PAC1_analisi_de_dades_omiques

Helena Álvarez Álvarez

2025-04-02

Url: https://github.com/helenaalvareza/Alvarez-Alvarez-Helena-PAC1

Taula de Continguts

Abstarct

En aquest treball s'analitza la base de dades GatricCancer_NMR, amb la qual es vol extreure un perfil metabòlic per diagnosticar el càncer gàstric amb mostres d'urina, mitjançant l'objecte de classe SummarizedExperiment i fent un anàlisi exploratori, el qual inclou una PCA.

En els resultats es veu que hi ha múltiples valors perduts i també veiem que no hi ha clústers diferenciats i per tant, no es pot extreure un perfil metabolòmic únic per els pacients amb càncer gàstric a partir dels anàlisi realitzats en aquest treball.

Introducció

L'objectiu d'aquest treball és realitzar, de manera general, un procés d'anàlisis de dades òmiques utilitzant les eines introduïdes durant el curs fins ara, en concret Bioconductor, git i l'exploració multivariant de dades òmiques, per després poder fer una interpretació dels resultats des d'un punt de vista biològic.

En concret, l'objectiu és fer un anàlisis exploratori d'una base de dades, de GatricCancer_NMR, utilitzant SummarizedExperiment, per després poder extreure uns resultats els quals es pugui interpretar des d'un punt de vista biològic. En el nostre cas, ens interessa investigar si hi ha un perfil metabolòmic urinari concret per persones amb càncer gàstric.

Material i mètodes

Materials

Per aquest treball s'ha utilitzat la base de dades anomenada GatricCancer_NMR utilitzada en l'estudi de Chan et al. (2016), que ha estat dipositada a Metabolomics Workbench data repository i posteriorment utilitzada en el tutorial Basic Metabolomics Data Analysis Workflow".

Es tracta d'una base de dades en el que intenten identificar un perfil metabolomic urinari únic per la detecció de càncer gàstric. Hi ha un total de 140 mostres de les quals 43 tenen càncer gàstric (GC), 40 tenen malaltia gàstrica benigna (BN), 40 son individus sans (HE) i 17 (QC) son controls. A més, s'ha analitzat la concentració 77 metabòlits diferents (149 mostres de metabòlits en total, alguns repetits).

Hem triat aquesta base de dades degut a que és un dataset bastant complet, amb informació addicional suficient per comprendre clarament les dades i amb una mostra realtivament gran (140 mostres) per poder extreure uns resultats concloents.
A més, es tracta d'un tema interessant, la perfilació metabòlica, amb la que si s'arriba a uns resultats concloents podria ajudar a una detecció més precoç de la malaltia, el que és de gran importància clínica.

S'ha utilitat la base de dades en format xlsx.

Metòdes

Per analitzar la base de dades s'ha utilitzat el software R studio amb Bioconductor, utilitzant les llibreries: SummarizedExperiment, Biobase, ggplot2, readxl, FactoMineR, factoextra i missMDA.

Primer s'han separat les dades en dos objectes, un anomenat data_df amb la informació de les mostres i un altre amb la informació sobre els metabòlits (quin és cada), peak_df.
Seguidament, hem convertit els resultats de la concentració dels metabòlits en una matriu i hem fet dos DataFrames, un amb la informació sobre les mostres (Idx, SampleID, SampleType, Class) i un altre amb la informació dels metabòlits (Name, Level). Finalment, hem creat l'objecte amb classe SummarizedEperiment, que conté les dades i metadades, anomenat se_object i hem fet el fitxer .Rda.

Un cop havíem creat l'objecte, hem fet un anàlisi exploratòria, per tenir un coneixement general de les dades de la base de dades. Primer, hem fet un resum general de l'objecte se_object i hem mirat quantes mostres hi havia de cada classe (tipus d'individu). També hem analitzat el percentatge de valors absents per mostra i s'ha fet una representació amb un gràfic de barres.
Finalment, hem fet un Anàlisis de Components Principals (PCA) per reduir la dimensionalitat de les dades, i hem generat un gràfic per poder veure i entendre millor els resultats.

Resultats

Exercici 1

```
library(SummarizedExperiment)
library(Biobase)
library(ggplot2)
library(readxl)
library(FactoMineR)
library(factoextra)
library(missMDA)

data <- read_excel("GastricCancer_NMR.xlsx")

head(data)
```

```
## # A tibble: 6 × 153
##     Idx SampleID SampleType Class    M1      M2    M3    M4    M5
M6    M7
##   <dbl> <chr>    <chr>      <chr> <dbl>   <dbl> <dbl> <dbl> <dbl>
<dbl> <dbl>
## 1     1 sample_1 QC         QC     90.1   492.  203.   35   164.
19.7  41
## 2     2 sample_2 Sample     GC     43     526.  130.   NA   694.  114.
37.9
## 3     3 sample_3 Sample     BN    214.  10703.  105.   46.8 483.  152.
110.
## 4     4 sample_4 Sample     HE     31.6    59.7  86.4  14    88.6
10.3 170.
```

```
## 5     5 sample_5 Sample     GC     81.9   259.  315.    8.7 243.
18.4 349.
## 6     6 sample_6 Sample     BN    197.    128.  862.   18.7 200.
4.7  37.3
## # ⓘ 142 more variables: M8 <dbl>, M9 <dbl>, M10 <dbl>, M11 <dbl>,
M12 <dbl>,
## #   M13 <dbl>, M14 <dbl>, M15 <dbl>, M16 <dbl>, M17 <dbl>, M18 <dbl>,
## #   M19 <dbl>, M20 <dbl>, M21 <dbl>, M22 <dbl>, M23 <dbl>, M24 <dbl>,
## #   M25 <dbl>, M26 <dbl>, M27 <dbl>, M28 <dbl>, M29 <dbl>, M30 <dbl>,
## #   M31 <dbl>, M32 <dbl>, M33 <dbl>, M34 <dbl>, M35 <dbl>, M36 <dbl>,
## #   M37 <dbl>, M38 <dbl>, M39 <dbl>, M40 <dbl>, M41 <dbl>, M42 <dbl>,
## #   M43 <dbl>, M44 <dbl>, M45 <dbl>, M46 <dbl>, M47 <dbl>, M48 <dbl>,
…
```

Hem Carregat els pàquets necessaris per resoldre tota l'activitat, i càrreguem les dades en format xlsx.

Exercici 2

```
data_path <- "GastricCancer_NMR.xlsx"
data_df <- read_excel(data_path, sheet = "Data")
peak_df <- read_excel(data_path, sheet = "Peak")

metabolite_data <- as.matrix(data_df[, 5:ncol(data_df)])
rownames(metabolite_data) <- data_df$SampleID
metabolite_data <- t(metabolite_data)



sample_info <- data_df[, 1:4]
rownames(sample_info) <- data_df$SampleID

## Warning: Setting row names on a tibble is deprecated.

sample_info <- as.data.frame(sample_info)
rownames(sample_info) <- sample_info$SampleID



feature_info <- data.frame(peak_df[, c("Name", "Label")])
rownames(feature_info) <- peak_df$Name

se_object <- SummarizedExperiment(
    assays = list(counts = metabolite_data),
    colData = sample_info,
    rowData = feature_info
)

save(se_object, file = "SummarizedExperiment_GastricCancer.Rda")
```

Generem un objecte a partir del document xlsx mitjançant SummarizedExperiment.

ExpressionSet es va desenvolupar abans de SummarizedExperiment, i a més es va dissenyar expressament per gestionar dades de microarrays, a diferència que SummarizedExperiment que està pensada per estructures més modernes amb alta dimensionalitat com RNA-seq. Els elements principals també són diferents, en ExpressionSet son: exprs, pData, fData, experimentData i annotation; mentre que en SummarizedExperiment son: assay, colData, rowData i metadata.

En general, SummarizedExperimet és més flexible que ExpressionSet.

Exercici 3

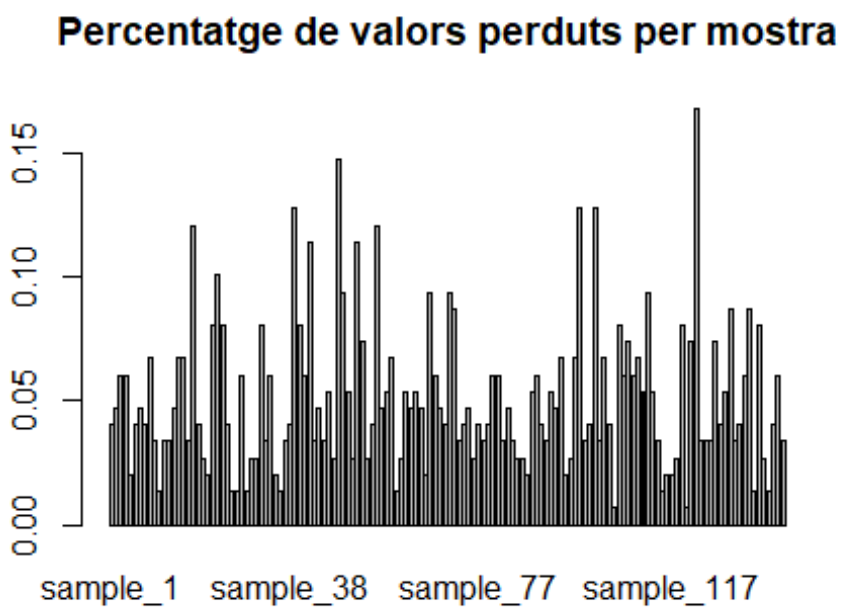
```
se_object
```

```
## class: SummarizedExperiment
## dim: 149 140
## metadata(0):
## assays(1): counts
## rownames(149): M1 M2 ... M148 M149
## rowData names(2): Name Label
## colnames(140): sample_1 sample_2 ... sample_139 sample_140
## colData names(4): Idx SampleID SampleType Class
```

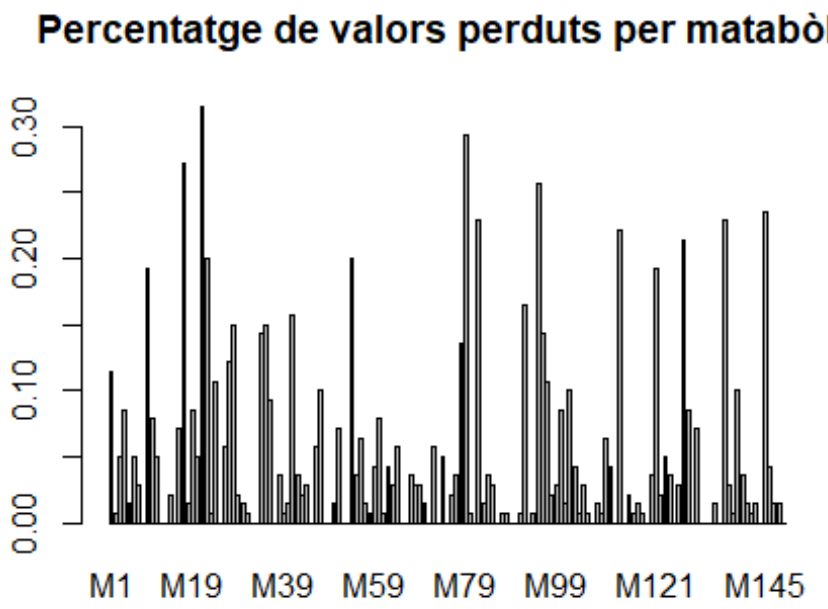
```
table(colData(se_object)$Class)
```

```
##
## BN GC HE QC
## 40 43 40 17
```

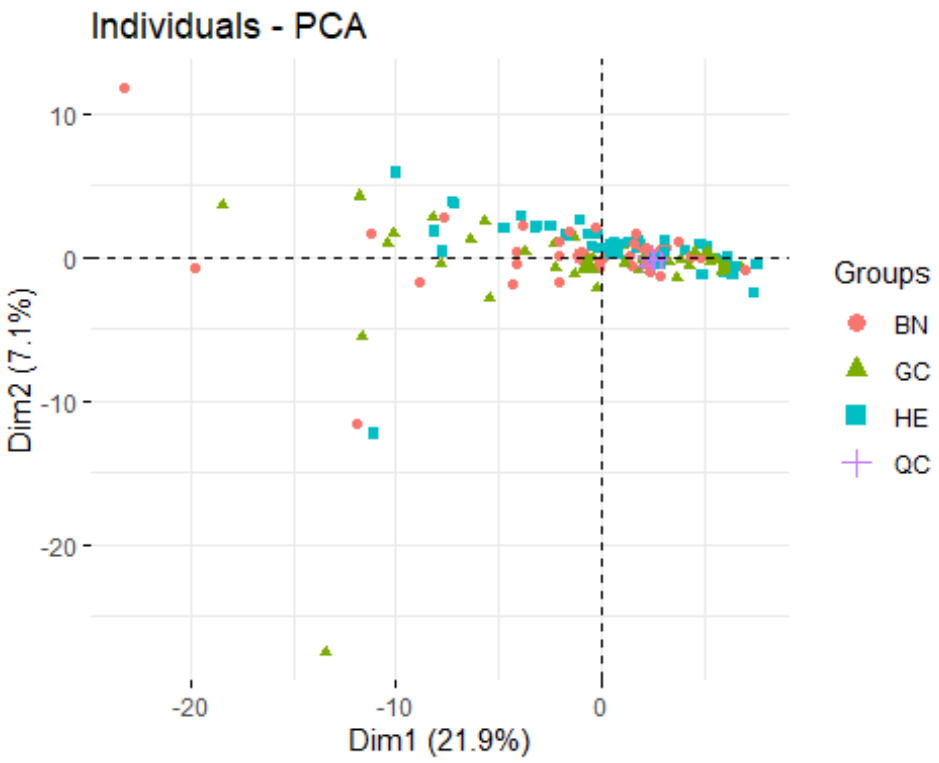
```
#Valors perduts per mostra
missing_data <- colMeans(is.na(assay(se_object)))
barplot(missing_data, main = "Percentatge de valors perduts per mostra")
```




```
mean_missing_data <- mean(missing_data)
mean_missing_data

## [1] 0.0512464

missing_data_1 <- rowMeans(is.na(assay(se_object)))
barplot(missing_data_1, main = "Percentatge de valors perduts per
matabòlit")
```




```
#PCA
corrected_data <- t(assay(se_object))
sum(is.na(corrected_data))

## [1] 1069

corrected_data[is.na(corrected_data)] <- apply(corrected_data, 2, mean,
na.rm = TRUE)

## Warning in corrected_data[is.na(corrected_data)] <-
apply(corrected_data, :
## número de elementos para sustituir no es un múltiplo de la longitud
del
## reemplazo

corrected_data <- imputePCA(corrected_data, ncp = 2)
pca_result <- prcomp(corrected_data, scale = TRUE)

fviz_pca_ind(pca_result, label = "none", habillage =
colData(se_object)$Class)
```



En se_objecte hi ha un total de 149 files que contenen els diferents metabòlits, i 140 columnes que són les diferents mostres dels individus. Trobem que en total hi ha un total de 1069 apartats amb valors absents. El màxim d'una mostra amb valors absents és un 15%. I la mitjana de valors perduts per cada individu és de 0,05. Veiem que tots els individus tenen algun valor perdut mentre que no tots els metabòlits tenen valors perduts, però hi ha que tenen bastants més, sent el màxim 0,3.

Veiem que en la PCA, el primer component explica un 21,9% de la diversitat de les mostres, mentre que la segona indica 7,1%. Veiem també que la majoria de mostres han format un clúster, menys algunes que es troben individualment separades, majoritàriament aquestes son GC i BN.

Discussió

Veient els resultats no podem extreure cap conclusió definitiva. Podem dir que hi ha més valors perduts per metabòlit que per mostra, el que es lògic, ja que si mirem els noms dels metabòlits hi ha alguns repetits i pot ser que en alguns dels casos no sempre s'hagi buscat aquet repetit, a demès i ha més metabòlits que mostres i aconseguir les concentracions de algun metabòlit en concret pot ser més complicat.
Tot i això, veient els resultat no hi ha cap mostra o cap metabòlit en concret que tingui una qualitat dolenta i s'hagi hagut d'eliminar. Això és degut a que aquestes dades ja han estat processades un cop (Metabolomics Workbench: NIH Data Repository, s.f.) i per tant, és probable que les mostres de més baixa qualitat ja hagin estat eliminades.

Per altra, banda podem dir que la PCA és poc indicativa. Això és degut a varis motius, el primer és que els dos components, Dim1 i Dim2 expliquen relativament poc de la diversitat de la mostra, un 21,9% i un 7,1% és menys del 30% de la diversitat. A més, no veiem clústers diferenciats segons la classe de pacient, sinó que, més aviat, hi ha un gran clúster on s'agrupen la majoria de mostres, i només algunes concretes que estan separades individualment.

Addicionalment, s'ha de tenir en compte que aquestes dades s'han aconseguit amb mostres d'orina, les quals sabem que reflecteixen una metabolització més ràpida que altres tipus de mostra, com la sang (Bouatra et al., 2013), i a més hi ha moltes diferències segons quan i com s'han agafat les mostres. Això implica que la mostra probablement no es tan exacte com altres tipus, i que si el mostreig no es fa adequadament pot suposar un biaix en els resultats.

Per aquests motius, amb només els primers anàlisis realitzats no podríem dir que hi ha una diferencia clara en el perfil metabolòmic de persones amb càncer gàstric amb mostres d'orina.
Tan mateix, els anàlisis realitzats només ens proporcionen una visió general de les dades. Si entrem en més detall, com fan en el article, s'ha vist que sí que les mostres d'orina d'aquest estudi sí tenen un perfil metabolòmic diferent (Chan et al., 2016), el qual té concretament tres metabòlits discriminatoris per individus amb càncer gàstric: 2-hydroxyisobutyrat, 3-indoxylsulfat, i alanina. La identificació d'aquest perfil metabòlic pot tenir, en un futur, un potencial clínic perla perfilació metabòlica, la qual ajudaria a una detecció precoç de GC.
No obstant les desavantatges de les mostres d'orina, aconseguir resultats fiables amb aquestes és un gran benefici ja que el seu mostreig és molt més flexible que altres tipus, el que simplifica la seva recuperació i faria molt més accessible la aplicació de futures aplicacions clíniques d'aquest perfils metabòlics.

Conclusió

D'aquest primer anàlisi exploratòria queda clar que les dades amb les que s'ha treballat han estat processades, ja que no sembla que hi hagi dades de baixa qualitat.
El que no queda clar és si realment podem dir que hi hagi un perfil metabòlic urinari determinat per persones amb càncer gàstric, ja que no veiem diferents clústers per les diferents classes de mostres. Però, per anàlisis més profunds, utilitzant estadística univariant i multivariant, l'estudi de Chan et al., 2016, del que provenen les dades, sí que han vist que hi ha un perfil metabòlic determinat, el qual té un gran potencial clínic en la detecció precoç d'aquest càncer, i a més que sigui mostres d'orina facilita la implementació de futurs plans de detecció.

Bibliografia

Bouatra S, Aziat F, Mandal R, Guo AC, Wilson MR, et al. (2013) The Human Urine Metabolome. PLOS ONE 8(9): e73076. https://doi.org/10.1371/journal.pone.0073076

Chan, A., Mercier, P., Schiller, D. et al. 1H-NMR urinary metabolomic profiling for diagnosis of gastric cancer. Br J Cancer 114, 59–62 (2016). https://doi.org/10.1038/bjc.2015.414

Metabolomics Workbench: NIH Data Repository. (s.f.). https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Project&ProjectID=PR000699

Tutorial1: basic Metabolomics Data Workflow. (s. f.). https://cimcb.github.io/MetabWorkflowTutorial/Tutorial1.html