
## 73  NETCR_019_V2
## 74   NETL_012_V1
## 75   NETL_012_V2
```

```
## 76   NETL_003_V1
## 77   NETL_003_V2
```

```
colData(se)
```

```
## DataFrame with 64 rows and 1 column
##                                              Sample
##                                         <character>
## Muscle.loss                              Muscle.loss
## X1.6.Anhydro.beta.D.glucose X1.6.Anhydro.beta.D...
## X1.Methylnicotinamide        X1.Methylnicotinamide
## X2.Aminobutyrate                  X2.Aminobutyrate
## X2.Hydroxyisobutyrate        X2.Hydroxyisobutyrate
## ...                                             ...
## cis.Aconitate                          cis.Aconitate
## myo.Inositol                            myo.Inositol
## trans.Aconitate                      trans.Aconitate
## pi.Methylhistidine                pi.Methylhistidine
## tau.Methylhistidine              tau.Methylhistidine
```

S'ha intentat utilitzar la funció metadata() per tal d'obtenir informació relacionada amb els mètodes experimentals i publicacions de referència de les dades utilitzades, però no s'ha trobat informació.

```
metadata(se)
```

```
## list()
```

Seguirem l'anàlisi de les dades observant si hi ha diferències significatives (amb un p-valor < 0.05) entre els diferents metabòlits estudiats depenent de si la mostra prové d'orina d'humans amb caquèxia o si prové d'orina d'humans sans.

Primer mirarem la mitjana d'expressió d'algun dels metabòlits per tal de veure si hi ha diferències aparents (sense necessitat de ser estadísticament significatives) entre els dos grups.

```
# És important assegurar-nos que la variable Muscle.loss és un factor aba
ns de començar l'anàlisi
dades$Muscle.loss <- as.factor(dades$Muscle.loss)

# Calculem la mitjana per cadascun dels metabòlits depenent de si són "co
ntrols" o individus caquèxics
mitjana <- aggregate(. ~ Muscle.loss, data = dades[, -1], FUN = mean)  #
Excloem la primera columna ja que és l'identificador del pacient i no ens
interessa en aquest moment.

# Degut a la gran quantitat de columnes que hi ha, i per tal de fer-ho re
produïble per a l'informe, mostrarem els resultats dels 5 primers metabòl
its únicament
mitjana_1_5 <- mitjana[, 1:6]
mitjana_1_5
```

```
##   Muscle.loss X1.6.Anhydro.beta.D.glucose X1.Methylnicotinamide
## 1    cachexic                   128.68894              70.56426
## 2     control                    69.50533              73.15500
##   X2.Aminobutyrate X2.Hydroxyisobutyrate X2.Oxoglutarate
## 1        23.669149              43.23766       183.11043
## 2         9.528333              27.87100        85.51733
```

A primera vista i observant únicament l'expressió en cadascun dels grups dels 5 primers metabòlits, ja es pot observar que és probable que hi hagi diferències en l'expressió entre els pacients sans i els pacients malalts. Això es pot veure, per exemple, en l'expressió de l'anhidro-beta-d-glucosa i de l'oxoglutarat, on els pacients caquèxics tenen, de mitjana, el doble d'expressió.

Ara crearem una taula obtenint el p-valor de cadascun dels metabòlits per tal de veure si aquestes diferències entre grups són, o no, significatives.

```r
# Crearem un data.frame per tal d'introduir els valors per cadascun dels metabòlits
resultats <- data.frame(
  Metabolits = character(),
  p_value = numeric(),
  stringsAsFactors = FALSE
)

# Creem un loop per cadascuna de les columnes que contenen dades d'expressió dels metabòlits. És per aquest motiu que comencem a partir de la tercera columna, ja que les dues primeres contenen l'identificador del pacient i a quin grup pertanyen.

for (metabolit in names(dades)[3:ncol(dades)]) {
  test_resultat <- t.test(
    dades[[metabolit]] ~ dades$Muscle.loss,
    data = dades,
    alternative = "two.sided"
  )

  # Guardem els resultats
  resultats <- rbind(resultats, data.frame(
    Metabolits = metabolit,
    p_value = test_resultat$p.value
  ))
}

# Mirem els resultats inicials
head(resultats)
```

```
##                   Metabolits     p_value
## 1 X1.6.Anhydro.beta.D.glucose 0.035319432
## 2       X1.Methylnicotinamide 0.943536744
## 3            X2.Aminobutyrate 0.007859048
```

```
## 4       X2.Hydroxyisobutyrate 0.004893295
## 5             X2.Oxoglutarate 0.158578752
## 6         X3.Aminoisobutyrate 0.102194569
```

L'important és veure quins són els metabòlits que estan expressats de manera significativament diferents entre els dos grups

```
resultats %>%
  filter(p_value < 0.05)
```

```
##                      Metabolits      p_value
## 1  X1.6.Anhydro.beta.D.glucose 3.531943e-02
## 2             X2.Aminobutyrate 7.859048e-03
## 3        X2.Hydroxyisobutyrate 4.893295e-03
## 4           X3.Hydroxybutyrate 1.334590e-04
## 5        X3.Hydroxyisovalerate 3.458034e-03
## 6            X3.Indoxylsulfate 4.849645e-03
## 7                      Acetate 1.740727e-03
## 8                      Adipate 7.515595e-03
## 9                      Alanine 2.960803e-04
## 10                  Asparagine 3.566603e-03
## 11                     Betaine 2.174920e-03
## 12                   Carnitine 3.192506e-02
## 13                     Citrate 5.098507e-03
## 14                    Creatine 1.984702e-02
## 15                  Creatinine 1.369605e-04
## 16               Dimethylamine 5.348990e-05
## 17                Ethanolamine 2.011237e-02
## 18                     Formate 4.909304e-03
## 19                      Fucose 1.863466e-03
## 20                    Fumarate 2.620524e-02
## 21                     Glucose 9.239445e-03
## 22                   Glutamine 3.391600e-04
## 23                     Glycine 1.304117e-02
## 24                   Glycolate 4.940718e-02
## 25                   Hippurate 1.015460e-02
## 26                   Histidine 4.541154e-03
## 27                     Leucine 2.662822e-05
## 28                 Methylamine 1.496975e-03
## 29         N.N.Dimethylglycine 2.114015e-05
## 30           O.Acetylcarnitine 2.418993e-02
## 31               Pyroglutamate 7.289814e-05
## 32                    Pyruvate 7.357362e-03
## 33                 Quinolinate 2.708732e-05
## 34                      Serine 8.173888e-04
## 35                   Succinate 3.676506e-03
## 36                     Taurine 1.418748e-02
## 37                   Threonine 1.183249e-03
## 38                Trigonelline 3.219917e-03
## 39      Trimethylamine.N.oxide 1.586788e-02
```

```
## 40                  Tryptophan 8.885986e-04
## 41                    Tyrosine 4.456642e-03
## 42                      Valine 1.574212e-05
## 43               cis.Aconitate 6.190184e-04
## 44                myo.Inositol 3.611776e-04
## 45             trans.Aconitate 1.953314e-02
## 46         tau.Methylhistidine 1.710741e-02
```

I això ens mostra que, dels 63 metabòlits estudiats, n'hi ha 46 d'ells en els quals s'observa una diferència d'expressió estadísticament significativa entre els dos grups, pel que amb l'estudi d'orina és possible diferencia pacients caquèxics de pacients sans.

Discussió, limitacions i conclusions de l'estudi

Discussió

La caquèxia és una síndrome caracteritzat per una pèrdua significativa de pes, i conseqüentment una pèrdua de massa muscular, que sol trobar-se relacionada amb certes malalties cròniques greus com poden ser insuficiències renals, cardíaques i hepàtiques i, amb el càncer, entre d'altres. És un síndrome causat per canvis metabòlics complexes en l'organisme que acaben comportant una inflamació continuada en el temps el que contribueix a una degradació de proteïnes musculars importants.

Els resultats obtinguts en l'anàlisi de les dades proporcionades ha mostrat que hi ha diferències significatives en la presència de diversos metabòlits entre pacients que pateixen caquèxia i controls sans, ja que de 63 metabòlits estudiats hi ha diferències significatives en 46 d'ells.

Els resultats mostren que hi ha diferències significatives en la presència d'aminoàcids, on es pot veure que els pacients amb caquèxia tenen una major presència d'aquests en orina, comparat amb els pacients sans. El desequilibri en la presència d'aminoàcids és un símptoma que es pot utilitzar com a un marcador biològic de la caquèxia degut a que indica destrucció muscular i l'ús d'aquests aminoàcids alliberats de les proteïnes musculars com a font d'energia.

La majoria d'altres elements que es troben alterats reforcen la idea de la continuada inflamació i degradació muscular a la que els pacients amb caquèxia estan sotmesos, el que duu a la necessitat d'estudiar i trobar teràpies que puguin intentar reduir el ritme de degradació muscular així com l'estat nutricional d'aquests pacients.

Limitacions

Tot i que l'estudi mostra diferències significatives en diversos metabòlits depenent de l'estat de salut dels pacients, conté diverses limitacions. La primera limitació és deguda a la poca informació dels pacients dels quals s'han obtingut les diverses mostres d'orina, així com informació de l'hora de recollida d'aquesta orina. La

coneixença d'altres factors i característiques dels pacients que s'estan estudiant donaria més garanties de que l'estudi realitzat s'ha fet de manera correcte, ja que els pacients amb caquèxia i els pacients sans es podrien emparellar segons aquestes característiques, per tal d'obtenir una comparació més fiable. Ara mateix, no sabem cap altra característica dels pacients, el que implica que aquests poden tenir altres malalties presents en el moment de l'estudi, poden ser d'edats i sexes diferents i, tot això, pot haver acabat influenciant els resultats introduïts inicialment en l'estudi i, per tant, un biaix.

Una altra limitació és el limitat nombre de pacients que s'han tingut en compte a l'hora de fer l'estudi, ja que per tal d'obtenir uns resultats més robustos seria ideal tenir una major mostra per a l'estudi dels diferents metabòlits.

Conclusions

La caquèxia és un síndrome greu que causa grans desregulacions metabòliques en la gent que la pateix, degut a una inflamació continuada en el temps que comporta una degradació important de proteïnes musculars.

El coneixement de la presència diferencial de certs metabòlits, com aminoàcids i 1,6-Anhidro-β-D-glucosa, entre d'altres, dona la possibilitat de trobar dianes terapèutiques permetin el desenvolupament de teràpies que ajudin a reduir la inflamació i la subseqüent degradació muscular i desnutrició i, per tant, que ajudin a millorar el nivell de vida de la gent que pateix la malaltia.

Link al repositori de github

https://github.com/miagba95/PAC1-dades-omiques.git