Anàlisis de dades òmiques – PEC1

Taula de continguts:

1. Abstract o “Resum Executiu”
2. Objectius de l’estudi
3. Materials i mètodes
4. Resultats
5. Discussió, limitacions i conclusions

1. Abstract o "Resum executiu":

Aquesta PEC és un exercici que permet repassar i ampliar els coneixements i habilitats adquirits treballant en les 3 activitats inicials. Ens centrem sobretot en la descàrrega, manipulació i anàlisis d’un conjunt de dades obtingudes, en aquest cas ‘Gastric_Cancer_NMR’, creant un contenidor *SummarizedExperiment* de Bioconductor. Seguidament fem una exploració descriptiva de les dades i creem un informe que podem compartir amb GitHub.

2. Objectius de l' estudi.

L’objectiu d’aquesta PEC és planificar i executar una versió simplificada de la anàlisis de dades òmiques, mentre que practiquem i treballem amb algunes eines i mètodes que hem treballat fins ara, com bioconductor, github, Expresion Set, etc.

L’objectiu serà descarregar un dataset (en el nostre cas ‘GastricCancer_NMR’), i un cop descarregades les dades haurem de crear un contenidor del tipus *SummarizedExperiment* que contingui les dades i metadades del dataset en qüestió.

A continuació durem a terme una exploració del dataset que ens doni una visió general de les dades.

Seguidament elaborarem l'informe que descrigui el procés mitjançant el qual hem realitzar l'estudi de les dades, incloent també la creació del repositori github.

Per últim crearem el repositori github, on hi inclourem:

- l' informe,
- l'objecte contenidor amb les dades i les metadades en format binari de R (arxiu amb extensió . Rda)
- el codi R per a l' exploració de les dades
- dades en format text
- les metadades sobre el dataset en un arxiu markdown.

3. Materials i Mètodes:

Treballarem amb les dades del dataset "GastricCancer_NMR.xlsx" que hem obtingut del repositori de github: https://github.com/nutrimetabolomics/metaboData/.

Utilitzem eines com R i Bioconductor, el contenidor *SummarizedExperiment,* i d'altres paquets utilitzats per la manipulació de dades. *SummarizedExperiment* permet, amb el contingut de les dades i metadades, emmagatzemar matrius rectangulars de diversos resultats experimentals, permetent gestionar simultàniament diversos resultats experimentals o *assays*, sempre i quan aquests tinguin les mateixes dimensions. Cada *assay* conté observacions d'una o més mostres, juntament amb metadates addicionals de les diferents observacions.

Seguidament creem un repositori github, que es una plataforma que permet emmagatzemar dades, compartir i treballar amb altres usuaris simultàniament les mateixes dades, així com fer un seguiment de tots els canvis que es van fent en el codi al llarg del temps.

4. Resultats:

A continuació s'exposen els passos i anàlisis que s'han anat seguint i els resultats obtinguts, així com l'explicació i interpretació d'aquests.

Primer s'instal·la *BiocManager* i *SummarizedExperiment*, amb el qual s'instal·len un conjunt de paquets predeterminats, i es carrega la llibreria per crear el contenidor:

```
> if (!require("BiocManager", quietly = TRUE))
+     install.packages("BiocManager")
> BiocManager::install(version = "3.18")
> BiocManager::install("SummarizedExperiment")
> library(SummarizedExperiment)
```

Carreguem el document "GastricCancer_NMR" amb la funció 'readxl', i en visualitzem l'estructura per sobre, amb funcions com 'colnames', 'head', 'summary', 'str', etc.

```
> library(readxl)
> GastricCancer_NMR <- read_excel("C:/Users/laura/OneDrive/Escriptori/GastricCancer_N
MR.xlsx")
> View(GastricCancer_NMR)
> colnames(GastricCancer_NMR)#visualitzem el nom de les columnes
  [1] "Idx"        "SampleID"   "SampleType" "Class"      "M1"
  [6] "M2"         "M3"         "M4"         "M5"         "M6"
 [11] "M7"         "M8"         "M9"         "M10"        "M11"
 [16] "M12"        "M13"        "M14"        "M15"        "M16"
 [21] "M17"        "M18"        "M19"        "M20"        "M21"

> head(GastricCancer_NMR) # mostrem les primeres files
# A tibble: 6 x 153
    Idx SampleID SampleType Class    M1     M2    M3    M4    M5    M6    M7    M8
  <dbl> <chr>    <chr>      <chr>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1     1 sample_1 QC         QC      90.1 4.92e2 203.    35   164.   19.7  41    46.5
2     2 sample_2 Sample     GC      43   5.26e2 130.    NA   694.  114.   37.9 126.
3     3 sample_3 Sample     BN     214.  1.07e4 105.    46.8 483.  152.  110.   85.1
4     4 sample_4 Sample     HE      31.6 5.97e1  86.4  14    88.6  10.3 170.   23.9
5     5 sample_5 Sample     GC      81.9 2.59e2 315.     8.7 243.   18.4 349.   61.1
6     6 sample_6 Sample     BN     197.  1.28e2 862.    18.7 200.    4.7  37.3 244.
# i 141 more variables: 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>,
#   M49 <dbl>, M50 <dbl>, M51 <dbl>, M52 <dbl>, M53 <dbl>, M54 <dbl>, …
# i Use `colnames()` to see all variable names

> summary(GastricCancer_NMR[, c("M1", "M2", "M3", "M4", "M5")]) #generem estadistiques
resum de les primeres columnes
       M1                M2                M3               M4
 Min.   :  0.40   Min.   :    3.1   Min.   :  0.1   Min.   :  0.10
 1st Qu.: 29.82   1st Qu.:  140.9   1st Qu.: 53.6   1st Qu.: 18.77
 Median : 60.35   Median :  270.2   Median :105.1   Median : 35.70
 Mean   :101.07   Mean   :  642.0   Mean   :146.4   Mean   : 43.83
 3rd Qu.:133.38   3rd Qu.:  480.9   3rd Qu.:198.8   3rd Qu.: 51.33
 Max.   :909.90   Max.   :26195.8   Max.   :862.5   Max.   :242.50
 NA's   :16       NA's   :1         NA's   :7       NA's   :12
       M5
 Min.   :   1.3
 1st Qu.:  67.0
 Median : 160.3
 Mean   : 231.1
 3rd Qu.: 253.1
 Max.   :2503.0
 NA's   :2
```

```
> str(GastricCancer_NMR) #classe de dades en cada columna
tibble [140 × 153] (S3: tbl_df/tbl/data.frame)
 $ Idx       : num [1:140] 1 2 3 4 5 6 7 8 9 10 ...
 $ SampleID  : chr [1:140] "sample_1" "sample_2" "sample_3"
"sample_4" ...
 $ SampleType: chr [1:140] "QC" "Sample" "Sample" "Sample"
...
 $ Class     : chr [1:140] "QC" "GC" "BN" "HE" ...
 $ M1        : num [1:140] 90.1 43 214.3 31.6 81.9 ...
 $ M2        : num [1:140] 491.6 525.7 10703.2 59.7 258.7 ..
 $ M3        : num [1:140] 202.9 130.2 104.7 86.4 315.1 ...
 $ M4        : num [1:140] 35 NA 46.8 14 8.7 18.7 NA 18.2 8.
36 ...
 $ M5        : num [1:140] 164.2 694.5 483.4 88.6 243.2 ...
 $ M6        : num [1:140] 19.7 114.5 152.3 10.3 18.4 ...
 $ M7        : num [1:140] 41 37.9 110.1 170.3 349.4 ...
```

Veiem que tenim 140 mostres (files), i 153 columnes: les quatre primeres columnes son “Idx”, “SampleID”, “SampleType” i “Class”, i la resta son de M1 a M149 (metabòlits)

Dividim les dades en *assays*, que son matrius de dades quantitatives, y *colData*, metadades de les mostres, tenint en compte que han de tenir les mateixes columnes per poder fer el *SummarizedExperiment*:

```
> assays <- t(as.matrix(GastricCancer_NMR[, 5:ncol(GastricCancer_NMR)]))#creem la matriu
transposada amb les columnes amb dades quantitatives; cada columna es una mostra
> dim(assays) #comprovem les dimensions de la matriu
[1] 149 140
```

Assignem nom a cada columna (mostra) per saber a quina mostra correspon (SampleID):

```
> colnames(assays) <- GastricCancer_NMR$SampleID #assignem un nom (sampleID) a cada c
olumna (monstra)
```

Creem el dataframe (*colData*) amb les dades associades a cada mostra, com el ID (SampleID), la classe (Class) i el tipus de mostra (SampleType), i en comprovem les dimensions:

```
> colData <- DataFrame(SampleID = GastricCancer_NMR$SampleID, SampleType = GastricCan
cer_NMR$SampleType, Class = GastricCancer_NMR$Class)
> dim(colData) #comprovem les dimensions
[1] 140   3
```

Veiem que tant *assays* com *colData* tenen les mateixes dimensions pel que fa al número de mostres (columnes en *assays* i files a *colData*), el que és una condició necessària per crear el contenidor *SummarizedExpermient*.

Creem el *SummarizesExpermient*, anomentat 'se':

```
> se <- SummarizedExperiment(assays = list(counts = assays), colData = colData)
> se #explorem el contingut de SummarizedExperiment
class: SummarizedExperiment
dim: 149 140
metadata(0):
assays(1): counts
rownames(149): M1 M2 ... M148 M149
rowData names(0):
colnames(140): sample_1 sample_2 ... sample_139 sample_140
colData names(3): SampleID SampleType Class
> summary(se) #veiem un resum de SummarizedExperiment
[1] "SummarizedExperiment object of length 149 with 0 metadata columns"
```

Veiem que les dimensions de la matriu principal son de 149 files (metabòlits, M1 – M149) i 140 columnes (mostres). De moment tenim 0 metadates.

Comprovem si les dades de *colData* s'han inclòs correctament:

```
> head(colData(se)) #mostra les primeres files de les metadates
DataFrame with 6 rows and 3 columns
           SampleID  SampleType       Class
         <character> <character> <character>
sample_1    sample_1          QC          QC
sample_2    sample_2      Sample          GC
sample_3    sample_3      Sample          BN
sample_4    sample_4      Sample          HE
sample_5    sample_5      Sample          GC
sample_6    sample_6      Sample          BN
```

Seguim fent una exploració del dataset:

```
> assays(se)$counts #visualitzem les dades quantitative de les mostres
     sample_1 sample_2 sample_3 sample_4 sample_5 sample_6 sample_7 sample_8
M1       90.1     43.0    214.3     31.6     81.9    196.9     45.5     91.0
M2      491.6    525.7  10703.2     59.7    258.7    128.2    190.4    231.9
M3      202.9    130.2    104.7     86.4    315.1    862.5     32.0    212.5
M4       35.0       NA     46.8     14.0      8.7     18.7       NA     18.2
M5      164.2    694.5    483.4     88.6    243.2    200.1    362.7     72.5
M6       19.7    114.5    152.3     10.3     18.4      4.7     35.7      6.7
M7       41.0     37.9    110.1    170.3    349.4     37.3     59.6     15.3
     sample_9 sample_10 sample_11 sample_12 sample_13 sample_14 sample_15
M1      480.6      62.2      36.5      93.5        NA      52.1        NA
M2      470.3     181.5     190.1     525.3     205.9      94.7     250.3
M3       60.7      75.5     153.1     215.8      19.6      68.2      59.1
M4        8.4      36.0      47.4      45.0      40.9      26.0      70.6
M5      270.2     203.4     146.5      62.6     106.9      95.5      65.4
M6       57.4      18.7      26.9      12.8      25.0       9.0      20.5
M7      213.8      44.4      20.6      42.4       9.5      10.4      26.2
     sample_16 sample_17 sample_18 sample_19 sample_20 sample_21 sample_22
```

```
> colData(se) #mostrem les metadates
DataFrame with 140 rows and 3 columns
             SampleID  SampleType       Class
           <character> <character> <character>
sample_1       sample_1          QC          QC
sample_2       sample_2      Sample          GC
sample_3       sample_3      Sample          BN
sample_4       sample_4      Sample          HE
sample_5       sample_5      Sample          GC
...                ...         ...         ...
sample_136   sample_136          QC          QC
sample_137   sample_137      Sample          GC
sample_138   sample_138      Sample          BN
sample_139   sample_139      Sample          HE
sample_140   sample_140          QC          QC
```

També podem fer subconjunts bidimensionals del *SummarizedExperiment*, per exemple amb les 5 primeres files i 3 primeres columnes:

```
> se [1:5, 1:3] #fem subconjunts bidimensionals p.e. 5 primeres files i 3 primeres co
lumnes
class: SummarizedExperiment
dim: 5 3
metadata(0):
assays(1): counts
rownames(5): M1 M2 M3 M4 M5
rowData names(0):
colnames(3): sample_1 sample_2 sample_3
colData names(3): SampleID SampleType Class
```

El document GastricCancer, a part del full "data" conte el full "peak", que conté les metadades que introduirem a *rowdata*, que fins ara està buit.

```
> peak_data <- read_excel("C:/Users/laura/OneDrive/Escriptori/GastricCancer_NMR.xls
x", sheet = "Peak")
> head(peak_data) #veiem les primeres files de peak_data
# A tibble: 6 × 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-β-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
> dim(peak_data) #comprovem les dimensions de peak_data
[1] 149   5
```

Veiem que les files de "peak" son iguals que les de 'se' (149), la qual cosa ens permet incorporar els valors de metadades de peak_data com a *rowData* en 'se'.

```
> rownames(se) #comprovem els noms de les files
  [1] "M1"   "M2"   "M3"   "M4"   "M5"   "M6"   "M7"   "M8"   "M9"   "M10"  "M11"
 [12] "M12"  "M13"  "M14"  "M15"  "M16"  "M17"  "M18"  "M19"  "M20"  "M21"  "M22"
 [23] "M23"  "M24"  "M25"  "M26"  "M27"  "M28"  "M29"  "M30"  "M31"  "M32"  "M33"
 [34] "M34"  "M35"  "M36"  "M37"  "M38"  "M39"  "M40"  "M41"  "M42"  "M43"  "M44"
 [45] "M45"  "M46"  "M47"  "M48"  "M49"  "M50"  "M51"  "M52"  "M53"  "M54"  "M55"
 [56] "M56"  "M57"  "M58"  "M59"  "M60"  "M61"  "M62"  "M63"  "M64"  "M65"  "M66"
 [67] "M67"  "M68"  "M69"  "M70"  "M71"  "M72"  "M73"  "M74"  "M75"  "M76"  "M77"
 [78] "M78"  "M79"  "M80"  "M81"  "M82"  "M83"  "M84"  "M85"  "M86"  "M87"  "M88"
 [89] "M89"  "M90"  "M91"  "M92"  "M93"  "M94"  "M95"  "M96"  "M97"  "M98"  "M99"
[100] "M100" "M101" "M102" "M103" "M104" "M105" "M106" "M107" "M108" "M109" "M110"
[111] "M111" "M112" "M113" "M114" "M115" "M116" "M117" "M118" "M119" "M120" "M121"
[122] "M122" "M123" "M124" "M125" "M126" "M127" "M128" "M129" "M130" "M131" "M132"
[133] "M133" "M134" "M135" "M136" "M137" "M138" "M139" "M140" "M141" "M142" "M143"
[144] "M144" "M145" "M146" "M147" "M148" "M149"

> rowData(se) <- as(peak_data, "DataFrame") #introduim rowData com a dataframe a Summ
arizedExperiment
> rowData(se) #comprovem que les metadates s'han afegit a rowdata
DataFrame with 149 rows and 5 columns
           Idx        Name                  Label Perc_missing    QC_RSD
     <numeric> <character>            <character>    <numeric> <numeric>
M1           1          M1      1_3-Dimethylurate    11.428571  32.20800
M2           2          M2 1_6-Anhydro-β-D-gluc..     0.714286  31.17803
M3           3          M3   1_7-Dimethylxanthine     5.000000  34.99060
M4           4          M4   1-Methylnicotinamide     8.571429  12.80420
M5           5          M5         2-Aminoadipate     1.428571   9.37266
...        ...         ...                    ...          ...       ...
M145       145        M145                  uarm1     23.57143   41.4070
M146       146        M146                  uarm2      4.28571   34.4582
M147       147        M147              β-Alanine      1.42857   27.6235
M148       148        M148      π-Methylhistidine      1.42857   16.5619
M149       149        M149      τ-Methylhistidine      0.00000    8.3518
> head(rowData(se)) #mostrem les primeres files de rowdata
DataFrame with 6 rows and 5 columns
          Idx        Name                  Label Perc_missing    QC_RSD
    <numeric> <character>            <character>    <numeric> <numeric>
M1          1          M1      1_3-Dimethylurate    11.428571  32.20800
M2          2          M2 1_6-Anhydro-β-D-gluc..     0.714286  31.17803
M3          3          M3   1_7-Dimethylxanthine     5.000000  34.99060
M4          4          M4   1-Methylnicotinamide     8.571429  12.80420
M5          5          M5         2-Aminoadipate     1.428571   9.37266
M6          6          M6        2-Aminobutyrate     5.000000  46.97715
```

'*Idx*' indica l'índex numèric de cada M, '*Name*' correspon el nom codificant de cadascun dels metabòlits (M1, M2, M3...), '*Label*' correspon al nom descriptiu dels metabòlits, '*Perc_missing*' el percentatge de valors perduts per a cada metabòlit, i '*QC_RSD*' és el percentatge de desviació estàndard relativa per a la qualitat del control.

```
> summary(se) #comprovem que hem afegit metadates a SummarizedExperiment
[1] "SummarizedExperiment object of length 149 with 5 metadata columns"
> se #comprovem al integració de rowData
class: SummarizedExperiment
dim: 149 140
metadata(0):
assays(1): counts
rownames(149): M1 M2 ... M148 M149
rowData names(5): Idx Name Label Perc_missing QC_RSD
colnames(140): sample_1 sample_2 ... sample_139 sample_140
colData names(3): SampleID SampleType Class
```

Podem comprovar que ara tenim les metadades correspondents a "*peak*" a *rowData*.

Un cop creat el contenidor *SummarizedExperiment* i feta l'exploració del dataset, creem el contenidor github:

Creem el repository Git a R:

```
> library(usethis)
```

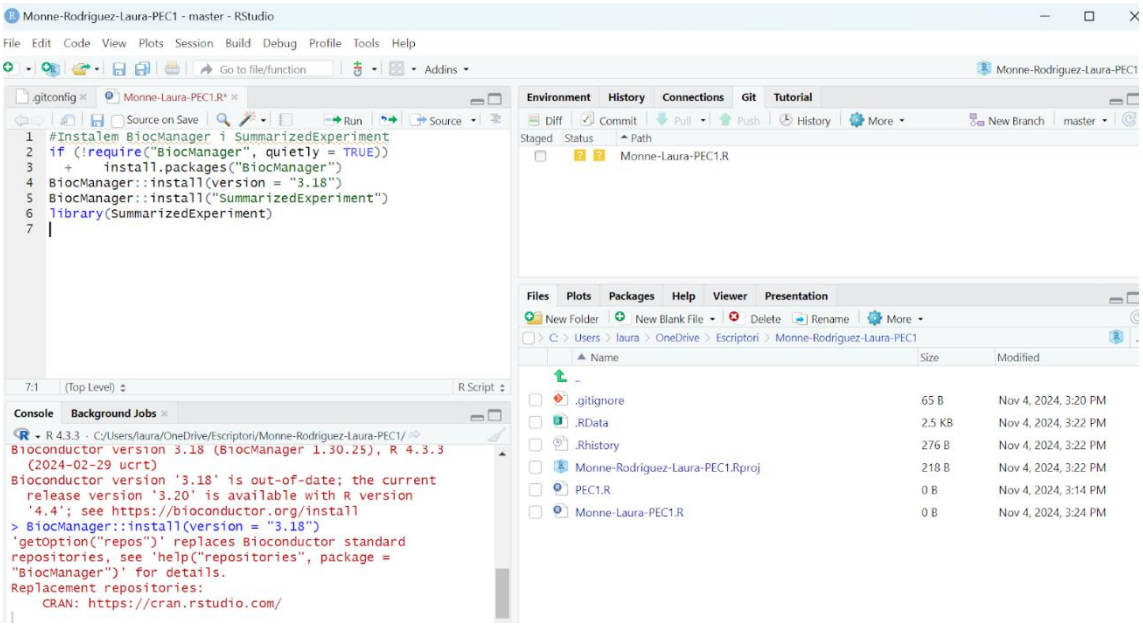
```
> usethis::edit_git_config()
```

I modifiquem les nostres dades (nom i e-mail).

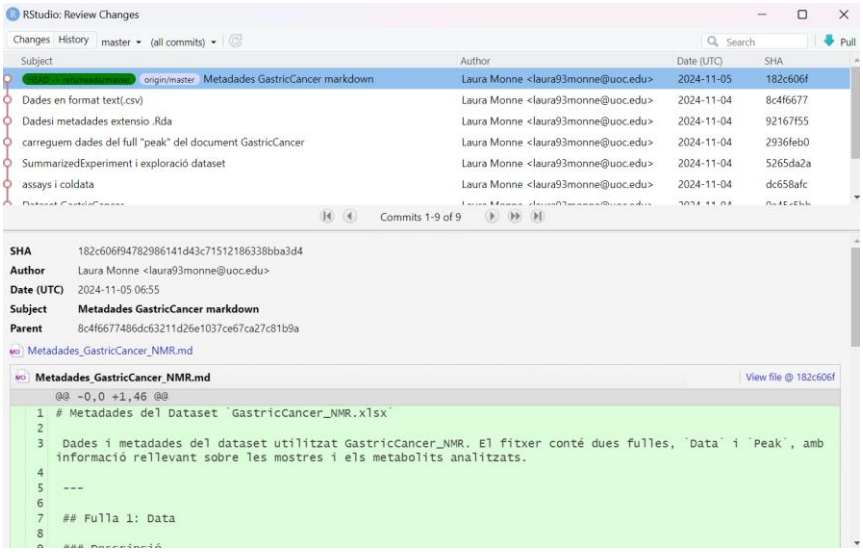
Llavors anem a `File -> New Project -> New Directory -> New Project -> Directory Name "Monne-Rodriguez-Laura-PEC1".`

Ara el nostre R studio té un repositori Git associat que podem veure a dalt a la dreta.

Un cop creat el nou projecte, creem un R script que es diu Monne-Laura-Pec1 i allí anem introduint el codi.



Cada vegada que introduïm codi nou cliquem a "commit". Si cliquem a "history" podem anar veient els canvis en el codi que anem fent al llarg del temps.



Seguidament ens hem de registrar a GitHub per poder connectar GitHub amb Rstudio. Un cop registrats posem a la consola

```
> usethis::create_github_token()
```

I se'ns obre un ellaç de GitHub on tindrem el nostre nom d'usuari i el token d'accés personal (PAT)

A R studio posem:

```
> library(gitcreds)
```

```
> gitcreds::git_set()
```

I ens demanarà una contrasenya que serà el token anteriorment creat. Ja tenim R Studio i GitHub connectats.

Link del repositori GitHub creat:

https://github.com/LauraMonne/Monne-Rodriguez-Laura-PEC1.git

5. Discussió i limitacions i conclusions de l' estudi.

Amb aquest estudi hem pogut construir i explorar un *SummarizedExperiment* amb les dades d'un repositori GtHub, en aquest cas "Gastric_Cancer_NMR", utilitzant Bioconductor en R. Amb la creació d'un nou repositori GitHub em pogut organitzar i controlar els diferents versions dels arxius, podent així compartir el treball que hem anat realitzant.

La limitació, o millor dit, proposta que podria haver-se fet és un estudi mes ampli i profund que es podria arribar a fer de les dades estudiades, com anàlisis clustering o altres tipus d'anàlisis que aportessin més informació sobre les dades estudiades.

Com a conclusió, l'exercici ens ha permès obtenir més experiència pràctica en l'ús de *SummarizedExperiment* per a l'estudi de dades metabolòmiques i en la gestió d'un projecte bioinformàtic en un repositori GitHub.