

PAC1

Joel Martínez Miralles

2024-11-06

Table of Contents

Introducció	1
Materials i Mètodes	2
Disseny Experimental	2
Obtenció de dades crues	2
Inici de l'anàlisi realitzat en l'activitat: selecció i lectura del dataset	2
Generació del contenidor SummarizedExperiment	3
Control de qualitat, preprocessat	3
Normalització i control d'outliers	4
Anàlisi descriptiva	4
Resultats	4
Discussió	10
Annex	10

Introducció

L'adenocarcinoma gàstric és el cinquè càncer més comú del món i el tercer més letal. Sovint, se sol diagnosticar tard i a través de símptomes no específics. La supervivència dels pacients diagnosticats en el primer estadi tumoral (IA) és molt major que els pacients que ja es diagnostiquen en l'estadi immediatament posterior (IB). Aquest fet accentua la importància de desenvolupar i aplicar tècniques de cribratge recurrents en les poblacions d'alt risc, per tal d'aconseguir una detecció precoç del càncer.

La metabolòmica és l'àrea d'estudis "òmics" més elemental en l'escala, i per tant, l'àrea que replica de manera més fidedigne el fenotip d'un organisme. La ressonància magnètica nuclear de protons és una tècnica que permet analitzar el perfil metabòlic d'una mostra d'orina.

Així doncs, l' objectiu d'aquest estudi és identificar si el càncer gàstric té un perfil metabolòmic en orina únic i específic, comparant-lo amb casos de malaltia gàstrica benigna i pacients sans. L'objecte final d'aquesta investigació és el desenvolupament

d'un procediment diagnòstic no invasiu, cost-efectiu, eficient i raonablement acurat pel càncer gàstric.

A nivell tècnic, l'objectiu d'aquesta activitat és posar en pràctica els coneixements adquirits en el primer apartat de l'assignatura a través de l'anàlisi de dades d'un microarray.

Materials i Mètodes

Les dades analitzades provenen de l'article de Chan et al. (2016), titulat "H-NMR urinary metabolomic profiling for diagnosis of gastric cancer".

Disseny Experimental

En aquest estudi es va escollir, amb uns criteris de selecció objectius i determinats, 43 individus amb càncer gàstric (GC), 40 individus amb malalties gàstriques benignes (BN), i 40 individus sans (HE).

El disseny experimental es va estructurar en blocs que inclouen aleatòriament un individu de cada grup: 1 GC, 1 BN, i 1 HE. Es va recol·lectar el perfil metabolòmic de cada individu amb mostres d'orina. L'obtenció d'aquest perfil es va realitzar mitjançant la ressonància magnètica nuclear de protons. El procés experimental es va dur a terme en 4 lots al llarg de 10 dies.

Per garantir la precisió i repetibilitat en la quantificació dels metabòlits, es van incloure mostres de control de qualitat (QC), que consistien en múltiples alíquotes d'orina d'un mateix individu sa, analitzades cada 10 mostres d'estudi.

Obtenció de dades crues

Es van normalitzar les dades utilitzant "Probabilistic Quotient Normalization" per corregir les diferències degudes a l'efecte de les diferents dilucions dels metabòlits. A més, es van unir tots els resultats obtinguts de les mostres de control de qualitat (QC) i es va calcular la desviació estàndard relativa (RSD) de la concentració de cada metabòlit. Recordem que totes les mostres QC provenen del mateix individu i moment d'extracció, i, per tant, haurien de tenir molt poca variabilitat.

Inici de l'anàlisi realitzat en l'activitat: selecció i lectura del dataset

He descarregat aquestes dades crues del repository github:

<https://github.com/nutrimetabolomics/metaboData/>

He escollit fitxer anomenat "GastricCancer_NMR.xlsx" perquè el tema d'anàlisi em semblava interessant i alineat amb la línia d'estudi on m'agradaria desenvolupar-me professionalment. L'arxiu contenia dos fulles de càlcul: una d'elles amb l'array de dades de concentració dels metabòlits, conjuntament amb metadades relacionades amb les mostres, i l'altra fulla de càlcul contenia informació sobre els metabòlits.

com ara el nom, el percentatge de “missing values” o els valors RSD de les mostres de control de qualitat (QC).

He llegit aquestes dades amb la llibreria “readxl”.

Generació del contenidor SummarizedExperiment

Per treballar amb aquestes dades he utilitzat un contenidor de tipus “SummarizedExperiment”, i he utilitzat la llibreria “POMA”.

Però, per poder generar el contenidor “SummarizedExperiment”, primer he hagut de fer un preprocessat de les dades. Com he mencionat abans, tenim dos “dataframes”, un amb metadades dels metabòlits, i un amb l’array de dades dels metabòlits i les mostres conjuntament amb metadades de les mostres. Per tant, per poder crear el “SummarizedExperiment”, he fet el següent:

- Primer he separat les metadades de les mostres i les dades de l’array.
- Seguidament, he transposat la matriu de l’array per tenir els metabòlits en files i les mostres en columnes.
- Després he netejat les dades reorganitzant i eliminat algunes columnes que em semblava que tenien informació redundant.

Per tant, he obtingut 3 datasets amb els quals he generat el “SummarizedExperiment”:

1. Array de concentració de metabòlits i mostres → assay
2. Informació dels metabòlits → rowData
3. Informació de les mostres → colData

Control de qualitat, preprocessat

Un cop format el SummarizedExperiment he visualitzat amb “head”, les primers línies de les dades per comprovar si tot ha funcionat correctament. Després d’aquesta visualització he factoritzat les variables “Class” i “SampleType” de l’objecte “colData”.

A continuació, aprofitant que tenia la informació de l’article, he generat 2 boxplots per visualitzar la distribució dels “missing values” i dels RSD de la concentració dels metabòlits en de les mostres control de qualitat.

- Tal i com realitzen a l’article, he prosseguit quedant-me només amb els metabòlits amb un RSD en les mostres de control de qualitat menor a 25%.
- Llavors, he tractat els “missing values” amb la llibreria “POMA” i la funció “PomaImpute”, utilitant el mètode de “k-nearest neighbors” (k-NN) per

reemplaçar els “missing values” (NA) per valors calculats a través de mostres molt similars. A més, si hi hagués trobat metabòlits amb més d’un 20% de NAs, els hauria eliminat directament (“remove_na = TRUE, cutoff = 20”).

Normalització i control d’outliers

Tot seguit he normalitzat les dades amb “PomaNorm” i el mètode “log_pareto”, que fa una transformació logarítmica, per reduir l’efecte dels valors extrems, i un escalat, per normalitzar la variabilitat de cada metabòlit i facilitar la comparabilitat.

Finalment, he processat els outilers amb “PomaOutliers”.

Anàlisi descriptiva

En paral·lel, com a anàlisi univariant, tal i com mostra el “workflow” de Bioconductor del paquet “POMA”, he generat un boxplot, amb “PomaBoxplot”, i un density-plot, amb “PomaDensity”, abans i després de la normalització, per veure’n l’efecte. D’altra banda, també he fet una visualització dels outliers amb “PomaOutliers\$polygon_plot”.

En últim lloc, per l’anàlisi multivariant he realitzat un anàlisi de components principals (PCA) amb “PomaPCA” per avaluar si les dades s’agrupen segons les classes “BN”, “HE”, “GC” i “QC” amb el gràfic “\$factors_plot” i he comprovat la varianza que explica cada component principa amb el gràfic “\$eigenvalues_plot”.

Resultats

Primer de tot visualitzem les dades que hem llegit de les dues fulles de l’arxiu “GastricCancer_NMR.xlsx”.

Idx	SampleID	SampleType	Class	M1
1	sample_1	QC	QC	90.1
2	sample_2	Sample	GC	43.0
3	sample_3	Sample	BN	214.3
4	sample_4	Sample	HE	31.6
5	sample_5	Sample	GC	81.9

Idx	Name	Label	Perc_missing	QC_RSD
1	M1	1_3-Dimethylurate	11.4285714	32.208005
2	M2	1_6-Anhydro-β-D-glucose	0.7142857	31.178028
3	M3	1_7-Dimethylxanthine	5.0000000	34.990605
4	M4	1-Methylnicotinamide	8.5714286	12.804201

Idx	Name	Label	Perc_missing	QC_RSD
5	M5	2-Aminoadipate	1.4285714	9.372664
6	M6	2-Aminobutyrate	5.0000000	46.977149

Com podem observar, tenim les dades de l'array juntament amb dades informatives sobre les mostres. Després de realitzar un primer preprocesat i generar el SummarizedExperiment veiem com queden els tres objectes: assay, colData i rowData.

	SampleType	Class
sample_1	QC	QC
sample_2	Sample	GC
sample_3	Sample	BN
sample_4	Sample	HE
sample_5	Sample	GC
sample_6	Sample	BN

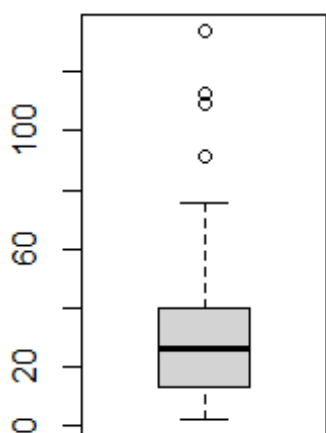
	Label	Perc_missing	QC_RSD
M1	1_3-Dimethylurate	11.4285714	32.208005
M2	1_6-Anhydro- β -D-glucose	0.7142857	31.178028
M3	1_7-Dimethylxanthine	5.0000000	34.990605
M4	1-Methylnicotinamide	8.5714286	12.804201
M5	2-Aminoadipate	1.4285714	9.372664
M6	2-Aminobutyrate	5.0000000	46.977149

	sample_1	sample_2	sample_3	sample_4	sample_5
M1	90.1	43.0	214.3	31.6	81.9
M2	491.6	525.7	10703.2	59.7	258.7
M3	202.9	130.2	104.7	86.4	315.1
M4	35.0	NA	46.8	14.0	8.7
M5	164.2	694.5	483.4	88.6	243.2

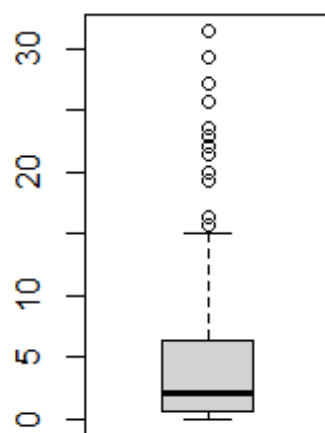
Com sabem gràcies a l'article, haurem de processar i filtrar els metabòlits amb "missing values" alts i valors alts d'RSD per la concentració dels metabòlits en les mostres de control de qualitat (QC).

Observem la distribució abans del processat.

QC %RSD

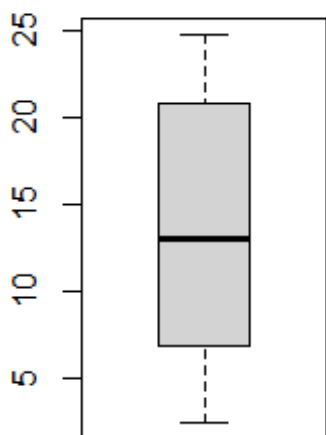


Missing values

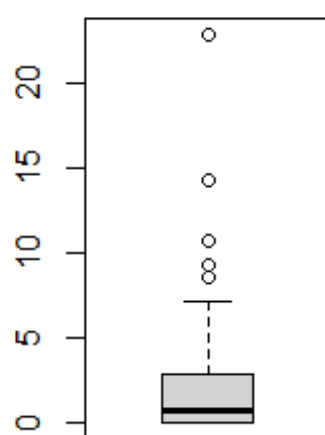


Observem la distribució després del processat.

QC %RSD

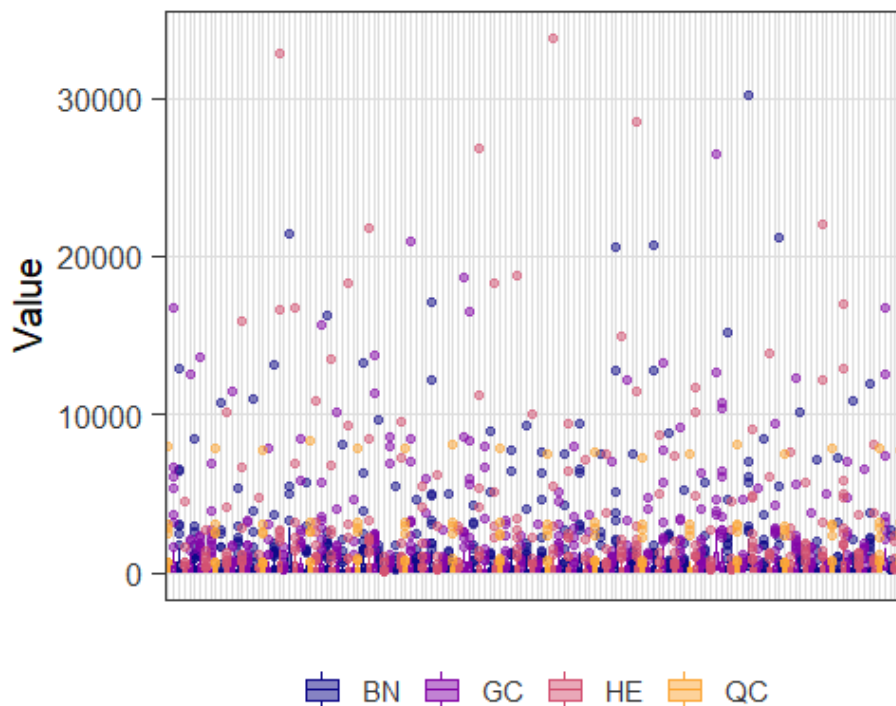


Missing values

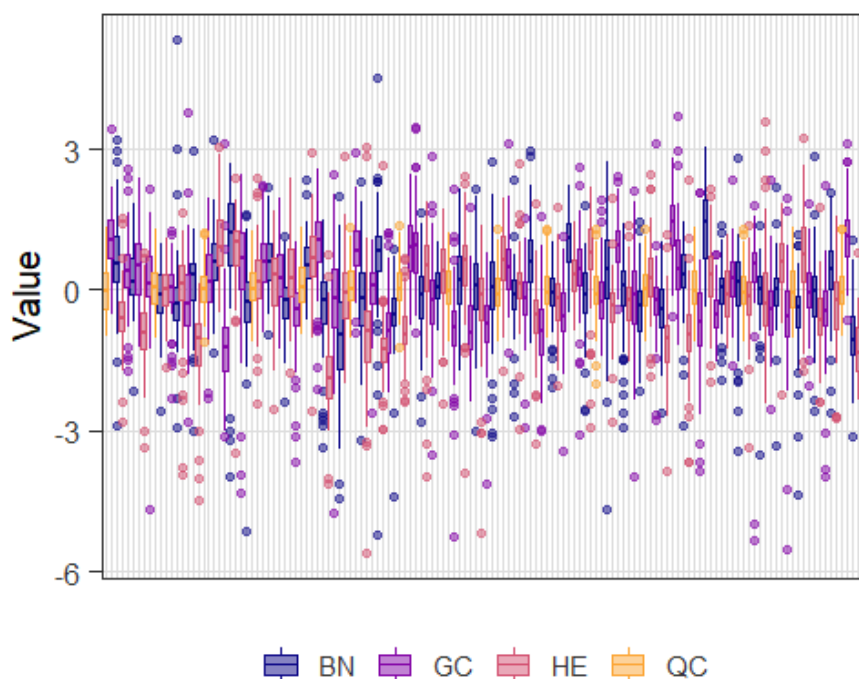


A continuació podem veure l'efecte de la normalització de les dades que ens permet reduir l'efecte dels valors extrems, i fer un escalat, per facilitar la comparabilitat. L'eix de les x són les mostres, i l'eix de les y la concentració de metabòlits.

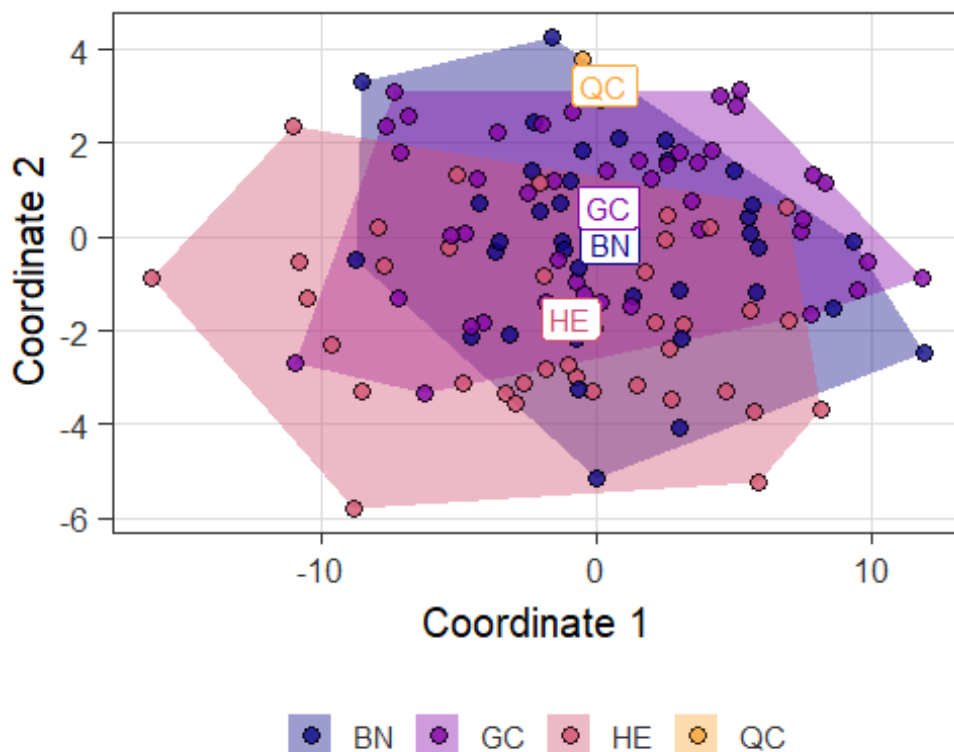
Dades abans de la normalització.



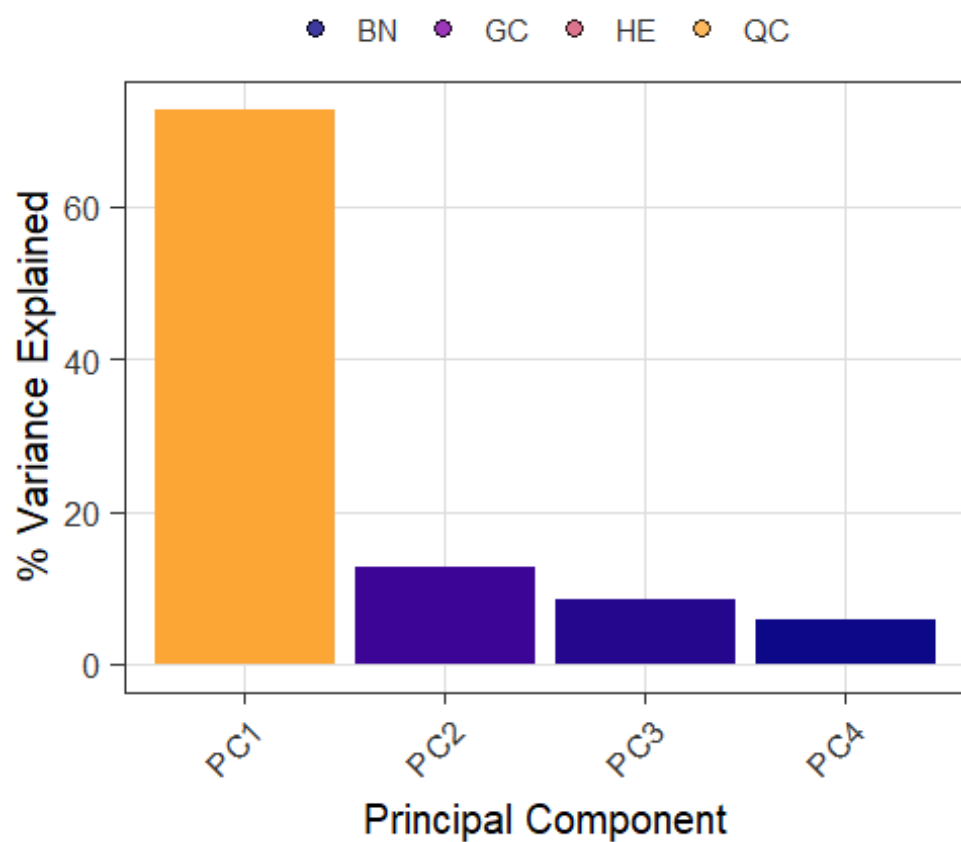
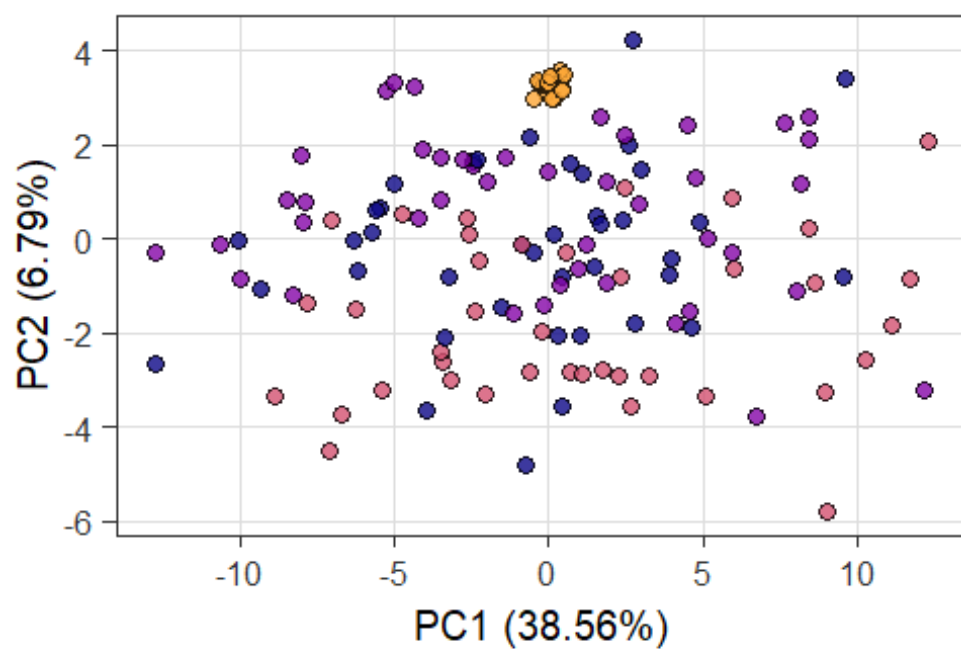
Dades després de la normalització



Amb el següent gràfic podem veure els outliers de les nostres dades. Com podem observar, les dades de les mostres de “QC” no presenten outliers. En canvi, en la resta de mostres sí que veiem outliers, destacant especialment en les mostres de pacients sans “HE”.



Finalment, amb l'anàlisi de components principals (PCA) s'observa que l'únic grup que es separa clarament és el de control de qualitat, cosa previsible ja que totes les mostres d'aquest grup provenen del mateix origen. Per a la resta de grups, les variàncies recollides en els components principals 1 i 2 no mostren una agrupació visible a simple vista. D'altra banda, es veu que el PC1 és el component que explica el percentatge més alt de variància en comparació amb els altres components.



Discussió

Després de realitzar aquest estudi, no podem dir que haguem identificat un perfil metabolòmic únic per als pacients amb càncer gàstric. Amb els resultats obtinguts a través de l'anàlisi de components principals (PCA) hem vist que la separació entre pacients amb càncer gàstric (GC), afeccions benignes (BN) i individus sans (HE) no és prou evident. Això indica que, tot i que el perfil metabolòmic pot mostrar algunes alteracions associades amb la malaltia, les diferències entre aquests grups no són suficients per permetre una separació clara basant-nos només en els dos primers components principals.

Per aprofundir en aquesta anàlisi i obtenir conclusions més sòlides, caldria complementar el PCA amb proves estadístiques com el T-test o l'ANOVA per identificar si hi ha diferències estadísticament significatives entre el grup de pacients amb càncer gàstric i els altres grups. Aquestes anàlisis podrien ajudar a determinar si certes concentracions de metabòlits són diferencials entre les condicions estudiades, cosa que ens acostaria a definir un conjunt inicial de biomarcadors útils per a un primer diagnòstic del càncer gàstric.

Amb aquests marcadors, es podria plantejar un primer test de cribratge per identificar possibles casos de càncer gàstric que posteriorment podrien confirmar-se amb proves addicionals per augmentar la precisió diagnòstica. Tot plegat, això contribuiria a establir una eina diagnòstica no invasiva, amb el potencial d'afavorir una detecció més precoç i accessible per a aquesta malaltia.

En relació amb l'anàlisi tècnica de les dades, aquest treball m'ha permès explorar en profunditat com tractar dades de microarrays utilitzant un contenidor SummarizedExperiment. Poder seguir el processament de dades detallat en l'article original ha estat molt valuós, ja que m'ha donat una visió més clara i pràctica del flux de treball habitual en aquest tipus d'estudis. A més, les instruccions disponibles al web del paquet "POMA" han estat de gran ajuda per simplificar l'anàlisi i comprendre millor cada pas.

En definitiva, aquesta activitat ha estat molt útil per establir una base sòlida en el flux de treball de l'anàlisi de dades òmiques, i en l'ús de la programació orientada a objectes aplicada a aquest camp.

Annex

Repositori Github: <https://github.com/joelmarmir/Martinez-Miralles-Joel-PEC1>

Codi utilitzat:

```
# Carrego Les dades i metadades del fitxer xlsx
dades_array_sample <- read_excel('GastricCancer_NMR.xlsx', sheet =
```

```

'Data')
dades_metabolites <- read_excel('GastricCancer_NMR.xlsx', sheet = 'Peak')

# Miro el que el nombre de columnes de Les dades (restant-li 4 columnes
informatives) i el nombre de files de Les metadades coincideixi

ncol(dades_array_sample)

## [1] 153

nrow(dades_metabolites)

## [1] 149

# Separo Les dades dels nivells de metabòlit (array) i els descriptors de
La mostra (metadades)
metadades <- subset(
  dades_array_sample, select = c(SampleType, Class)
)

rownames(metadades) <- dades_array_sample$SampleID

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

dades_assay <- subset(
  dades_array_sample, select = -c(Idx, SampleID, SampleType, Class)
)

# Afegeixo els noms de Les mostres a Les files del nou dataframe
rownames(dades_assay) <- dades_array_sample$SampleID

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

# Ara transposo creo La matriu de dades pel SummarizedExperiment,
transposant el dataframe amb Les dades de concentració dels metabòlits
assay_matriu <- t(as.matrix(dades_assay))

# Preparo La matriu de dades dels metabòlits
dades_metabolites_prep <- subset(
  dades_metabolites, select = -c(Idx, Name)
)

rownames(dades_metabolites_prep) <- dades_metabolites$Name

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

# Genero el SummarizedExperiment
se <- SummarizedExperiment(
  assays = list(counts = assay_matriu),
  rowData = dades_metabolites_prep,
  colData = metadades
)

```

```
# Previsualitzem les 3 matrius de dades que tenim: assay, colData, rowData.
```

```
# Comencem amb colData
```

```
head(colData(se))
```

```
## DataFrame with 6 rows and 2 columns
##           SampleType      Class
##           <character> <character>
## sample_1          QC          QC
## sample_2        Sample          GC
## sample_3        Sample          BN
## sample_4        Sample          HE
## sample_5        Sample          GC
## sample_6        Sample          BN
```

```
dim(colData(se))
```

```
## [1] 140  2
```

```
# Comprovem quin és el nombre de mostres dins de cada classe
```

```
table(colData(se)$Class)
```

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

```
# Factoritzem les dades
```

```
colData(se)$Class <- factor(colData(se)$Class)
```

```
colData(se)$SampleType <- factor(colData(se)$SampleType)
```

```
# Comprovem que s'hagi relitzat correctament
```

```
head(colData(se))
```

```
## DataFrame with 6 rows and 2 columns
##           SampleType      Class
##           <factor> <factor>
## sample_1          QC          QC
## sample_2        Sample          GC
## sample_3        Sample          BN
## sample_4        Sample          HE
## sample_5        Sample          GC
## sample_6        Sample          BN
```

```
assay(se)[1:5, 1:5]
```

```
##   sample_1 sample_2 sample_3 sample_4 sample_5
## M1    90.1    43.0   214.3    31.6    81.9
## M2   491.6   525.7  10703.2    59.7   258.7
## M3   202.9   130.2   104.7    86.4   315.1
## M4    35.0     NA    46.8    14.0    8.7
## M5   164.2   694.5   483.4    88.6   243.2
```

```
dim(assay(se))
```

```
## [1] 149 140
```

```
head(rowData(se))
```

```
## DataFrame with 6 rows and 3 columns
```

##		Label	Perc_missing	QC_RSD
##		<character>	<numeric>	<numeric>
## M1		1_3-Dimethylurate	11.428571	32.20800
## M2		1_6-Anhydro- β -D-gluc..	0.714286	31.17803
## M3		1_7-Dimethylxanthine	5.000000	34.99060
## M4		1-Methylnicotinamide	8.571429	12.80420
## M5		2-Aminoadipate	1.428571	9.37266
## M6		2-Aminobutyrate	5.000000	46.97715

```
dim(rowData(se))
```

```
## [1] 149 3
```

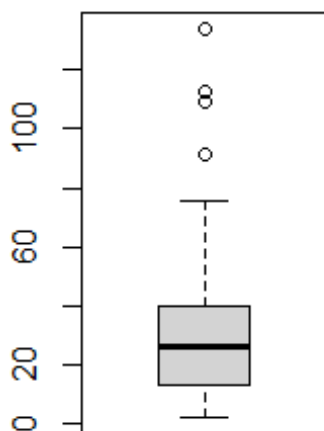
Recordem que en l'estudi, es fan servir mesures de control de qualitat per avaluar si els nivells de concentració dels metabòlits són reproduïbles o presenten una variabilitat intrínseca (QC_RSD). Serà interessant visualitzar aquestes dades, així com el percentatge de missing values per cada metabòlit (Perc_missing)

```
par(mfrow = c(1,2))
```

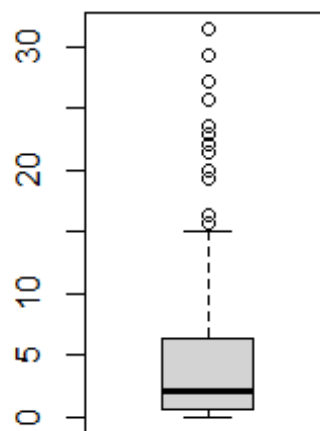
```
boxplot(rowData(se)$QC_RSD, main = "QC %RSD")
```

```
boxplot(rowData(se)$Perc_missing, main = "Missing values")
```

QC %RSD



Missing values



```
# Comencem el Pre-processat.
```

```
# Com podem observar, hi ha diversos metabòlits amb concentracions molt  
variabes en els QC (valors d'RSD alts). Per altra banda, no tenim missing  
values elevats. En l'anàlisi realitzat amb aquestes dades (Chan et al,  
2016), es van incloure només els metabòlits amb un QC_RSD < 25%. Per  
tant, començarem fent aquest filtratge. Pel que fa als missing values, no  
realitzarem cap acció en concret de moment.
```

```
nrow(assay(se))
```

```
## [1] 149
```

```
# Com veiem comecem amb 149 metabòlits
```

```
# Realitzem el filtratge.
```

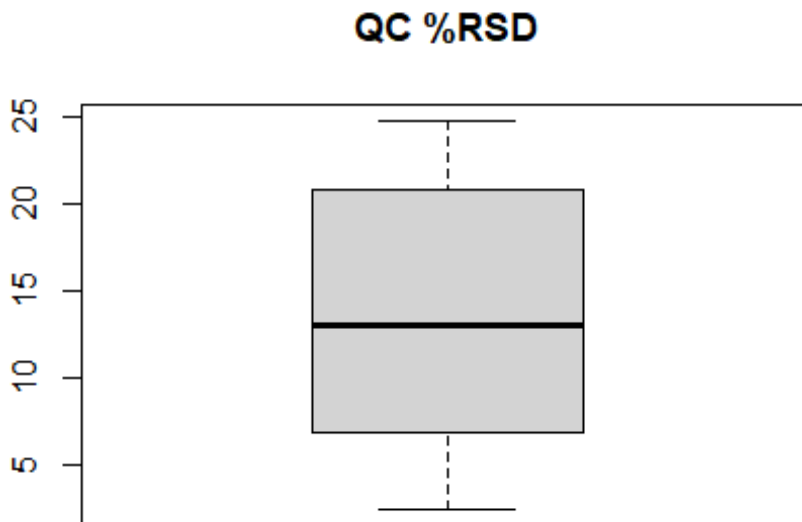
```
clean_se <- se[rowData(se)$QC_RSD<25, ]
```

```
nrow(assay(clean_se))
```

```
## [1] 73
```

```
# Ara ens quedem amb 73 metabòlits
```

```
boxplot(rowData(clean_se)$QC_RSD, main = "QC %RSD")
```



```

# Seguim amb el pre-processat amb "POMA". Utilitzem el mètode de k-
nearest neighbors (k-NN) per reemplaçar els missing values (NA) per
valors calculats a través de mostres molt similars. A més, ara sí que
eliminem les mostres que tenen més d'un 20% de NAs.
imputed <- clean_se %>% PomaImpute(method = "knn", zeros_as_na = TRUE,
remove_na = TRUE, cutoff = 20)

## 0 features removed.

# Visualitzem el resultat
imputed

## class: SummarizedExperiment
## dim: 73 140
## metadata(0):
## assays(1): ''
## rownames(73): M4 M5 ... M148 M149
## rowData names(0):
## colnames(140): sample_1 sample_2 ... sample_139 sample_140
## colData names(2): SampleType Class

assay(imputed)[1:5, 1:5]

##      sample_1 sample_2 sample_3 sample_4 sample_5
## M4          35.0  83.37778      46.8      14.0      8.7
## M5          164.2 694.50000     483.4      88.6     243.2
## M7           41.0  37.90000     110.1     170.3     349.4
## M8           46.5 125.70000      85.1      23.9      61.1
## M11          61.7 490.60000    2441.2     140.7      48.7

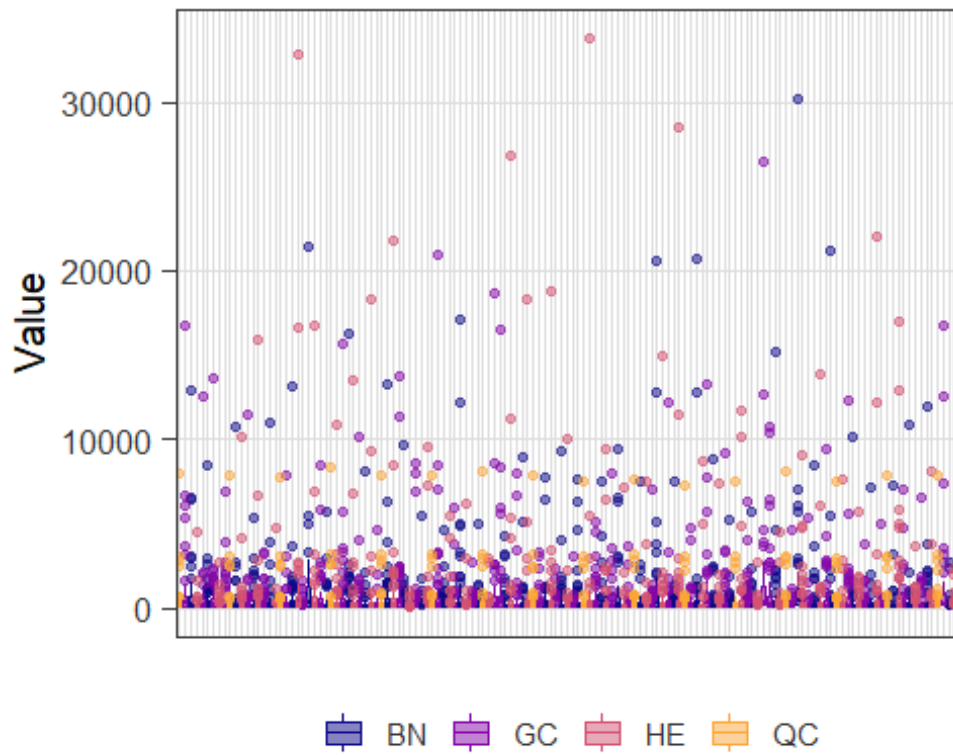
# Realitzem la normalització de les dades.
# "log_pareto" fa una transformació logarítmica, per reduir l'efecte dels
valors extrems, i un escalat, per normalitzar la variabilitat de cada
metabòlit i facilitar la comparabilitat.
normalized <- imputed %>%
  PomaNorm(method = "log_pareto")

normalized

## class: SummarizedExperiment
## dim: 73 140
## metadata(0):
## assays(1): ''
## rownames(73): M4 M5 ... M148 M149
## rowData names(0):
## colnames(140): sample_1 sample_2 ... sample_139 sample_140
## colData names(2): SampleType Class

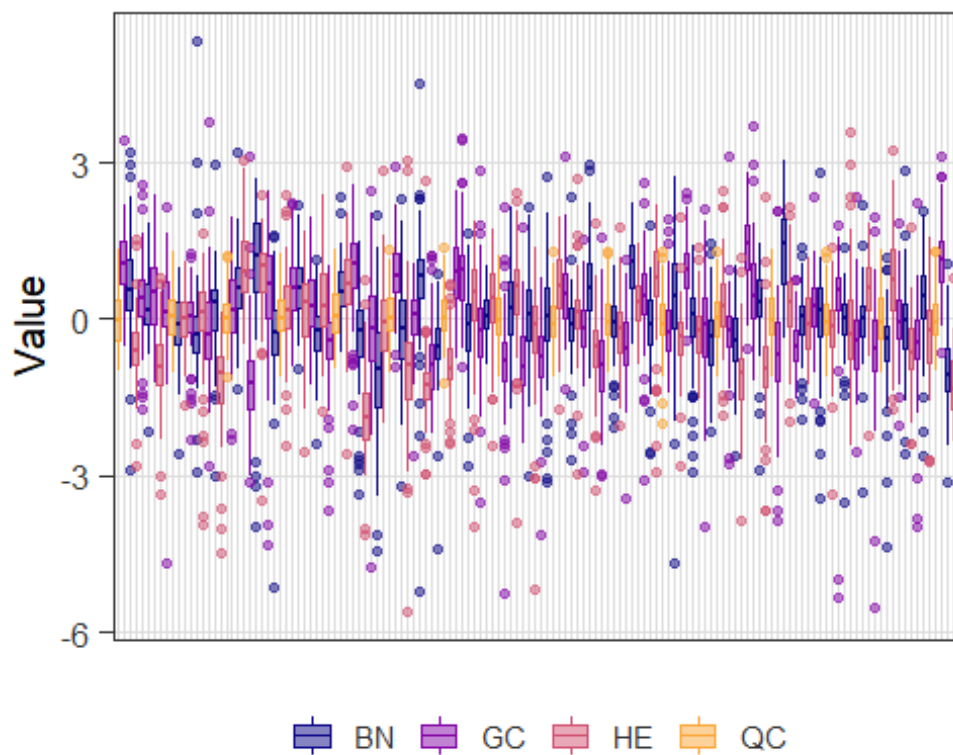
# Visualitzem l'efecte de la normalització. Primer representem les dades
abans de la normalització.
PomaBoxplots(imputed, x = "samples", outcome = "Class", theme_params =
list(axistext = "y"))

```



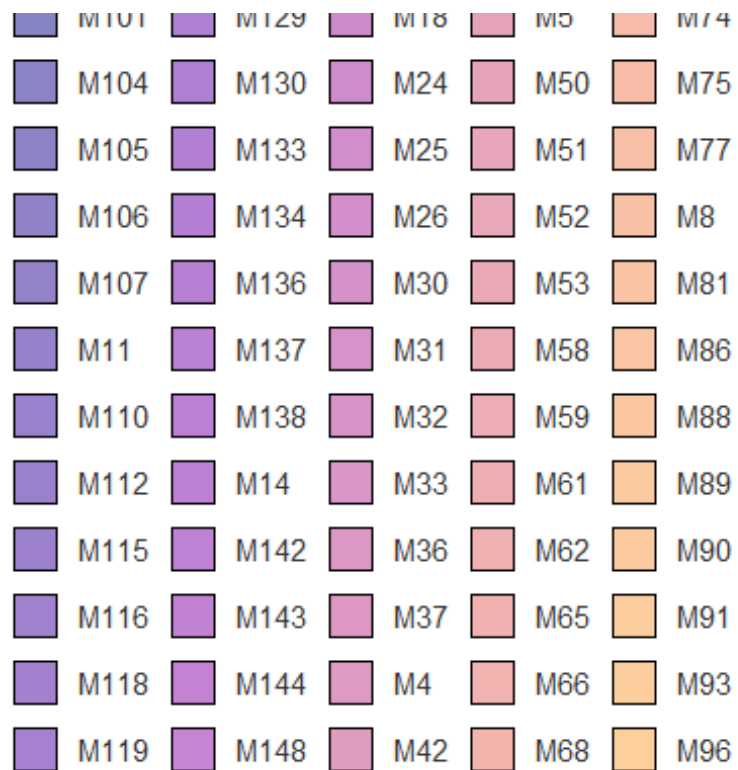
Ara després de la normalització

```
PomaBoxplots(normalized, x = "samples", outcome = "Class", theme_params =
list(axistext = "y"))
```




```
# Genero el gràfic
```

```
PomaDensity(imputed, x = "features")
```



```
# Deso el gràfic
```

```
png("poma_density_i.png", width = 800, height = 600)
```

```
PomaDensity(imputed, x = "features")
```





























































```
dev.off()
```

```
## png
```

```
## 2
```

```
# Genero el gràfic
```

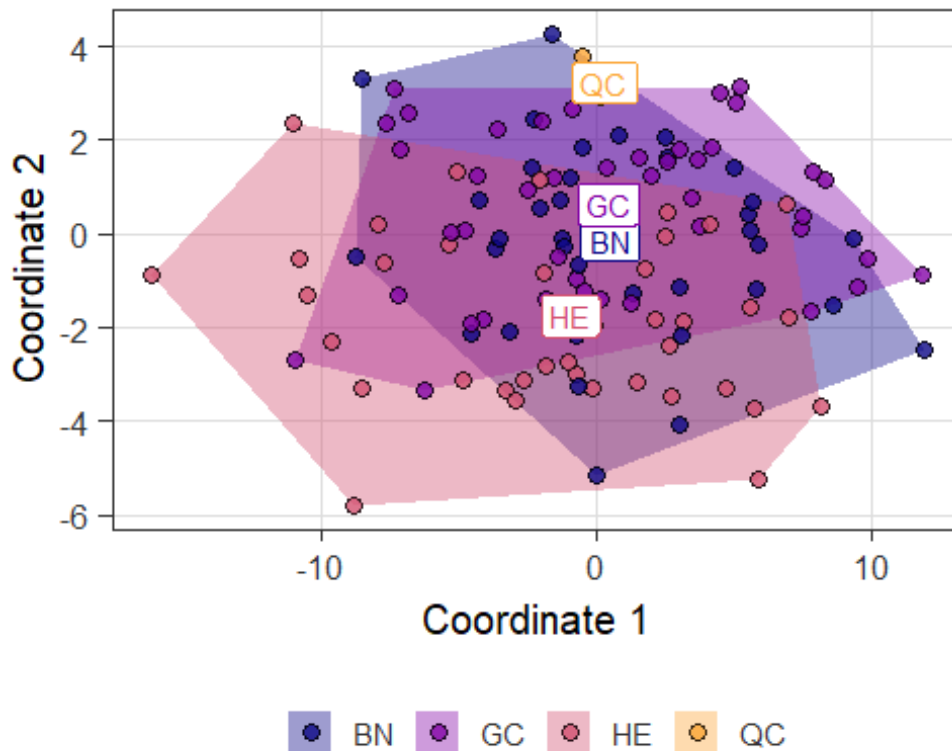
```
PomaDensity(normalized, x = "features")
```

	M101		M129		M18		M5		M74
	M104		M130		M24		M50		M75
	M105		M133		M25		M51		M77
	M106		M134		M26		M52		M8
	M107		M136		M30		M53		M81
	M11		M137		M31		M58		M86
	M110		M138		M32		M59		M88
	M112		M14		M33		M61		M89
	M115		M142		M36		M62		M90
	M116		M143		M37		M65		M91
	M118		M144		M4		M66		M93
	M119		M148		M42		M68		M96

```
# Deso el gràfic
png("poma_density_n.png", width = 800, height = 600)
PomaDensity(normalized, x = "features")
dev.off()

## png
## 2

PomaOutliers(normalized, outcome = 'Class')$polygon_plot
```



```
# Deso el gràfic
png("poma_outliers.png", width = 1000, height = 600)
PomaOutliers(normalized, outcome = 'Class')$polygon_plot
dev.off()

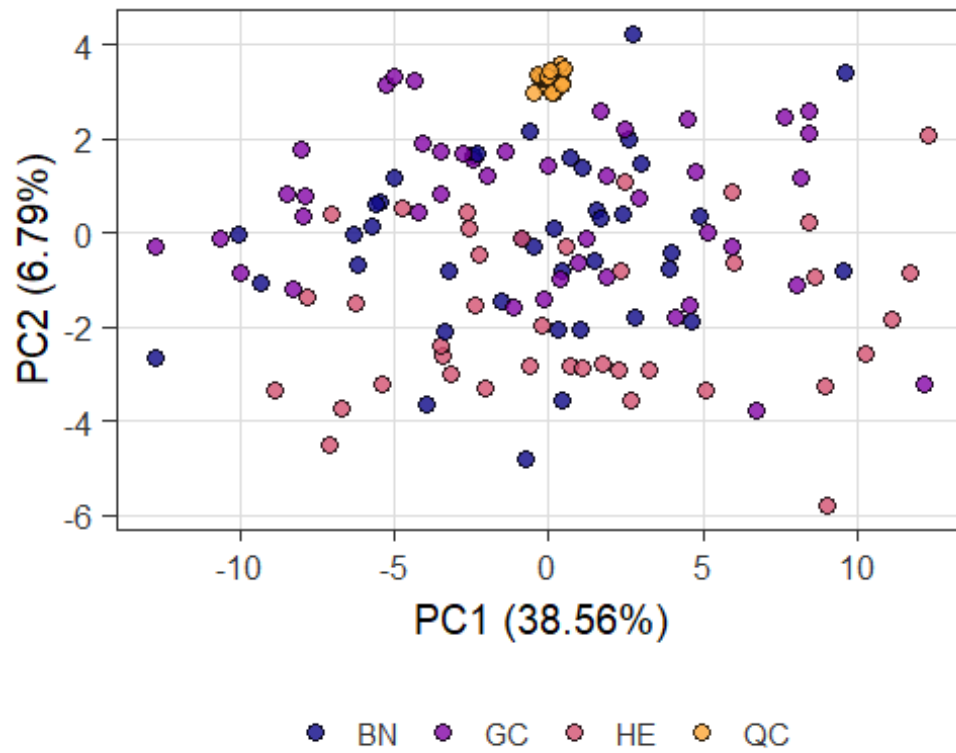
## png
## 2

pre_processed <- PomaOutliers(normalized)$data
pre_processed

## class: SummarizedExperiment
## dim: 73 138
## metadata(0):
## assays(1): ''
## rownames(73): M4 M5 ... M148 M149
## rowData names(0):
## colnames(138): sample_1 sample_2 ... sample_139 sample_140
## colData names(2): SampleType Class

# Fem un anàlisi de components principals

# Visualitzem la gràfica general
PomaPCA(pre_processed, outcome = "Class")$factors_plot
```



```
# Ara veiem el pes de cada component principal
PomaPCA(pre_processed, outcome = "Class")$eigenvalues_plot
```

