Informe de la PEC1 d'Anàlisi de dades òmiques

Aida Perramon Malavez

2024-11-03

Contents

Abstract **2**

Objectius de l'Estudi **2**

Materials i Mètodes **3**
Requeriments computacionals per a l'anàlisi . . . 3
Obtenció de les dades . . . 3
Estructura i descripció de les dades . . . 4
Creació del contenidor de dades . . . 5
Neteja de dades i control de qualitat . . . 6
Anàlisi descriptiu i multivariant . . . 6

Resultats **8**
Neteja de dades i control de qualitat . . . 8
Anàlisi descriptiu i multivariant . . . 11

Discussió i limitacions i conclusions de l'estudi **18**

Generació d'outputs **18**

Referències **19**

Abstract

En aquesta PAC, hem analitzat les dades metabolòmiques de pacients amb càncer gàstric (GC), malaltia gàstrica benigna (BN) i sans (HE), tenint les mostres de control de qualitat (QC), per identificar metabòlits diferencials associats al càncer gàstric, principalment. Creant un contenidor de dades tipus `SummarizedExperiment`, hem preprocessat els registres filtrant-los per garantir que els metabòlits complissin els estàndards de qualitat (variabilitat <20% i menys del 10% dels valors faltants). Per la resta de dades, hem imputat els valors faltants amb el mètode kNN i normalitzat les dades mitjançant una transformació d'estabilització de la variància (VSN) per minimitzar els efectes no biològics. Posteriorment hem realitzat un estudi descriptiu i multivariant, amb un anàlisi de components principals (PCA) per comprovar la qualitat de les mostres i un de correlació per revelar associacions entre metabòlits. A més, en aquest estudi hem intentat identificar els metabòlits més diferencials entre grups amb un test Mann Whitney-U, centrant-nos en els associats al teixit tumoral (GC) en comparació amb el teixit sa (HE), i hem dissenyat un model estadístic de Bayes per a comprendre la diferència en l'expressió metabòlica entre pacients GC i HE per una banda i BN i HE per una altra. Hem trobat que la sobreexpressió dels metabolits u233 i N-AcetylglutamineDerivative són indicadors de patologia gàstrica, conjuntament amb la infraexpressió de Creatinine. Tanmateix, és la sobreexpressió de u144 i infraexpressió de 2-Hydroxyisobutyrate que diferencien el GC de BN. Altres marcadors que ens podrien indicar GC en comptes de BN serien l'expressió de 2-Furoylglycine i ATP i la infraexpressió de 1-Methylnicotinamide.

Objectius de l'Estudi

L'objectiu d'aquesta PEC és planificar i executar una versió simplificada del procés d'anàlisi de dades òmiques, practicant les metodologies treballades a classe.

Els objectius específics són:

- Aprendre a treballar amb repositoris github (clonar-los i obtenir-ne dades).
- Crear contenidors de dades i metadades òmiques, en particular en format *SummarizedExperiment*.
- Explorar bases de dades òmiques, per a poder saber-ne l'estructura, realitzar un control de qualitat bàsic que ens permeti saber la necessitat de pre-processat previ a ànalisi que requereixen les dades i fer-ne una breu descripció estadística multivariant.
- Presentar de forma clara i organitzada la informació recollida.

Pel que fa a l'objectiu de l'estudi amb les dades triades, és determinar si el càncer gàstric (GC) té un perfil metabolòmic urinari únic en comparació amb els pacients amb malaltia gàstrica benigna (BN) i sans (HE).

Materials i Mètodes

Requeriments computacionals per a l'anàlisi

Hem treballat amb RStudio, amb la versió d'R 4.3.2. Han estat necessàries les següents llibreries per a l'anàlisi i col · lecció de dades:

```
The BiocManager package from Bioconductor and:
- SummarizedExperiment
- pcaMethods
- pmp
- scater
- vsn
- limma
- qmtools
- ComplexHeatmap

Other R packages:
- readxl
- dplyr
- kableExtra
- tidyr
- ggplot2
- reshape2
- gridExtra
- reshape2
- circlize
```

Obtenció de les dades

Hem accedit al `github` de dades que se'ns proporciona a l'enunciat de la PEC. Això ho fem de manera automatitzada accedint al link de descàrrega `.zip` del repositori i fent l'extracció d'aquest en R. D'entre els datasets possibles, hem triat les dades de càncer gàstric (GC) del tutorial del Centre per la Metabolòmica Integrativa i Biologia Computacional (CICMB).

```r
# Descarreguem el contingut del github
download.file(url =
"https://github.com/nutrimetabolomics/metaboData/archive/refs/heads/main.zip"
                                          , destfile = "pec1_metabodata.zip")

# Extraiem el contingut de l'arxiu .zip
unzip(zipfile = "pec1_metabodata.zip")

# Busquem el directori correcte
carpeta_descomprimida <- list.dirs(path = ".", recursive = FALSE,
                  full.names = TRUE)[grepl("metaboData", list.dirs(path = ".",
                                                    recursive = FALSE))]

directori <- list.files(path = carpeta_descomprimida,
                        recursive = TRUE, full.names = TRUE)

# Escollim i carreguem l'arxiu de dades
directori_dades <- directori[grepl("GastricCancer_NMR\\.xlsx$", directori)]
if (length(directori_dades) > 0) {
  data <- read_excel(directori_dades[1], sheet = "Data")
```

```
  peak <- read_excel(directori_dades[1], sheet = "Peak")
}
```

Aquestes dades van ser publicades originalment a l'article de Chan *et al.* (2016) [1] al *British Journal of Cancer*. L'objectiu del seu estudi era identificar si el càncer gàstric té un perfil urinari metabolòmic únic comparat amb la malaltia gàstrica benigna (BN) i pacients sans (HE). Les dades que se'ns presenten són doncs dades urinàries de 43 GC, 40 BN i 40 HE aparellats, que es van analitzar amb resonànica magnètica nuclear ^{1}H (^{1}H-RMN).Els espectres de ^{1}H-RMN es van obtenir al National High Field Nuclear Magnetic Resonance Centre de Canadà (NANUC) amb un espectròmetre Varian Inova de 600 MHz. L'anotació de metabòlits i la deconvolució espectral es van realitzar amb el programari Chenomx NMR Suite v7.6.

Les dades deconvolucionades i anotades estan disponibles al repositori *Metabolomics Workbench* amb l'identificador de projecte `PR000699` i es poden accedir per mitjà del DOI: [10.21228/M8B10B]. Cal destacar que les dades brutes originals de RMN no estan disponibles.

Estructura i descripció de les dades

Les dades obtingudes presenten un format `tibble` de 140x153, és a dir, contenen 140 mostres i 153 *features*, entre les quals trobem 149 metabolits, dues variables identificador, una variable de tipus de mostra (control de qualitat (QC) o mostra (Sample)) i una última variable de classe (QC, GC, BN o HE). Les mostres de control de qualitat agrupades, conegudes com a *pooled QC samples*, serveixen per a assegurar la qualitat en l'anàlisi de les dades metabolòmiques. Aquestes mostres es creen combinant una petita quantitat de cadascuna de les mostres d'un experiment per formar una mescla representativa. La seva importància recau en el fet que proporcionen un punt de referència per controlar la consistència i la fiabilitat de les mesures al llarg de tot el procés analític, és a dir, que ens permeten verificar que l'instrument ha mantingut una resposta estable durant tot l'anàlisi [2].

```
## # A tibble: 6 x 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
## # i 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>, ...
```

A part d'aquestes dades se'ns proporciona la mesura de qualitat per cada metabolit representant la variació en mesura que aquest metabolit ha generat en totes les mostres (QC_RSD), així com el percentatge de valors faltants *missings* per cada metabolit i el seu nom científic o un identificador.

```
## # A tibble: 6 x 5
##     Idx Name  Label                   Perc_missing QC_RSD
##   <dbl> <chr> <chr>                          <dbl>  <dbl>
## 1     1 M1    1_3-Dimethylurate              11.4    32.2
## 2     2 M2    1_6-Anhydro- -D-glucose         0.714  31.2
## 3     3 M3    1_7-Dimethylxanthine            5      35.0
## 4     4 M4    1-Methylnicotinamide            8.57   12.8
## 5     5 M5    2-Aminoadipate                  1.43    9.37
```

```
## 6     6 M6    2-Aminobutyrate                5      47.0
```

Creació del contenidor de dades

A continuació, hem creat un contenidor del tipus `SummarizedExperiment` que contingui les dades i les metadades (informació sobre el dataset, les files i les columnes). És a dir, hem reorganitzat la informació aportada pels dos excels en un sol objecte. I és que la classe `SummarizedExperiment` emmagatzema dades experimentals i les seves metadades, oferint una sincronització entre les dades i les metadades, cosa que facilita el treball amb subgrups i redueix errors. A diferència d' `ExpressionSet`, aquesta classe és més flexible, admetent tant dades basades en GRanges com en DataFrames. [3]

Hem preparat les nostres dades de la següent manera, per incorporar-les al `SummarizedExperiment` que anomenarem `se`. Hem seguit les indicacions i informació de Sánchez, A. [4] i Morgan, M. [5], a part de la guia de Bioconductor [3]. També, ens hem basat en la guia de Li, Y. *et al.* [6], que explica que les dades de metabolòmica es presenten amb les mostres en columnes i les variables en files, és a dir, la matriu que se'ns dóna transposada.

```
# Pas 1: Dades de "assays", és a dir, els valors experimentals d'expressió de
# cada metabolit, com a "counts"

counts_expressio <- as.matrix(data %>% select(starts_with("M")))
counts_expressio <- t(counts_expressio)
colnames(counts_expressio) <- data$SampleID

# Pas 2: Preparem la informació de cada mostra, que es correspon amb les files
# dels "counts" i serien les columnes del SummarizedExperiment. Han de
# coincidir les files amb les columnes de "counts"

info_mostres <- DataFrame(
  TipusMostra = data$SampleType,
  Classe = data$Class,
  row.names = data$SampleID
)

# Pas 3: Tot i que la RowData sol ser la informació de regions d'interès,
# posarem les dades dels metabolits en aquest apartat ja que crec que encaixa
# més que no pas el de metadades

dades_metabolits <- DataFrame(
  Metabolit = peak$Name,
  Label = peak$Label,
  Perc_missing = peak$Perc_missing,
  QC_RSD = peak$QC_RSD,
  row.names = peak$Name
)

# Pas 4: Creem l'objecte SummarizedExperiment amb els "counts", la informació de
# les mostres i les dades dels metabolits

se <- SummarizedExperiment(
  assays = list(counts = counts_expressio),
  colData = info_mostres,
  rowData = dades_metabolits)
```

Un cop construït l'objecte, podem accedir als seus diferents continguts a partir d'uns comands concrets, com

`array(se)` per accedir a l'expressió de metabolits, `colData(se)` per a accedir a la informació de les mostres, i `rowData(se)` per a aaccedir a la informació dels metabolits.

Neteja de dades i control de qualitat

Com hem comentat, se'ns proporciona el percentatge de *missings* presents per a cada metabolit, així com a una mesura de qualitat. Hem representat aquestes variables per metabolit en scatterplots, com es pot veure a l'apartat de resultats. Seguint les consideracions de Broadhurst, D. *et al.* [7] i en vista a l'observat, conservarem només els metabòlits que compleixin amb un QC_RSD inferior al 20% i dels quals faltin menys del 10% dels valors.

```
# Trobem quins metabolits compleixen amb les condicions de filtratge
metabolits_nets <- (rowData(se)$QC_RSD < 20) & (rowData(se)$Perc_missing < 10)

# Filtrem l'objecte agafant només aquests metabolits
se_filtrat <- se[metabolits_nets, ]
```

Hi ha moltes maneres de fer aquest filtratge, infinites llibreries que treballen amb objectes `SummarizedExperiment` amb funcions predefinides però no he aconseguit treballar com volia amb elles. Un exemple n'és `pmp`.

També, a Jankevics, A. i Weber, R.J.M. [8], filtraven les mostres tal que es quedaven amb aquelles que tinguessin menys d'un 10% de *missings*. Hem seguit aquesta mateixa estratègia, ara sí utilitzant la llibreria `pmp` de `Bioconductor`. Després d'aquest filtratge, hem representat els valors faltants per identificar si depenen de la classe (GC, HE, BN, QC).

```
se_filtrat_2 <- filter_samples_by_mv(df=se_filtrat, max_perc_mv=0.1)
```

I, finalment, hem imputat la resta de valors faltants utilitzant el mètode dels k-nearest neighbours (KNN) com explica Joo, J. [9].

```
se_final <- imputeIntensity(se_filtrat_2, i = "counts", name = "knn_counts",
                            method = "knn")
```

Hem fet un gràfic comparatiu entre la distribució inicial dels valors dels metabolits i la distribució post-imputació per a comprovar que no varien substancialment i s'ha dut a terme de forma correcta.

Per últim, hem normalitzat els valors d'expressió dels metabolits, ja que
aquest procediment és necessari per reduir la variabilitat tècnica i assegurar la comparabilitat entre mostres, ja que les diferències en les escales de mesura poden distorsionar els anàlisis.

```
se_norm <- normalizeIntensity(se_final, i = "knn_counts", name = "norm_counts",
                          method = "vsn")
```

I hem verificat la normalització amb un boxplot comparatiu.

D'altra banda, per a comprovar que el control de qualitat estigués ben fet, hem calculat i representat les 2 components principals dels valors experimentals d'expressió genètica.

```
pca <- reduceFeatures(se_norm, i = "norm_counts", method = "pca", ncomp = 2)
```

Anàlisi descriptiu i multivariant

Hem calculat els estadístics descriptius bàsics (mitjana, mediana, desviació estàndard, màxim i mínim) per a cada metabòlit en el conjunt de mostres. A continuació, hem representat aquests valors en boxplots diferenciant per tipus de mostra (GC, HE i BN), excloent els controls de qualitat per evitar possibles interferències en els resultats.

Posteriorment, hem realitzat un anàlisi de correlació per examinar les associacions entre els diferents metabòlits, amb l'objectiu de revelar patrons d'associació entre ells. Els resultats d'aquesta anàlisi s'han visualitzat

en un *heatmap* per facilitar-ne la interpretació, utilitzant la llibreria `ComplexHeatmap` de `Bioconductor`, que també ens proporciona dendograma.

Per cada metabòlit, hem aplicat la prova no paramètrica de Mann-Whitney U per comparar parelles de grups, concretament GC *vs.* HE i BN *vs.* HE. Igual que en l'article de referència [1], hem aplicat la correcció per comparacions múltiples mitjançant el mètode de Benjamini i Hochberg [10]. Establim la significança al 95%.

Finalment, per identificar les característiques metabòliques diferencials entre les classes, hem utilitzat la funció `compareSamples` a través de la interfície del paquet `limma`, que permet calcular estadístiques empíriques de Bayes. D'aquesta manera hem ajustat tres models, per a indicar-nos els metabolits característics de:

- Càncer gàstric respecte pacient sa(na).
- Malaltia gàstrica benigna respecte pacient sa(na).
- Càncer gàstric respecte malaltia gàstrica benigna.

```
fit <- compareSamples(se_norm, i = "norm_counts", group = "Classe",
                      class1 = "HE", class2 = "GC")

fit2 <- compareSamples(se_norm, i = "norm_counts", group = "Classe",
                      class1 = "HE", class2 = "BN")

fit3 <- compareSamples(se_norm, i = "norm_counts", group = "Classe",
                      class1 = "BN", class2 = "GC")
```

Resultats

Neteja de dades i control de qualitat

Hem representat el percentatge de valors faltants per metabolit i la desviació estàndard relativa del control de qualitat a la **Figura 1**.

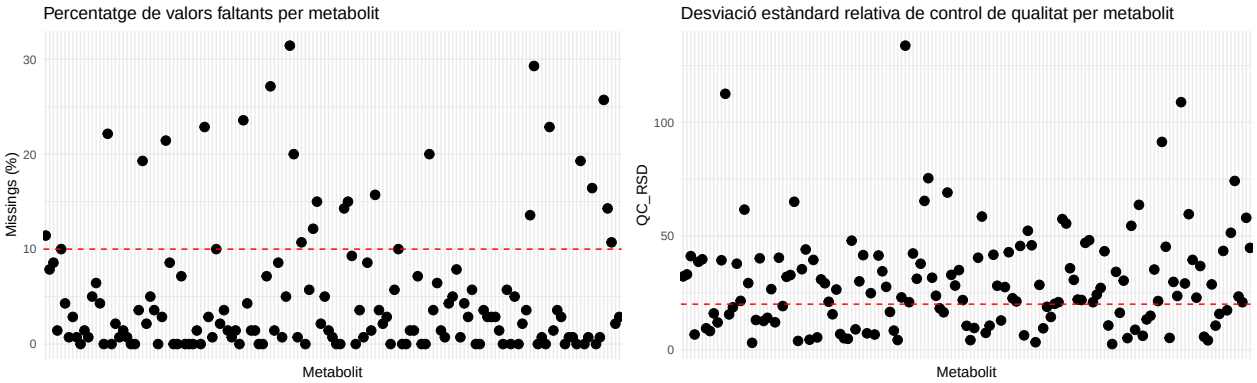

Figure 1: Percentatge de valors faltants i QC-RSD per metabolit a les dades sense pre-processar.

Veiem com la condició sobre QC_RSD és molt més restrictiva que sobre els valors faltants, ja que els *missings* estan concentrats sota el llindar del 10% però per la desviació relativa del control de qualitat visualment triaríem el llindar del 50% per extreure'n *outliers*. Tanmateix, amb la informació que tenim i degut a la naturalesa de la hipòtesi d'investigació, podem ser més restrictius en la qualitat de les dades.

Amb el preprocessat de dades ens hem quedat amb 52 metabolits descrits a la **Taula 1**, en comptes dels 149 d'inici. També hem disminuït el nombre de mostres de 140 a 139.

Table 1: Metabòlits restants després del filtratge

Metabolit	Label
M4	1-Methylnicotinamide
M5	2-Aminoadipate
M7	2-Furoylglycine
M8	2-Hydroxyisobutyrate
M11	3-Aminoisobutyrate
M14	3-Hydroxyisobutyrate
M15	3-Hydroxyisovalerate
M25	6-Hydroxynicotinate
M26	ATP
M31	Adipate
M32	Alanine
M33	Anserine
M36	Asparagine
M37	Azelate
M45	Citrate
M48	Creatinine
M50	Ethanol
M51	Ethanolamine
M65	Glycylproline
M66	Hippurate

M68	Histidine
M71	Ibuprofen
M73	Indole-3-lactate
M74	Isoleucine
M75	Lactate
M88	N-Acetylglutamine
M89	N-AcetylglutamineDerivative
M90	N-Acetylornithine
M91	N-Acetylserotonin
M93	N-Methylhydantoin
M101	Pantothenate
M104	Proline
M105	Propylene glycol
M106	Pyridoxine
M107	Pyroglutamate
M110	Serotonin
M115	Trigonelline
M116	Trimethylamine
M118	Tropate
M119	Tryptophan
M120	Tyrosine
M122	Valine
M126	trans-Aconitate
M129	u11
M130	u1125
M134	u144
M137	u217
M138	u233
M142	u43
M144	u87
M148	-Methylhistidine
M149	-Methylhistidine

Un cop filtrades les dades, hem fet l'anàlisi de components principals i les hem representat a la **Figura 2**.

El fet que les mostres de control de qualitat estiguin tan separades respecte a les altres, afegit a que estan agrupades de manera compacte, suggereix que les dades són de bona qualitat i que les mesures són fiables, ja que no han tingut variació temporal. Tanmateix, es pot observar que les classes BN, GC i HE estan bastant solapades, sobretot BN i GC, el qual indica que no tenen perfils de metabolits molt diferenciats (almenys en les dues primeres components principals). Per tant, tot i que l'experiment estigui ben realitzat, no sembla que hi hagi una diferenciació entre classes directa.

Com a pas final del pre-processat de dades, hem normalitzat l'expressió dels metabolits per a que siguin comparables entre ells. En la **Figura 3** hem representat l'evolució de la distribució dels valors d'expressió dels metabolits per al set original de dades, el filtrat i el finalment normalitzat.

Entre les dades originals i les filtrades la diferència és inapreciable, mentre que veiem com s'ha fet correctament la normalització i els valors d'expressió dels metabolits són comparables entre sí. Previ a aquest pre-processat, l'expressió de per exemple M48 era molt superior a la de la resta de metabolits i, per tant, no permetia fer un anàlisi comparatiu rigurós.

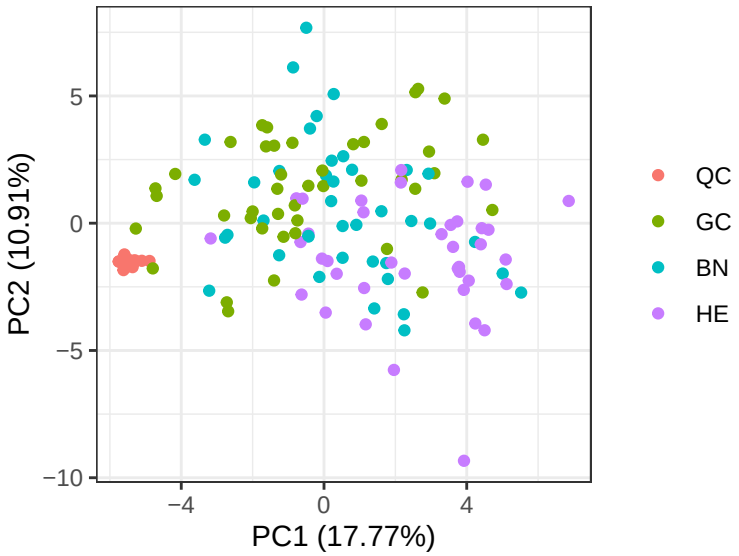

Figure 2: Anàlisi de components principals de les dades filtrades per missings i QC RSD.

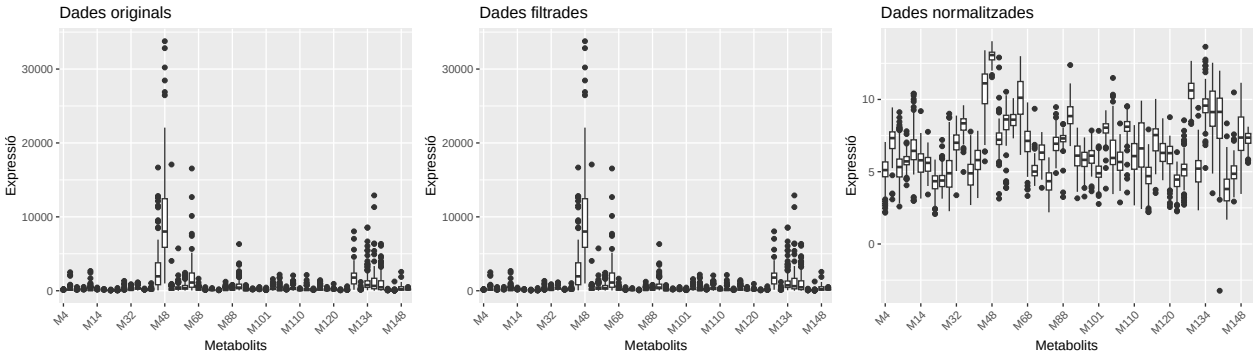

Figure 3: Distribució dels valors d'expressió dels metabolits per a el set de dades (esquerra) original, (centre) filtrat i (dreta) normalitzat.

Anàlisi descriptiu i multivariant

Un cop hem pre-processat les dades, representem l'expressió de cada metabolit per classe de pacient, excloses les mostres de controls de qualitat (**Figura 4**).

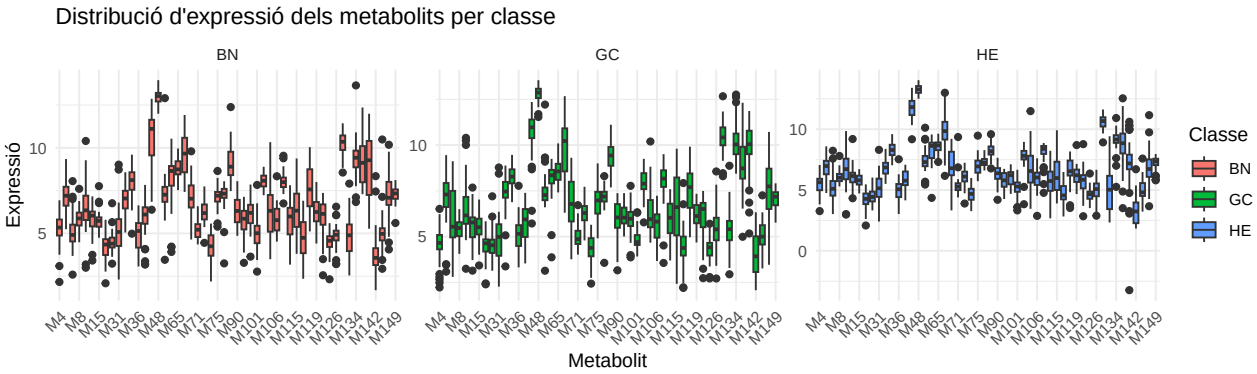

Figure 4: Boxplot de l'expressió dels metabolits per a pacients amb (esquerra) malaltia gàstrica benigna, (centre) càncer gàstric i (dreta) pacients sans.

D'aquesta figura veiem que sembla interessant que l'expressió de la majoria dels metabolits en pacients amb patologia gàstrica sigui benigna o maligna està per sota d'aquella dels pacients sans.

Podem calcular els valors estadístics bàsics per a l'expressió dels diversos metabolits del set final. Trobem a la **Taula 2**, **Taula 3** i **Taula 4** resumits els valors de mitjana, mediana, desviació estàndard (SD), màxim i mínim per a cada metabolit, per a pacients sans, amb càncer gàstric i amb malaltia gàstrica benigna, respectivament.

Table 2: Estadístiques bàsiques dels Metabolits - HE

Metabolit	Mitjana	Mediana	SD	Max	Min
M4	5.399061	5.640917	0.7796095	7.058596	3.267294
M5	6.938587	6.970123	0.8720667	8.591374	4.752164
M7	5.218341	5.100090	1.1542873	8.230811	3.020044
M8	6.072599	6.010363	0.5022027	7.800831	5.297351
M11	6.683694	6.719310	1.4563509	9.848155	2.983502
M14	6.115165	6.157441	0.8516236	9.192350	4.323388
M15	5.796927	5.798067	0.5002042	6.712961	4.717044
M25	4.250925	4.270868	0.7118244	5.463971	2.083696
M26	4.469817	4.431425	0.6034793	6.025711	3.324788
M31	5.162824	5.141672	1.2154787	8.303359	2.411396
M32	6.816808	6.857161	0.6531226	8.382307	5.503455
M33	8.282719	8.292967	0.5965631	9.594678	7.032140
M36	4.988091	5.060729	0.9123973	7.530452	3.027396
M37	5.885125	5.767262	0.7828730	7.079131	4.147025
M45	11.788816	11.816118	0.8789316	13.403254	9.175735
M48	13.249296	13.273142	0.3877200	14.029173	12.505660
M50	7.410062	7.299044	0.9248231	10.122128	5.446148
M51	8.248684	8.628418	0.9835628	9.566685	4.353236
M65	8.585556	8.670951	0.5472969	9.631099	7.308585
M66	9.751992	9.845605	1.4435173	12.995524	6.164718
M68	7.071815	6.791292	1.3812826	9.208128	3.323544

M71	5.378445	5.278514	0.8026187	9.356822	4.391373
M73	6.032678	6.110924	0.6849686	7.242588	3.888002
M74	4.642137	4.683908	0.6689722	5.993642	3.263127
M75	6.945001	6.940445	0.7361415	9.470960	5.726532
M88	7.290248	7.255750	0.3713567	8.281937	6.605097
M89	8.253528	8.193366	0.6328476	9.588751	6.773404
M90	6.272685	6.381333	0.8058070	7.648492	4.197376
M91	5.733665	5.795634	0.6956460	6.739386	4.259325
M93	6.015981	6.219041	0.6486678	7.114179	4.680837
M101	5.253312	5.284585	0.7900653	6.667716	3.323544
M104	7.739310	7.811689	0.9008302	8.956855	3.831432
M105	6.845603	6.560477	1.7075048	11.488921	3.452833
M106	5.881140	6.030301	0.8245836	7.603593	2.875043
M107	8.055478	8.379658	0.8988797	8.865674	4.766593
M110	5.624848	5.738641	1.2993859	8.225457	2.671509
M115	6.289919	5.983491	1.5624968	9.914077	3.896587
M116	4.654935	4.559208	0.7735067	6.483996	2.825190
M118	6.566576	6.547675	1.1014961	8.167888	3.524001
M119	6.164201	6.213584	0.7651035	8.695583	4.667806
M120	5.676613	5.840886	1.1188691	8.772677	2.844086
M122	4.628567	4.649842	0.5592098	6.354332	3.511686
M126	4.987670	5.078142	0.7545402	6.664361	2.888307
M129	10.576650	10.671098	0.5745570	11.622466	8.916369
M130	4.944166	5.025000	1.7369539	9.420591	2.328757
M134	9.101956	9.149483	0.9313625	10.920276	5.259324
M137	8.648830	8.848534	1.8839230	12.540009	3.509472
M138	7.068469	7.185022	2.4160518	10.612284	-3.234942
M142	3.261814	3.237051	1.0864378	6.708644	1.834260
M144	4.981614	4.807754	0.9215220	7.586355	3.225170
M148	6.873138	6.726554	1.4321475	11.154954	3.667273
M149	7.218691	7.342970	0.5578179	7.959022	5.797334

Table 3: Estadístiques bàsiques dels Metabolits - GC

Metabolit	Mitjana	Mediana	SD	Max	Min
M4	4.564822	4.655937	0.9741822	6.513964	2.223900
M5	7.038936	7.301870	1.3280311	9.449837	3.085801
M7	5.977708	5.539621	1.6028886	9.120219	2.800392
M8	5.396466	5.468907	0.5524217	6.499078	4.061069
M11	6.585126	6.158962	1.4785831	10.396614	3.245408
M14	5.552537	5.775617	0.9064547	7.680136	3.142167
M15	5.509330	5.514534	0.7095875	7.026434	3.425359
M25	4.467689	4.636038	0.7542409	5.839416	2.394109
M26	4.591531	4.526968	0.8082657	7.202870	3.136707
M31	4.876643	4.983320	1.3929115	8.776941	2.268399
M32	7.351920	7.462565	0.9941698	8.882072	3.373183
M33	8.240500	8.275795	0.5973853	9.464315	6.882209
M36	5.160581	5.153753	0.9073574	7.776985	3.318894
M37	5.724376	5.942026	1.0885561	7.393387	3.535512

M45	10.635018	10.974613	1.3478726	12.355523	5.717713
M48	12.787611	12.869048	0.4330377	13.545120	11.536438
M50	7.452008	7.268528	1.1779776	12.213048	3.137599
M51	8.100008	8.349310	0.9936400	9.249042	3.870282
M65	8.599716	8.566136	0.6029829	10.091575	7.413288
M66	9.764865	10.223881	1.5884121	12.671928	6.306917
M68	6.866635	6.784038	1.4245152	9.309838	3.763257
M71	4.979240	4.862622	0.6136513	6.719843	4.018178
M73	6.254544	6.290453	0.6900290	7.723453	4.793183
M74	4.383235	4.381587	0.6810166	5.647746	2.433423
M75	6.942118	6.977828	0.9303027	8.999684	4.907423
M88	7.121402	7.192462	0.6780576	8.304603	3.568356
M89	9.394397	9.433445	0.7765327	11.111130	7.688838
M90	6.013300	6.045685	0.9339754	7.374356	3.161845
M91	6.114306	6.034832	0.6064726	7.375829	5.007588
M93	5.896846	5.974918	0.8166469	7.773639	3.867275
M101	4.843808	4.701300	0.5983041	6.328796	3.914970
M104	7.919093	7.890407	0.7150013	9.250060	5.776811
M105	6.314208	5.951871	1.1699691	10.197869	4.670244
M106	5.653836	5.815982	0.9033625	7.234019	3.670620
M107	8.003130	8.169528	1.0517985	9.503769	3.560982
M110	5.953801	6.030378	1.0649596	7.591075	3.696700
M115	6.679518	6.612521	1.9404920	9.804612	2.422135
M116	4.442467	4.378009	1.1053072	7.926793	2.200063
M118	7.563519	7.699778	1.2441911	9.913189	4.967427
M119	6.162191	6.123356	0.6998345	7.339283	4.682202
M120	6.127418	6.506531	1.0738460	7.493398	2.731221
M122	4.363909	4.412825	0.6484644	5.702988	2.716278
M126	5.448458	5.395493	1.2252891	8.560097	2.712766
M129	10.416642	10.418322	0.8698845	12.668141	8.302107
M130	5.381607	5.404258	1.1382385	8.871772	2.943239
M134	10.270572	10.033605	1.0858063	12.760059	8.437880
M137	8.912913	8.727857	1.3800802	12.312047	4.992128
M138	9.804187	10.048761	1.4182030	11.862246	5.172889
M142	4.110044	3.924919	1.2655249	6.327487	2.074947
M144	5.031030	4.999125	0.8405761	7.255506	3.261920
M148	7.569343	7.745087	1.4722781	10.720209	3.462683
M149	7.070406	7.181640	0.4827025	7.910834	5.763308

Table 4: Estadístiques bàsiques dels Metabolits - BN

Metabolit	Mitjana	Mediana	SD	Max	Min
M4	5.265891	5.342231	1.0019836	6.935145	2.163778
M5	7.220313	7.199833	0.8515597	9.347718	5.247833
M7	5.088664	4.910185	1.0311089	8.053636	2.591502
M8	5.863061	5.872883	0.6009290	7.604384	4.483417
M11	6.511096	6.346966	1.4373967	10.399816	2.995865
M14	5.788454	6.038339	0.8355010	7.164282	3.413312

M15	5.691215	5.729295	0.5913469	7.762465	4.746548
M25	4.257035	4.349804	0.7280649	5.796903	2.087107
M26	4.505622	4.398730	0.6176145	6.762071	3.231627
M31	5.107127	5.081667	1.4174922	9.015531	2.306756
M32	6.966467	7.040129	0.8277120	8.571755	5.050754
M33	7.978179	8.114401	0.8699570	9.598561	4.983520
M36	4.939271	5.140677	0.9861046	6.636818	3.069923
M37	5.859527	6.103637	1.0544070	7.833756	3.179863
M45	10.575412	11.105151	1.6322213	12.864961	6.383640
M48	12.959767	13.024347	0.3891255	13.949731	11.992667
M50	7.333020	7.261244	1.2559406	12.901815	3.459051
M51	8.194373	8.643287	1.4506799	10.532898	3.917000
M65	8.821528	8.739135	0.5730330	9.960517	7.530997
M66	9.650003	9.662630	1.2776411	11.919779	7.421878
M68	7.178840	7.025819	1.0491996	9.795415	4.644271
M71	5.192441	5.214311	0.4926422	6.586776	4.357559
M73	6.244203	6.187480	0.6699263	7.744502	4.449270
M74	4.209545	4.250033	0.8115024	5.574254	2.191514
M75	7.083956	7.201737	0.6631203	8.646161	5.392060
M88	7.210945	7.322756	0.8201191	8.047439	3.247629
M89	9.116970	8.898118	1.0253436	12.379440	7.470899
M90	6.374127	6.312022	0.8331289	8.050058	4.859969
M91	5.740890	5.857671	0.8858554	7.139312	3.280771
M93	6.068206	6.198745	0.8063596	7.853096	3.643699
M101	4.940255	5.012460	0.9251786	7.831064	2.768587
M104	8.077532	8.073864	0.3980577	8.891030	7.229618
M105	6.271537	6.319053	1.4730923	10.331353	3.492011
M106	5.819695	5.774095	0.8032645	8.338205	4.520109
M107	8.017870	7.995442	0.5472340	9.537242	6.749618
M110	5.772762	5.986820	1.2157229	7.815018	3.175163
M115	6.445546	6.326950	1.3897209	9.376952	3.807057
M116	4.703590	4.740031	1.1535189	6.907512	2.349071
M118	7.524258	7.569506	1.3685794	10.037848	5.114437
M119	6.223866	6.252502	0.7748570	8.001405	4.411208
M120	5.943604	6.139418	1.1346576	7.365893	2.553238
M122	4.516754	4.570754	0.6282010	5.598058	2.320206
M126	4.828913	4.919966	0.7177505	6.559353	3.205693
M129	10.264669	10.350566	0.6907240	11.432048	8.555056
M130	4.820081	4.890723	1.2455993	7.908581	2.542419
M134	9.452012	9.419388	1.1602453	13.645441	6.833897
M137	9.175286	9.122003	1.3329180	12.355226	7.290015
M138	8.912481	9.275707	1.9683680	11.990312	5.273470
M142	3.757890	3.570562	1.3078733	8.344720	1.683928
M144	5.054812	4.972156	1.1827073	10.486750	2.935072
M148	7.283321	7.080060	1.3103761	10.177434	4.062280
M149	7.274699	7.334859	0.4584072	8.115265	5.608270

Amb les taules descriptives podem confirmar quantitativament l'observat a la **Figura 4**. Per seguir amb l'exemple anterior, el metabolit M48 té una expressió mitjana de 13.25 per a HE, però de 12.79 i 12.96 per a GC i BN respectivament. D'altra banda, també sembla haver-hi metabolits diferencials de patologia gàstrica,

com l'M118, que té valors mitjans d'expressió superiors per a GC i BN que per a HE.

Per a complementar aquesta informació, hem representat la matriu de correlació dels metabolits en la **Figura 5**. Una manca de correlació es visualitza de color blanc, mentre que la correlació positiva s'indica de vermell i la negativa de color blau.

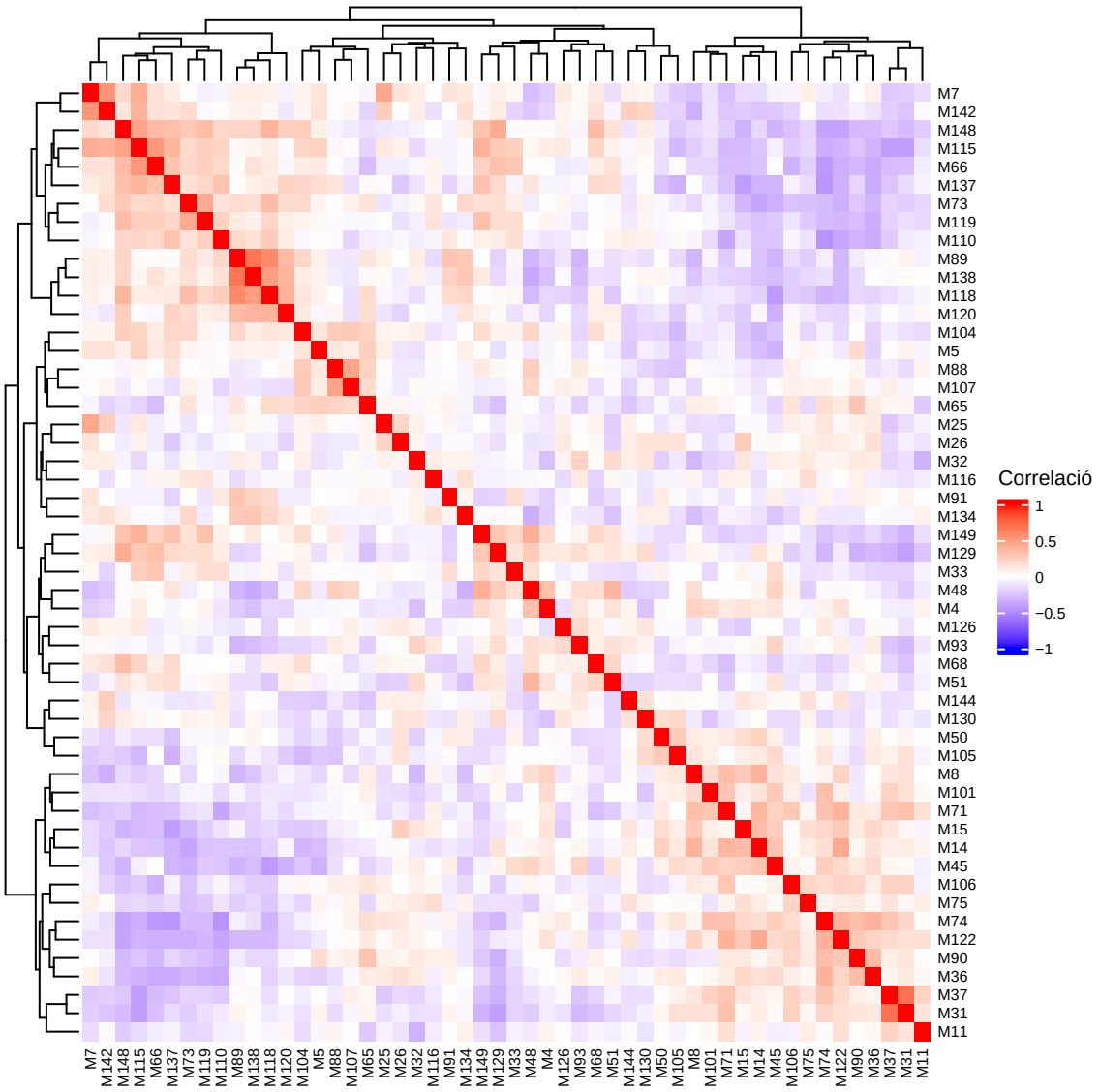

Figure 5: Matriu de correlació i dendogrames entre metabolits, de blau (-1) a vermell (1), passant per blanc (0).

Podem observar que els metabolits s'han reendreçat en aquesta representació. Això és perquè, quan s'utilitza `ComplexHeatmap`, aquest reendreça automàticament les files i columnes per reflectir agrupaments basats en similitud. És a dir, aplica un clustering jeràrquic per defecte, reorganitzant els metabòlits en funció de les seves correlacions, de manera que aquells amb patrons de correlació similars queden més propers al *heatmap*. És d'aquesta manera com ens és més fàcil identificar que hi ha dos grups bastant separats d'expressió de metabolits. Els metabolits M7, M142, M148, M115, M66, M137, M73, M119, M110, M89, M138, M118, M120 i M104 tenen una forta correlació entre ells, i alhora una relació inversa amb l'expressió dels metabolits M105, M8, M101, M71, M15, M14, M45, M106, M75, M74, M122, M90, M36, M37, M31 i M11, que entre ells també estan fortament correlacionats. Això es pot comprovar mirant els dendogrames, on destaquen dos grups principals.

Són particularment intenses les correlacions entre l'M37 i l'M31, és a dir, entre Azelate i Adipate, tot i que aquesta relació no està descrita a la literatura (o no l'he trobat). També veiem una correlació molt forta entre M7 i M142, identificats com 2-Furoylglycine i u43, així com entre l'M89 i M138, el N-AcetylglutamineDerivative i l'u233. Al ser l'M142 i l'M138 metabolits inespecífics no he pogut trobar informació a la literatura sobre aquestes relacions. Pel que fa a correlacions negatives, destacarien per exemple l'M115 amb l'M31, sent l'M115 el metabolit Trigonelline, o de l'M110 amb l'M74 (serotonina amb isoleucina). Per a aquesta darrera interacció, he trobat algun article interessant on semblen descriure l'antagonisme de la serotonina amb la leucina, però relatant de manera inespecífica aquesta relació amb la isoleucina [11][12].

Comprovem quantitativament l'existència d'aquestes relacions mitjançant la prova de Mann-Whitney U. Mostrem els metabolits que surten significativament (p-val ajustat < 0.05) diferents per a GC *vs.* HE a la **Taula 5** i per a BN *vs.* HE a la **Taula 6**.

Table 5: Metabolits significativament diferents per a malalts de càncer gàstric respecte pacients san(e)s.

Metabolit	Estadístic	p-value	p-value ajustat
M4	401	0.0000496	0.0003683
M8	303	0.0000007	0.0000118
M14	518	0.0029662	0.0128534
M32	1230	0.0002829	0.0018389
M45	364	0.0000108	0.0001120
M48	369	0.0000133	0.0001155
M71	523	0.0034468	0.0137873
M89	1458	0.0000000	0.0000005
M101	518	0.0029662	0.0128534
M118	1196	0.0009172	0.0052993
M134	1362	0.0000012	0.0000156
M138	1436	0.0000000	0.0000008
M142	1168	0.0022522	0.0117115

Table 6: Metabolits significativament diferents per a pacients amb malaltia gàstrica benigna respecte els san(e)s.

Metabolit	Estadístic	p-value	p-value ajustat
M45	408	0.0002696	0.0070103
M48	482	0.0035316	0.0367284
M89	1210	0.0000254	0.0013186
M118	1087	0.0026516	0.0344711
M138	1115	0.0010379	0.0179905

Veiem que el grup de metabolits que inclouen l'M48, M89, M138 i M118, entre d'altres, sembla estar relacionat amb malaltia gàstrica. És curiós com el càncer gàstric i la malaltia gàstrica benigna comparteixen metabolits representatius, mentre que el càncer gàstric té més indicadors alterats que la malaltia gàstrica benigna.

Finalment, utilitzem estadística Bayesiana per a fer un model que identifiqui quins metabolits són més representatius per al càncer gàstric respecte els HE, i per a la malaltia benigna respecte els HE també. Així com un darrer model que compara GC i BN. Els 5 valors més representatius de cada model estan recollits a la **Taula 7**, **Taula 8** i **Taula 9**, respectivament.

Table 7: Metabolits més representatius de càncer gàstric respecte pacients san(e)s.

	logFC	CI.L	CI.R	p-value	p-value ajustat	B
M138	2.7357170	1.9437191	3.5277149	0e+00	0.0e+00	13.140124
M89	1.1408688	0.8005798	1.4811578	0e+00	0.0e+00	12.124617
M8	-0.6761326	-0.9094442	-0.4428211	1e-07	1.0e-06	7.784863
M134	1.1686161	0.7333365	1.6038957	4e-07	5.5e-06	5.882699
M48	-0.4616848	-0.6385218	-0.2848478	8e-07	8.5e-06	5.245285

Table 8: Metabolits més representatius de malaltia benigna gàstrica respecte pacients san(e)s.

	logFC	CI.L	CI.R	p-value	p-value ajustat	B
M89	0.8634422	0.5171372	1.2097472	0.0000023	0.0001193	4.4738018
M138	1.8440115	1.0380118	2.6500112	0.0000129	0.0003343	2.8424543
M45	-1.2134037	-1.7615307	-0.6652767	0.0000234	0.0004051	2.2791794
M118	0.9576818	0.4428075	1.4725560	0.0003348	0.0043525	-0.2081924
M48	-0.2895290	-0.4694924	-0.1095657	0.0018087	0.0188103	-1.7579627

Table 9: Metabolits més representatius de càncer gàstric respecte malaltia benigna gàstrica.

	logFC	CI.L	CI.R	p-value	p-value ajustat	B
M8	-0.4665952	-0.6983724	-0.2348180	0.0001101	0.0057232	0.9672402
M134	0.8185601	0.3861431	1.2509771	0.0002649	0.0059902	0.1596530
M4	-0.7010690	-1.0788938	-0.3232443	0.0003456	0.0059902	-0.0838570
M7	0.8890444	0.3661084	1.4119805	0.0010002	0.0130020	-1.0514049
M126	0.6195450	0.2381330	1.0009570	0.0016381	0.0170360	-1.4970206

Discussió i limitacions i conclusions de l'estudi

Hem aconseguit identificar quina expressió o manca d'expressió de metabolits s'identifica amb càncer gàstric i malaltia gàstrica benigna, i quins els diferencien. La sobreexpressió dels metabolits u233 i N-AcetylglutamineDerivative són indicadors de patologia gàstrica, conjuntament amb la infraexpressió de Creatinine. Tanmateix, és la sobreexpressió de u144 i infraexpressió de 2-Hydroxyisobutyrate que diferencien el càncer gàstric de la patologia gàstrica benigna, més característica per la sobreexpressió de Tropate i infraexpressió de Citrate. Altres marcadors que ens podrien indicar que és càncer gàstric en comptes de malaltia benigna serien l'expressió de 2-Furoylglycine i ATP i la infraexpressió de 1-Methylnicotinamide.

Els nostres resultats són consistents amb l'article envers el 2-Hydroxyisobutyrate, però a Chan, A. *et al.* [1] els surten especialment rellevants per a identificar càncer gàstric els metabolits 3-indoxylsulfate i alanine. A l'anàlisi de significança estadística, a nosaltres també ens surt un efecte significatiu de l'alanina per a la identificació de càncer gàstric comparat amb pacients san(e)s, però aquest no queda registrat entre les 5 components principals del model de Bayes. Pel que fa la 3-indoxylsulfate, no és un dels 52 metabolits inclosos a l'estudi final, doncs no ha passat el filtratge de *missings* i de qualitat, el que ens pot indicar que potser hem estat massa estrictes en la filtració de les dades i hauríem d'haver admès una QC_RSD més elevada. Els gràfics de dispersió de QC_RSD ja ens indicaven que la majoria de metabolits tenien una QC_RSD sota el 50%, essent la decisió d'incloure només els inferiors al 20% molt restrictiva.

D'altra banda, a l'article original també troben rellevant la infraexpressió de 1-Methylnicotinamide, que raonen és degut a que els pacients BN i GC tenen una pèrdua del sistema de protecció mucós gàstric; i de la creatinina, una reducció deguda a la pèrdua de massa muscular esquelètica. Pel que fa a la infraexpressió del citrat a pacients de GC, essent aquest metabolit rellevant al cicle de Kreb, Chan, A. *et al.* raonen que la seva falta podria estar relacionada amb l'habilitat del GC d'escapar a la mort cel · lular programada.

Amb tot, podem dir que els nostres resultats són consistents amb els de l'estudi original, tot i que el model bayesià difereixi del LASSO o el PLS de l'article en algun metabolit. Això és una mostra més de la inespecificitat de les dades, doncs tot i que hem vist que s'intuïa un perfil de metabolits diferent a pacients amb patologia gàstrica i sans, les divergències són subtils tal com hem vist a l'anàlisi de PCA (Figura 2), el heatmap de correlació (Figura 5) o el boxplot de l'expressió de metabolits per classe (Figura 4).

A més, aquest estudi té limitacions. Les dades són escases amb tan sols un(e)s 40 pacients per grup, el qual limita la potència estadística dels anàlisis. També, la tria d'aquest(e)s pacients finals es va fer per aparellament per edat, sexe i índex de massa corporal, però podria haver-hi altres variables que no s'hagin tingut en compte per falta d'informació. Àdhuc, com ja hem comentat, les restriccions en la neteja de dades han pogut limitar la informació analitzada, tot i garantir que aquesta fos de qualitat. La imputació de missings també ha pogut afectar parcialment en els resultats, tot i que no significativament. Així mateix, la limitada experiència de l'autora d'aquest document en l'anàlisi de dades òmiques, així com la comprensió parcial de l'enunciat d'aquesta PAC, poden haver influit substancialment en el present estudi que, tanmateix, s'ha fet amb la major implicació possible.

En conjunt, s'ha fet un pre-processat extensiu de dades de metabolits, construït un objecte `SummarizedExperiment` per a treballar-hi amb llibreries de `Bioconductor`, analitzat de manera descriptiva i multivariant les dades del present anàlisi i identificat metabolits rellevants en el diagnòstic diferencial de càncer gàstric respecte patologia gàstrica benigna i pacients saludables.

Es pot trobar tota la informació complementària al github:

https://github.com/aid-pm/AnalisiOmiques/tree/main/PEC1

Generació d'outputs

```
save(se_norm, file = "objecte_dades_meta.Rda")
write.table(data, file = "dades_originals.txt", sep = "\t", row.names = FALSE,
            quote = FALSE)
```

```r
write.table(peak, file = "peak_original.txt", sep = "\t", row.names = FALSE,
            quote = FALSE)
```

Referències

[1] 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

[2] Broeckling CD, Beger RD, Cheng LL, et al. Current Practices in LC-MS Untargeted Metabolomics: A Scoping Review on the Use of Pooled Quality Control Samples. Anal Chem. 2023 Dec 26;95(51):18645-18654. doi: 10.1021/acs.analchem.3c02924.

[3] SummarizedExperiment for Coordinating Experimental Assays, Samples, and Regions of Interest [Internet]. [cited 2024 Oct 30]. Available from: https://www.bioconductor.org/packages/devel/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html

[4] Sánchez Pla, Alex. ASPteaching/Omics_Data_Analysis-Case_Study_0-Introduction_to_BioC [Internet]. [cited 2024 Oct 30]. Available from: https://github.com/ASPteaching/Omics_Data_Analysis-Case_Study_0-Introduction_to_BioC?tab=readme-ov-file

[5] Morgan M. Working With Data: SummarizedExperiment [Internet]. 2015 [cited 2024 Oct 30]. Available from: https://www.bioconductor.org/help/course-materials/2015/Uruguay2015/V2-WorkingWithData.html

[6] Li, Y., Wang, C. and Chen, L.: SDAMS: an R Package for differential expression analysis of single-cell RNA sequencing data (Manuscript).

[7] Broadhurst, D., Goodacre, R., Reinke, S.N. et al. Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. Metabolomics 14, 72 (2018). https://doi.org/10.1007/s11306-018-1367-3

[8] Jankevics A, Weber RJM. Peak Matrix Processing for metabolomics datasets [Internet]. [cited 2024 Nov 1]. Available from: https://www.bioconductor.org/packages/release/bioc/vignettes/pmp/inst/doc/pmp_vignette_peak_matrix_processing_for_metabolomics_datasets.html

[9] Joo J. Processing quantitative metabolomics data with the qmtools package [Internet]. 2024 [cited 2024 Nov 1]. Available from: https://bioconductor.org/packages/devel/bioc/vignettes/qmtools/inst/doc/qmtools.html#1_Introduction

[10] Benjamini YHY (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 57 (1): 289–300.

[11] Krishnaswamy K, Raghuram TC. Effect of leucine and isoleucine on brain serotonin concentration in rats. Life Sci. 1972 Dec 22;11(24):1191–7.

[12] Wessels AG, Kluge H, Hirche F, Kiowski A, Schutkowski A, Corrent E, et al. High Leucine Diets Stimulate Cerebral Branched-Chain Amino Acid Degradation and Modify Serotonin and Ketone Body Concentrations in a Pig Model. PLoS One [Internet]. 2016 Mar 1 [cited 2024 Nov 2];11(3). Available from: https://pubmed.ncbi.nlm.nih.gov/26930301/