Treball Final de Màster: Cerca dels gens involucrats en el grau histològic de tumors de mama

Autor: Ernest Albets Moreno

02 de diciembre de 2024

Contents

1 Introducció

En aquest treball s'apliquen un conjunt de **tècniques de regressió de machine-learning de tipus supervisat** diferents sobre un conjunt de dades de pacients de tumors de mama on figuren com a característiques:

- Conjunt d'**expressions genètiques** dels **tumors de mama** de cada una de les pacients extretes amb la tècnica de microarrays (variables independents de tipus quantitatives contínues)
- **Grau histològic** del tumor (variable objectiu, de tipus categòric ordinal de 3 classes)

Es tracta, per tant, d'un conjunt de dades supervisat.

Les tècniques d'aprenentatge automàtic que s'apliquen per a obtenir les importàncies (ja siguin positives, com a factors de risc, o negatives, com a protectors) dels conjunts de gens en cada un dels diferents graus histològics són les següents:

- TÈCNIQUES BASADES EN ABRES DE DECISIÓ I/O DESCENS DE GRADIENT:
  - **Random Forest**
  - **XGBoost** (arbres de decisió i descens de gradient)
  - **LightGBM** (arbres de decisió i descens de gradient, més veloç que XGBoost)
- TÈCNIQUES BASADES EN REGULARITZACIÓ:
  - **Ridge Regression o L2**
  - **Lasso Regression o L1**
  - **Elastic Net** (combinació de les dues anteriors L1 i L2)

2 Objectiu

Detectar quins són els gens o conjunts de gens candidats a tenir **més rellevància** en relació a cada un dels diferents graus histològics dels tumors de mama. En alguns casos la importància dels gens pot ser de caràcter positiu i actuar com a elements protectors, o pel contrari, de caràcter negatiu i actuar com a factors de risc.

Establir rànkings i comparatives dels resultats en cada una de les tècniques de machine-learning supervisat aplicades així com gràfiques que n'il · lustrin els resultats.

3 0. Càrrega de funcions genèriques

En primer lloc, es carreguen una sèrie de funcions de caràcter general que seran utilitzades i accedides per a diversos processos de les diferents seccions del present treball i que eviten la redundància de codi i el contingut excessiu.

```
## FUNCIONS GENÈRIQUES

## Funció de càrrega, processament i unió amb les dades gèniques de la plataforma utilitzada per extrac
# 1. Carrega i llegeix el contingut de la taula de la ruta especificada
# 2. Captura i reanomena la variable que ens interessa, l'odentificador i el símbol del gen: GENE_SYMBO
# 3. Uneix la matriu d'expressions gèniques passada per paràmetre amb el seus GENE_SYMBOL que figuren e
process_platform_table <- function(origin_path, platform_name, gene_expression_matrix) {

  # filePath
  filePath <- paste0(origin_path, platform_name)
  # Read lines
  paltform_file <- readLines(filePath)
  paltform_file <- read.table(text = paltform_file, header = TRUE, sep = "\t", quote = "\"")

  # Selecció només de la informació de l'ID i el Gene.Symbol
  gene_info_selected <- paltform_file[, c("ID", "Gene.Symbol")]
  # Substituir '///' per ', ' en la columna Gene.Symbol, pels que tenen més d'un símbol
  gene_info_selected$Gene.Symbol <- gsub("///", ", ", gene_info_selected$Gene.Symbol)
  # Reanomenem GEN_SYMBOL
  colnames(gene_info_selected)[colnames(gene_info_selected) == "Gene.Symbol"] <- "GENE_SYMBOL"

  # Unió de les dues taules emprant la columna ID_REF de matriu_expressio_genetica_4 y ID de gene_info_s
  matriu_genetica_final <- merge(gene_expression_matrix, gene_info_selected[, c("ID", "GENE_SYMBOL")],
                          by.x = "ID_REF", by.y = "ID", all.x = TRUE)
```

```r
  # Reorganitzar les columnes per a que ID i GENE_SYMBOL apareixin primer
  matriu_genetica_final <- matriu_genetica_final[, c("ID_REF", "GENE_SYMBOL", setdiff(names(matriu_genet

  return(matriu_genetica_final)

}

## Funció de processament i neteja de la matriu final d'expressió genètica unida a les dades de la plata
# 1. Comprova si existeixen registres repetits que coincideixen amb el mateix GENE_SYMBOL. En cas afirma
#  - S'observa si existeixen files l'identificador de les quals no es troba present entre els existents
#     * Els eliminem
#  - S'agrupen totes les files per GENE_SYMBOL i a les repetides s'assigna el valor de la mitja aritmèti

process_merged_gene_expression <- function(gene_expression_matrix) {

  # Comprovació de repeticions en la columna de GENE_SYMBOL per tal d'aplicar la mitja en cas afirmatiu
  repeated_genes <- gene_expression_matrix %>%
    group_by(GENE_SYMBOL) %>%
    summarise(count = n()) %>%
    filter(count > 1)

  if (nrow(repeated_genes) > 0) {
    cat("S'han trobat símbols de gens repetits en la columna GENE_SYMBOL, procedim a realitzar les mitja

    # Comprovem si existeixen registres sense relació amb l'identificador de la plataforma
    missing_genes <- gene_expression_matrix %>%
      filter(is.na(GENE_SYMBOL) | GENE_SYMBOL == "")

    if (nrow(missing_genes) > 0) {
      cat("Tenim", nrow(missing_genes)," registres NO associats a cap identificador de la plataforma que

    # Eliminació registres no associats a cap identificador de la plataforma
    gene_expression_matrix <- gene_expression_matrix %>%
      filter(!is.na(GENE_SYMBOL) & GENE_SYMBOL != "")

      cat("Els registres sense identificador de plataforma s'han eliminat correctament.\n")

  }  else {
      cat("Tots els registres estan associats amb un identificador.\n")
  }

    # Realitzem l'agrupació per columna GENE_SYMBOL i la mitjana aritmètica dels valors que coincideixen
    # - Desapareixen els registres amb gene_symbol repetit quedant-ne només un com amb els valors mitjos
    gene_expression_matrix <- gene_expression_matrix %>%
      group_by(GENE_SYMBOL) %>%       # Agrupem per GENE_SYMBOL
      summarise(across(starts_with("GSM"), ~ round(mean(.x, na.rm = TRUE), 3))) %>%  # Calcula la mitja
      ungroup()

  } else {
    cat("No s'han trobat símbols de gens repetits en la columna GENE_SYMBOL.\n")
  }

  return(gene_expression_matrix)
```

```
}

## Funció de bolcatge dels resultats en un fitxer .csv a una ruta destí
write_to_csv <- function(gene_expression_matrix_processed, destination_path, filename) {

  # Bolquem els resultats en un fitxer .csv
  ruta_destino <- paste0(destination_path,filename)
  write.csv(gene_expression_matrix_processed, ruta_destino, row.names = FALSE)

}


## Funció de normalització
# Realitza la normalització de tots els valors continguts en una matriu de tipus llista @matrix passada
# La normalització posa tots els valors en un interval entre 0 i 1 tenint en compte el màxim i el mínim
# Funció per normalitzar una matriu entre 0 i 1
normalize_matrix <- function(matrix_list) {
  # Comprovar que la matriu d'entrada és una llista
  if (!is.list(matrix_list)) {
    stop("L'entrada ha de ser una llista.")
  }

  # Aplanar la llista en un sol vector per calcular el mínim i el màxim
  vector_values <- unlist(matrix_list)
  min_val <- min(vector_values, na.rm = TRUE)
  max_val <- max(vector_values, na.rm = TRUE)

  # Funció de normalització entre 0 i 1
  normalize_value <- function(x) {
    (x - min_val) / (max_val - min_val)
  }


  # Aplicar la normalització a cada element de la llista mantenint la seva estructura
  matrix_norm <- lapply(matrix_list, function(sublist) {
    # Normalitzar cada subllista o vector individualment
    sapply(sublist, normalize_value)
  })

  # Retornar la matriu normalitzada
  return(matrix_norm)
}


# Funció genèrica que crea un diagrama de barres horitzonals a partir de la llibreria ggplot2 per visual
# el rànking de resultats de les importàncies dels gens en cada una de les tècniques de machine-learning
# aplicades i per a un determinat grau histològic específic
# Paràmetres:
# - data: dades que contenen les importàncies de cada característica (gens)
# - X: columna de dades en eix X
# - y: columna de dades en eix Y
# - color: color de les barres horitzontals
# - title: títol de la gràfica
```

```r
# - top_num: nombre d'elements (gens) a visualitzar en les y's
# - xlablel: etiqueta en l'eix 'X'
# - ylabel: etiqueta en l'eix 'Y'
create_horitzontal_barchart_plot <- function(data, X, Y, color, title, top_num, xlabel, ylabel) {

  plot <- ggplot(data[1:top_num,], aes(x = reorder(.data[[X]], .data[[Y]]), y = .data[[Y]])) +
    geom_bar(stat = "identity", fill = color) +
    coord_flip() +
    labs(title = title, x = xlabel, y = ylabel) +
    theme_minimal() +
    theme(
      panel.background = element_rect(fill = "white", color = "white"), # Fons del gràfic
      plot.background = element_rect(fill = "white", color = "white"),  # Fons del quadre del gràfic
      panel.grid.major = element_line(color = "grey80"),                # Línies de la quadrícula
      panel.grid.minor = element_blank()                                # Línies de quadrícula menors
    )

  return(plot)
}


# Funció genèrica que crea un diagrama de punts a partir de la llibreria ggplot2 per visualitzar
# el rànking de resultats de les importàncies dels gens en cada una de les tècniques de machine-learning
# aplicades i per a un determinat grau histològic específic
# Paràmetres:
# - data: dades que contenen les importàncies de cada característica (gens)
# - X: columna de dades en eix X
# - y: columna de dades en eix Y
# - color: color de les barres horitzontals
# - title: títol de la gràfica
# - top_num: nombre d'elements (gens) a visualitzar en les y's
# - xlablel: etiqueta en l'eix 'X'
# - ylabel: etiqueta en l'eix 'Y'
create_point_chart_plot <- function(data, X, Y, color, title, top_num, xlabel, ylabel) {

  plot <- ggplot(data[1:top_num,], aes(x = .data[[Y]], y = reorder(.data[[X]], .data[[Y]]))) +
    geom_point(size = 3, color = color) +
    labs(title = title, x = xlabel, y = ylabel)
  theme_minimal() +
    theme(
      panel.background = element_rect(fill = "white", color = "white"), # Fons del gràfic
      plot.background = element_rect(fill = "white", color = "white"),  # Fons del quadre del gràfic
      panel.grid.major = element_line(color = "grey80"),                # Línies de la quadrícula
      panel.grid.minor = element_blank()                                # Línies de quadrícula menors
    )

  return(plot)
}


# Funció genèrica que crea un diagrama de barres horitzonals amb valors positius i/o negatius a partir d
# per visualitzar el rànking de resultats de les importàncies dels gens en cada una de les tècniques de
# basades en la regularització i a partir d'un grau histològic específic
```

```r
# Paràmetres:
# - data: dades que contenen les importàncies de cada característica (gens)
# - X: columna de dades en eix X
# - y: columna de dades en eix Y
# - z: distinció del signe
# - color1: color de les barres horitzontals amb valors positius
# - color2: color de les barres horitzontals amb valors negatius
# - title: títol de la gràfica
# - subtitle: subtitol
# - top_num: nombre d'elements (gens) a visualitzar en les y's
# - xlablel: etiqueta en l'eix 'X'
# - ylabel: etiqueta en l'eix 'Y'
create_horitzontal_barchart_with_sign_plot <- function(data, X, Y, z, color1, color2, title, subtitle, 

  plot <- ggplot(data[1:top_num,], aes(x = reorder(.data[[X]], .data[[Y]]), y = .data[[Y]], fill= .data
    geom_bar(stat = "identity") +
    coord_flip() +
    scale_fill_manual(values = c("Positiu (Risc)" = color1, "Negatiu (Protecció)" = color2)) +
    labs(
      title = title,
      subtitle = subtitle,
      x = xlabel,
      y = ylabel
    ) +
    theme_minimal() +
    theme(
      panel.background = element_rect(fill = "white", color = "white"), # Fons del gràfic
      plot.background = element_rect(fill = "white", color = "white"),  # Fons del quadre del gràfic
      panel.grid.major = element_line(color = "grey80"),                # Línies de la quadrícula
      panel.grid.minor = element_blank(),                               # Línies de quadrícula menors
      legend.position = "top"                                           # Llegenda a la part superior
    )

  return(plot)
}
```

4 1. Càrrega i processament dels conjunts de dades

Originalment es tenen **dos conjunts de dades** diferents amb informació extreta de diferents pacients amb tumors de mama.

- **Conjunt de dades 1**:
  - Repositori d'origen: https://ftp.ncbi.nlm.nih.gov/geo/series/GSE140nnn/GSE140494/matrix/
  - Reanomenat com a: data_part_1.txt
- **Conjunt de dades 2**:
  - Repositori d'origen: https://ftp.ncbi.nlm.nih.gov/geo/series/GSE43nnn/GSE43365/matrix/
  - Reanomenat com a: data_part_2.txt

Ubicats dins la carpeta '/data/'

Ambdós conjunts de dades contenen una sèrie de caraterístiques descriptives de les pacients i els tumors de mama analitzats, entre elles, el **grau histològic** del tumor que és la variable classificadora que actua com a objectiu en la regressió. A més a més, contenen les **matrius d'expressions genètiques** que han

estat extretes amb la tècnica de microarrays i serviran per a definir la importància en cada un dels graus histològics.

La **plataforma** tecnològica utilitzada per a l'extracció de les dades genètiques en els dos conjunts de dades pertany a l'empresa *Affymetrix Probe Set*. Aquesta serveix d'enllaç entre les sondes corresponents a les mostres d'un valor d'expressió i la informació genètica (nom del gen o conjunts de gens).

La taula d'equivalències genètiques de la plataforma emprada en la tecnologia de microarrays es troba en la URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL570

S'ha anomenat 'GPL570-55999.txt' i s'ha ubicat dins la carpeta 'data/platform_microarrays'

4.1 1.1 Càrrega i processament conjunt de dades 1 (data_part_1)

A continuació es mostra tot el procés de càrrega, filtratge de variables i processament del conjunt de dades 1

```
#### FILE 1 PROCESSMENT: Processat i captura dels atributs identificadors dels pacients, el grau histol

### CÀRREGA DE DADES ###
# Ruta al fitxer de dades 1
file_path_1 <- paste0(ruta_input,"data_part_1.txt")

# Lectura de línies
lines_file_1 <- readLines(file_path_1)

cat("DATASET 1, PAS 1: Filtratge i captura de camps")
```

```
## DATASET 1, PAS 1: Filtratge i captura de camps
```

```
####### PAS 1: Filtratge #########
# Filtrem els camps que ens interessen en el fitxer.
# Id del pacient ve a !Sample_title
sample_title_1 <- grep("^!Sample_title", lines_file_1, value = TRUE)
# Capturem informació sobre el grau histològic del tumor en '!Sample_characteristics_ch1' (només el que
sample_characteristics_1 <- grep("^!Sample_characteristics_ch1.*tumor grade:", lines_file_1, value = TRU
# 'geo_accession' coné un identificador únic (en aquest cas GSM, per tant corresponent a una mostra indi
sample_geo_accession_1 <- grep("^!Sample_geo_accession", lines_file_1, value = TRUE)
# Es captura el tipus d'expressió genètica que es sempre RNA
sample_type_1 <- grep("^!Sample_type", lines_file_1, value = TRUE)
# Es captura el país de residència de la pacient en 'Sample_contact_country'
sample_country_1 <- grep("^!Sample_contact_country", lines_file_1, value = TRUE)
# En 'sample_source_name_ch1' es captura la font biològica d'origen de la mostra (tumor de mama, biòpsia
sample_name_1 <- grep("^!Sample_source_name_ch1", lines_file_1, value = TRUE)


cat("DATASET 1, PAS 2: Transformació de camps")
```

```
## DATASET 1, PAS 2: Transformació de camps
```

```
###### PAS 2: Transformació ######
# Separar els elements resultants pel delimitador "\t"
sample_title_split_1 <- strsplit(sample_title_1, "\t")[[1]]
sample_characteristics_split_1 <- strsplit(sample_characteristics_1, "\t")[[1]]
sample_geo_accession_split_1 <- strsplit(sample_geo_accession_1, "\t")[[1]]
sample_type_split_1 <- strsplit(sample_type_1, "\t")[[1]]
sample_country_split_1 <- strsplit(sample_country_1, "\t")[[1]]
sample_name_split_1 <- strsplit(sample_name_1, "\t")[[1]]
```

```r
# Per a 'sample_characteristics_2' tenim dues columnes on apareix la característica 'tumor_grade' i són
# els índex de posicions on surt el 'tumor_grade' en la primera és on no surt a la segona i viceversa. (
# en la primera, indica el 'cm_stage / er_status'. Quan no informa del 'tumor_grade' en la segona, indi

# De la primera charactersitics_ch1 ens quedem amb el 'tumor_grade' i on NO hi hagi un 'tumor_grade' pos
sample_characteristics_1_1 <- sample_characteristics_1[1]
sample_characteristics_1_1_split_2 <- strsplit(sample_characteristics_1_1, "\t")[[1]]
# En la segona characteristics_ch1 es procedeix de la mateixa manera
sample_characteristics_1_2 <- sample_characteristics_1[2]
sample_characteristics_1_2_split_2 <- strsplit(sample_characteristics_1_2, "\t")[[1]]

# Eliminar el primer element corresponent al nom de la columna o del camp
sample_title_clean_1 <- sample_title_split_1[-1]
sample_characteristics_1_1_clean_2 <- sample_characteristics_1_1_split_2[-1]
sample_characteristics_1_2_clean_2 <- sample_characteristics_1_2_split_2[-1]
sample_geo_accession_clean_1 <- sample_geo_accession_split_1[-1]
sample_type_1 <- sample_type_split_1[-1]
sample_country_clean_1 <- sample_country_split_1[-1]
sample_name_clean_1 <- sample_name_split_1[-1]

# Eliminació de les cometes entre els elements
sample_title_clean_1 <- gsub("\"", "", sample_title_clean_1)
sample_characteristics_1_1_clean_2 <- gsub("\"", "", sample_characteristics_1_1_clean_2)
sample_characteristics_1_2_clean_2 <- gsub("\"", "", sample_characteristics_1_2_clean_2)
sample_geo_accession_clean_1 <- gsub("\"", "", sample_geo_accession_clean_1)
sample_type_clean_1 <- gsub("\"", "", sample_type_1)
sample_country_clean_1 <- gsub("\"", "", sample_country_clean_1)
sample_name_clean_1 <- gsub("\"", "", sample_name_clean_1)

# En el cas dels graus histològics, capturem només el que es descriu en 'tumor_grade:'. Si no es conté
# la cadena 'tumor_grade:', aleshores el valor en aquesta posició es deixa a buit ja que no és el grau h
sample_characteristics_1_1_clean_2 <- ifelse(!grepl("tumor grade:", sample_characteristics_1_1_clean_2)
sample_characteristics_1_2_clean_2 <- ifelse(!grepl("tumor grade:", sample_characteristics_1_2_clean_2)
# Finalment, unificarem els resultats fent que en les posicions buides quedin els 'tumor_grade' compleme
sample_characteristics_1_clean_final <- paste0(sample_characteristics_1_1_clean_2, sample_characteristi
# Extraiem l'últim caràcter i el convertim en número. Si es troba '' es substitueix per NA
tumor_grades_1 <- as.numeric(sub(".*grade: ", "", sample_characteristics_1_clean_final))

cat("DATASET 1, PAS 3: Generació del dataset parcial sense la informació de l'expressió genètica")
```

```
## DATASET 1, PAS 3: Generació del dataset parcial sense la informació de l'expressió genètica
```

```r
# Genererem el dataset1 parcial sense la informació de l'expressió genètica del pacient amb les variables
# - TITLE (identificador del pacient amb tumor de mama, obligatori)
# - GEO_ACCESSION (identificador de la mostra del pacient, obligatori)
# - COUNTRY (país de residència del pacient, opcional)
# - TYPE (tipus estrucutral de la mostra: ARN, obligatori)
# - NAME (font biològica d'origen de la mostra, obligatori)
# - HISTOLOGICAL_GRADE (grau histològic del tumor que serà la variable objectiu obligatòria i podrà teni
dataset1_parcial <- data.frame(
  TITLE = sample_title_clean_1,
  GEO_ACCESSION = sample_geo_accession_clean_1,
  COUNTRY = sample_country_clean_1,
  TYPE = sample_type_clean_1,
```

```
  NAME = sample_name_clean_1,
  HISTOLOGICAL_GRADE = tumor_grades_1
)

head(dataset1_parcial)
```

```
##                       TITLE GEO_ACCESSION COUNTRY TYPE                NAME
## 1 BC_Patient1007_Genearray1    GSM4171495 Germany  RNA Pretreatment biopsy
## 2 BC_Patient1008_Genearray1    GSM4171496 Germany  RNA Pretreatment biopsy
## 3 BC_Patient1010_Genearray1    GSM4171497 Germany  RNA Pretreatment biopsy
## 4 BC_Patient1011_Genearray1    GSM4171498 Germany  RNA Pretreatment biopsy
## 5 BC_Patient1013_Genearray1    GSM4171499 Germany  RNA Pretreatment biopsy
## 6 BC_Patient1018_Genearray1    GSM4171500 Germany  RNA Pretreatment biopsy
##   HISTOLOGICAL_GRADE
## 1                  2
## 2                  2
## 3                  3
## 4                  3
## 5                  2
## 6                  3
```

```
tail(dataset1_parcial)
```

```
##                       TITLE GEO_ACCESSION COUNTRY TYPE                NAME
## 86 BC_Patient4005_Genearray5    GSM4171580 Germany  RNA Pretreatment biopsy
## 87 BC_Patient2019_Genearray5    GSM4171581 Germany  RNA Pretreatment biopsy
## 88 BC_Patient6001_Genearray5    GSM4171582 Germany  RNA Pretreatment biopsy
## 89 BC_Patient4012_Genearray5    GSM4171583 Germany  RNA Pretreatment biopsy
## 90 BC_Patient4003_Genearray5    GSM4171584 Germany  RNA Pretreatment biopsy
## 91 BC_Patient9021_Genearray5    GSM4171585 Germany  RNA Pretreatment biopsy
##    HISTOLOGICAL_GRADE
## 86                  2
## 87                 NA
## 88                  2
## 89                  2
## 90                  3
## 91                  2
```

Amb tot això, ja es té una part del conjunt de dades procedent de la primera font d'origen filtrat i processat amb la següent informació:

- TITLE (identificador del pacient amb tumor de mama, obligatori)
- GEO_ACCESSION (identificador de la mostra del pacient, obligatori)
- COUNTRY (país de residència del pacient, opcional)
- TYPE (tipus estrucutral de la mostra: per exemple ARN, obligatori)
- NAME (font biològica d'origen de la mostra, obligatori)
- HISTOLOGICAL_GRADE (grau histològic del tumor que serà la variable objectiu obligatòria i podrà tenir tres valors: 1, 2 o 3)

Posteriorment s'obté, es **filtra i es processa la matriu d'expressions genètiques extreta per microarrays**. En ella, cada fila representa un gen o conjunts de gens i cada columna una mostra de la pacient referida a través del GEO_ACCESSION.

```
cat("DATASET 1, PAS 4: Captura i càrrega de la matriu d'expressions genètiques extreta per microarrays")
```

```
## DATASET 1, PAS 4: Captura i càrrega de la matriu d'expressions genètiques extreta per microarrays
```

```r
### MATRIU DE DADES DE L'EXPRESSIÓ GENÈTICA ###
# En aquest fitxer els ID_REF apunten a la taula de la plataforma 'GPL570-55999'

# Ens situem en la zona de la matriu de dades, que s'inicia en '!series_matrix_table_begin' i conclou en
inici_matriu_1 <- grep("!series_matrix_table_begin", lines_file_1)
fi_matriu_1 <- grep("!series_matrix_table_end", lines_file_1)

# Capturem el contingut entre inici i fi
# Extraiem la part desitjada
matriu_expressio_genetica_1 <- lines_file_1[(inici_matriu_1 + 1):(fi_matriu_1 - 1)]

# Ho convertim a un dataframe on l'espai '\t' serà el delimitador de columna
matriu_expressio_genetica_1 <- read.table(text = matriu_expressio_genetica_1, header = TRUE, sep = "\t"

# Mostrem estadístiques bàsiques
cat("** DADES :** \n")
```

```
## ** DADES :**
```

```r
head(matriu_expressio_genetica_1[1:10],n=5)
```

```
##      ID_REF GSM4171495 GSM4171496 GSM4171497 GSM4171498 GSM4171499 GSM4171500
## 1 1007_s_at   9.607295   9.242755   9.004883  10.218012   9.792149   9.387552
## 2   1053_at   7.342585   8.140602   8.295291   7.703011   7.837335   7.949209
## 3    117_at   6.515292   6.287939   6.216547   6.057173   7.284177   6.245449
## 4    121_at   7.645063   7.637049   7.953448   7.681222   7.782360   7.692700
## 5 1255_g_at   3.644651   3.611982   3.435436   3.526985   3.860871   3.911908
##   GSM4171501 GSM4171502 GSM4171503
## 1   9.586211  10.267303  10.100368
## 2   7.186469   7.227751   7.309617
## 3   5.850214   6.167704   5.749173
## 4   7.505446   8.175508   7.795066
## 5   3.449997   3.646126   3.482868
```

```r
tail(matriu_expressio_genetica_1[1:10],n=5)
```

```
##                ID_REF GSM4171495 GSM4171496 GSM4171497 GSM4171498 GSM4171499
## 54671  AFFX-ThrX-5_at   7.112670   7.784184   7.465943   7.744283   8.339830
## 54672  AFFX-ThrX-M_at   7.707325   8.025933   7.883073   8.153580   8.642472
## 54673 AFFX-TrpnX-3_at   2.930642   2.995437   2.886472   2.945537   3.088699
## 54674 AFFX-TrpnX-5_at   3.409822   3.513638   3.518748   3.380261   3.474563
## 54675 AFFX-TrpnX-M_at   3.058811   3.107095   3.353181   3.205128   3.300049
##       GSM4171500 GSM4171501 GSM4171502 GSM4171503
## 54671   7.555737   7.761287   7.014222   7.980122
## 54672   7.895892   8.243721   7.326720   8.457530
## 54673   2.975360   3.126392   3.057823   2.865499
## 54674   3.449950   3.420131   3.403961   3.381593
## 54675   3.221628   3.298365   3.247559   3.117706
```

```r
cat("** ESTRUCTURA: **\n")
```

```
## ** ESTRUCTURA: **
```

```r
str(matriu_expressio_genetica_1)
```

```
## 'data.frame':    54675 obs. of  92 variables:
##  $ ID_REF    : chr  "1007_s_at" "1053_at" "117_at" "121_at" ...
```

```
##  $ GSM4171495: num  9.61 7.34 6.52 7.65 3.64 ...
##  $ GSM4171496: num  9.24 8.14 6.29 7.64 3.61 ...
##  $ GSM4171497: num  9 8.3 6.22 7.95 3.44 ...
##  $ GSM4171498: num  10.22 7.7 6.06 7.68 3.53 ...
##  $ GSM4171499: num  9.79 7.84 7.28 7.78 3.86 ...
##  $ GSM4171500: num  9.39 7.95 6.25 7.69 3.91 ...
##  $ GSM4171501: num  9.59 7.19 5.85 7.51 3.45 ...
##  $ GSM4171502: num  10.27 7.23 6.17 8.18 3.65 ...
##  $ GSM4171503: num  10.1 7.31 5.75 7.8 3.48 ...
##  $ GSM4171504: num  9.86 7.53 6.33 7.65 3.5 ...
##  $ GSM4171505: num  9.94 7.48 5.8 7.51 3.46 ...
##  $ GSM4171506: num  9.85 8.94 6.01 7.94 3.65 ...
##  $ GSM4171507: num  10 7.01 5.14 7.79 3.46 ...
##  $ GSM4171508: num  9.5 8.65 5.81 7.72 3.65 ...
##  $ GSM4171509: num  10.36 7.49 5.78 7.79 3.65 ...
##  $ GSM4171510: num  10.35 7.86 6.19 7.54 4.22 ...
##  $ GSM4171511: num  10.18 7.8 7.21 7.78 4.13 ...
##  $ GSM4171512: num  9.37 7.24 5.55 7.59 3.43 ...
##  $ GSM4171513: num  9.78 7.4 5.67 7.79 3.42 ...
##  $ GSM4171514: num  9.26 8.18 5.6 7.79 3.61 ...
##  $ GSM4171515: num  9.77 7.04 6.43 8.01 3.62 ...
##  $ GSM4171516: num  9.6 7.98 6.39 7.76 3.42 ...
##  $ GSM4171517: num  9.9 7.66 6.7 7.53 3.69 ...
##  $ GSM4171518: num  9.29 7.97 6 7.69 3.44 ...
##  $ GSM4171519: num  9.98 6.96 5.2 7.94 3.49 ...
##  $ GSM4171520: num  10.05 6.74 6.12 7.62 3.48 ...
##  $ GSM4171521: num  9.31 8.48 5.98 7.41 3.68 ...
##  $ GSM4171522: num  10.05 8.44 5.98 7.88 3.62 ...
##  $ GSM4171523: num  10.02 7.53 6.05 7.55 3.58 ...
##  $ GSM4171524: num  9.44 7.51 6.09 7.45 3.46 ...
##  $ GSM4171525: num  9.79 7.88 5.99 7.73 3.54 ...
##  $ GSM4171526: num  9.95 7.54 6.55 7.51 3.55 ...
##  $ GSM4171527: num  9.57 7.28 5.82 7.6 3.51 ...
##  $ GSM4171528: num  9.19 7.04 6.54 7.42 3.38 ...
##  $ GSM4171529: num  9.97 7.54 5.53 7.53 3.38 ...
##  $ GSM4171530: num  9.62 7.88 7.21 7.78 3.76 ...
##  $ GSM4171531: num  8.78 8.78 5.83 7.84 3.61 ...
##  $ GSM4171532: num  9.43 7.09 6.19 7.67 3.5 ...
##  $ GSM4171533: num  9.19 7.98 5.98 7.67 3.6 ...
##  $ GSM4171534: num  9.37 7.53 7.03 7.94 3.69 ...
##  $ GSM4171535: num  10.04 8.87 5.57 7.52 3.71 ...
##  $ GSM4171536: num  10.04 7.75 5.83 7.57 3.55 ...
##  $ GSM4171537: num  9.46 8.16 6.22 7.74 3.6 ...
##  $ GSM4171538: num  10.27 7.3 6.33 7.81 3.52 ...
##  $ GSM4171539: num  9.8 7.58 6.22 7.69 3.44 ...
##  $ GSM4171540: num  8.87 7.53 7.36 7.59 4.28 ...
##  $ GSM4171541: num  9.94 8.8 6.54 7.42 3.7 ...
##  $ GSM4171542: num  9.47 7.29 6.3 7.67 3.6 ...
##  $ GSM4171543: num  9.76 7.67 6.82 7.51 3.6 ...
##  $ GSM4171544: num  8.75 6.98 6.09 7.69 4.32 ...
##  $ GSM4171545: num  9.38 7.57 6.19 7.54 3.53 ...
##  $ GSM4171546: num  8.83 7.23 6.8 7.73 3.99 ...
##  $ GSM4171547: num  9.19 7.44 6.12 7.65 3.55 ...
##  $ GSM4171548: num  8.79 7.34 7.24 7.73 3.51 ...
```

```
##  $ GSM4171549: num  10.08 7.8 5.74 7.82 3.72 ...
##  $ GSM4171550: num  9.64 7.28 5.76 7.67 3.55 ...
##  $ GSM4171551: num  10.23 7.86 6.23 7.59 3.43 ...
##  $ GSM4171552: num  9.67 7.19 6.14 7.29 3.49 ...
##  $ GSM4171553: num  10.78 7.8 6.33 7.57 3.54 ...
##  $ GSM4171554: num  9.4 7.15 6.88 7.35 3.48 ...
##  $ GSM4171555: num  9.86 7.2 6.41 7.93 3.62 ...
##  $ GSM4171556: num  10.06 7.35 6.16 7.6 3.54 ...
##  $ GSM4171557: num  10.12 7.35 5.95 7.67 3.66 ...
##  $ GSM4171558: num  9.97 7.55 5.99 7.73 3.71 ...
##  $ GSM4171559: num  10.19 7.29 5.34 7.81 3.49 ...
##  $ GSM4171560: num  10.19 7.34 5.97 7.6 3.68 ...
##  $ GSM4171561: num  10.05 7.33 6.06 7.75 3.61 ...
##  $ GSM4171562: num  10.11 8.13 5.84 7.79 3.62 ...
##  $ GSM4171563: num  9.8 7.52 6.19 7.81 3.57 ...
##  $ GSM4171564: num  9.59 6.98 5.97 7.54 3.5 ...
##  $ GSM4171565: num  9.15 7.97 6.43 7.4 3.62 ...
##  $ GSM4171566: num  10.03 7.96 7.03 7.57 3.54 ...
##  $ GSM4171567: num  9.66 7.36 6.69 7.75 3.68 ...
##  $ GSM4171568: num  10.11 7.37 6.37 7.45 3.52 ...
##  $ GSM4171569: num  9.28 7.64 7.83 7.63 3.71 ...
##  $ GSM4171570: num  9.39 7.42 6.14 7.45 3.44 ...
##  $ GSM4171571: num  9.71 7.05 5.87 7.59 3.58 ...
##  $ GSM4171572: num  9.41 8.05 6.87 7.67 3.71 ...
##  $ GSM4171573: num  10.46 7.59 7.18 7.63 3.92 ...
##  $ GSM4171574: num  9.16 8.27 6.77 7.9 4.81 ...
##  $ GSM4171575: num  9.65 7.37 6.01 7.54 3.55 ...
##  $ GSM4171576: num  10.37 9.41 6.05 7.55 3.59 ...
##  $ GSM4171577: num  10.45 7.73 6.6 8.18 3.43 ...
##  $ GSM4171578: num  8.58 7.08 6.59 7.69 3.49 ...
##  $ GSM4171579: num  9.34 7.31 6.72 7.73 3.31 ...
##  $ GSM4171580: num  9.07 6.87 6.17 7.76 3.55 ...
##  $ GSM4171581: num  9.44 7.08 7.72 7.62 3.56 ...
##  $ GSM4171582: num  9.33 6.73 5.47 7.63 3.54 ...
##  $ GSM4171583: num  9.66 7.62 6.01 7.76 3.59 ...
##  $ GSM4171584: num  9.4 6.84 5.25 7.78 3.5 ...
##  $ GSM4171585: num  9.67 6.65 6.09 7.36 3.38 ...
```

```
cat("** ESTADÍSTIQUES BÀSIQUES: **\n")
```

```
## ** ESTADÍSTIQUES BÀSIQUES: **
```

```
summary(matriu_expressio_genetica_1)
```

```
##     ID_REF             GSM4171495        GSM4171496        GSM4171497
##  Length:54675       Min.   : 2.701   Min.   : 2.696   Min.   : 2.627
##  Class :character   1st Qu.: 4.039   1st Qu.: 4.051   1st Qu.: 4.060
##  Mode  :character   Median : 5.143   Median : 5.163   Median : 5.168
##                     Mean   : 5.717   Mean   : 5.735   Mean   : 5.722
##                     3rd Qu.: 7.131   3rd Qu.: 7.146   3rd Qu.: 7.105
##                     Max.   :14.544   Max.   :14.217   Max.   :14.429
##    GSM4171498        GSM4171499        GSM4171500        GSM4171501
##  Min.   : 2.664   Min.   : 2.677   Min.   : 2.687   Min.   : 2.691
##  1st Qu.: 4.036   1st Qu.: 4.049   1st Qu.: 4.047   1st Qu.: 4.040
##  Median : 5.149   Median : 5.150   Median : 5.149   Median : 5.166
```

```
##  Mean   : 5.717   Mean   : 5.698   Mean   : 5.728   Mean   : 5.733
##  3rd Qu.: 7.130   3rd Qu.: 7.060   3rd Qu.: 7.133   3rd Qu.: 7.161
##  Max.   :14.480   Max.   :14.471   Max.   :14.370   Max.   :14.459
##    GSM4171502       GSM4171503       GSM4171504       GSM4171505
##  Min.   : 2.668   Min.   : 2.698   Min.   : 2.709   Min.   : 2.680
##  1st Qu.: 4.051   1st Qu.: 4.054   1st Qu.: 4.044   1st Qu.: 4.037
##  Median : 5.156   Median : 5.144   Median : 5.149   Median : 5.164
##  Mean   : 5.714   Mean   : 5.723   Mean   : 5.707   Mean   : 5.739
##  3rd Qu.: 7.089   3rd Qu.: 7.125   3rd Qu.: 7.083   3rd Qu.: 7.181
##  Max.   :14.460   Max.   :14.458   Max.   :14.513   Max.   :14.358
##    GSM4171506       GSM4171507       GSM4171508       GSM4171509
##  Min.   : 2.714   Min.   : 2.674   Min.   : 2.697   Min.   : 2.698
##  1st Qu.: 4.086   1st Qu.: 4.073   1st Qu.: 4.058   1st Qu.: 4.048
##  Median : 5.143   Median : 5.144   Median : 5.164   Median : 5.160
##  Mean   : 5.689   Mean   : 5.715   Mean   : 5.719   Mean   : 5.717
##  3rd Qu.: 6.977   3rd Qu.: 7.070   3rd Qu.: 7.093   3rd Qu.: 7.106
##  Max.   :14.257   Max.   :14.772   Max.   :14.311   Max.   :14.265
##    GSM4171510       GSM4171511       GSM4171512       GSM4171513
##  Min.   : 2.683   Min.   : 2.628   Min.   : 2.673   Min.   : 2.655
##  1st Qu.: 4.032   1st Qu.: 4.052   1st Qu.: 4.040   1st Qu.: 4.056
##  Median : 5.141   Median : 5.156   Median : 5.186   Median : 5.167
##  Mean   : 5.733   Mean   : 5.710   Mean   : 5.730   Mean   : 5.708
##  3rd Qu.: 7.185   3rd Qu.: 7.090   3rd Qu.: 7.141   3rd Qu.: 7.067
##  Max.   :14.367   Max.   :14.473   Max.   :14.478   Max.   :14.507
##    GSM4171514       GSM4171515       GSM4171516       GSM4171517
##  Min.   : 2.663   Min.   : 2.680   Min.   : 2.682   Min.   : 2.648
##  1st Qu.: 4.047   1st Qu.: 4.065   1st Qu.: 4.063   1st Qu.: 4.041
##  Median : 5.154   Median : 5.176   Median : 5.160   Median : 5.163
##  Mean   : 5.710   Mean   : 5.681   Mean   : 5.713   Mean   : 5.723
##  3rd Qu.: 7.094   3rd Qu.: 7.001   3rd Qu.: 7.062   3rd Qu.: 7.135
##  Max.   :14.464   Max.   :14.502   Max.   :14.620   Max.   :14.675
##    GSM4171518       GSM4171519       GSM4171520       GSM4171521
##  Min.   : 2.682   Min.   : 2.656   Min.   : 2.711   Min.   : 2.649
##  1st Qu.: 4.071   1st Qu.: 4.069   1st Qu.: 4.066   1st Qu.: 4.053
##  Median : 5.178   Median : 5.173   Median : 5.183   Median : 5.164
##  Mean   : 5.715   Mean   : 5.694   Mean   : 5.734   Mean   : 5.729
##  3rd Qu.: 7.063   3rd Qu.: 7.008   3rd Qu.: 7.115   3rd Qu.: 7.118
##  Max.   :14.420   Max.   :14.824   Max.   :14.705   Max.   :14.530
##    GSM4171522       GSM4171523       GSM4171524       GSM4171525
##  Min.   : 2.651   Min.   : 2.653   Min.   : 2.725   Min.   : 2.657
##  1st Qu.: 4.056   1st Qu.: 4.045   1st Qu.: 4.031   1st Qu.: 4.076
##  Median : 5.156   Median : 5.174   Median : 5.152   Median : 5.167
##  Mean   : 5.721   Mean   : 5.735   Mean   : 5.739   Mean   : 5.728
##  3rd Qu.: 7.107   3rd Qu.: 7.150   3rd Qu.: 7.206   3rd Qu.: 7.104
##  Max.   :14.318   Max.   :14.368   Max.   :14.290   Max.   :14.537
##    GSM4171526       GSM4171527       GSM4171528       GSM4171529
##  Min.   : 2.603   Min.   : 2.689   Min.   : 2.690   Min.   : 2.697
##  1st Qu.: 4.064   1st Qu.: 4.065   1st Qu.: 4.039   1st Qu.: 4.053
##  Median : 5.191   Median : 5.155   Median : 5.181   Median : 5.178
##  Mean   : 5.720   Mean   : 5.738   Mean   : 5.747   Mean   : 5.711
##  3rd Qu.: 7.072   3rd Qu.: 7.148   3rd Qu.: 7.197   3rd Qu.: 7.072
##  Max.   :14.564   Max.   :14.473   Max.   :14.395   Max.   :14.430
##    GSM4171530       GSM4171531       GSM4171532       GSM4171533
##  Min.   : 2.648   Min.   : 2.750   Min.   : 2.667   Min.   : 2.612
```

```
##  1st Qu.: 4.056   1st Qu.: 4.062   1st Qu.: 4.057   1st Qu.: 4.049
##  Median : 5.174   Median : 5.157   Median : 5.178   Median : 5.182
##  Mean   : 5.738   Mean   : 5.711   Mean   : 5.739   Mean   : 5.741
##  3rd Qu.: 7.141   3rd Qu.: 7.069   3rd Qu.: 7.147   3rd Qu.: 7.160
##  Max.   :14.352   Max.   :14.301   Max.   :14.431   Max.   :14.500
##    GSM4171534       GSM4171535       GSM4171536       GSM4171537
##  Min.   : 2.623   Min.   : 2.667   Min.   : 2.614   Min.   : 2.681
##  1st Qu.: 4.069   1st Qu.: 4.052   1st Qu.: 4.046   1st Qu.: 4.064
##  Median : 5.185   Median : 5.158   Median : 5.170   Median : 5.146
##  Mean   : 5.714   Mean   : 5.721   Mean   : 5.726   Mean   : 5.712
##  3rd Qu.: 7.066   3rd Qu.: 7.116   3rd Qu.: 7.123   3rd Qu.: 7.067
##  Max.   :14.347   Max.   :14.518   Max.   :14.142   Max.   :14.469
##    GSM4171538       GSM4171539       GSM4171540       GSM4171541
##  Min.   : 2.655   Min.   : 2.665   Min.   : 2.690   Min.   : 2.697
##  1st Qu.: 4.058   1st Qu.: 4.053   1st Qu.: 4.056   1st Qu.: 4.039
##  Median : 5.171   Median : 5.164   Median : 5.180   Median : 5.163
##  Mean   : 5.721   Mean   : 5.718   Mean   : 5.721   Mean   : 5.729
##  3rd Qu.: 7.081   3rd Qu.: 7.102   3rd Qu.: 7.109   3rd Qu.: 7.135
##  Max.   :14.523   Max.   :14.215   Max.   :14.286   Max.   :14.408
##    GSM4171542       GSM4171543       GSM4171544       GSM4171545
##  Min.   : 2.710   Min.   : 2.657   Min.   : 2.668   Min.   : 2.637
##  1st Qu.: 4.063   1st Qu.: 4.044   1st Qu.: 4.054   1st Qu.: 4.032
##  Median : 5.184   Median : 5.166   Median : 5.175   Median : 5.166
##  Mean   : 5.717   Mean   : 5.725   Mean   : 5.707   Mean   : 5.723
##  3rd Qu.: 7.081   3rd Qu.: 7.125   3rd Qu.: 7.073   3rd Qu.: 7.140
##  Max.   :14.328   Max.   :14.370   Max.   :14.389   Max.   :14.625
##    GSM4171546       GSM4171547       GSM4171548       GSM4171549
##  Min.   : 2.655   Min.   : 2.613   Min.   : 2.674   Min.   : 2.674
##  1st Qu.: 4.058   1st Qu.: 4.055   1st Qu.: 4.041   1st Qu.: 4.062
##  Median : 5.161   Median : 5.177   Median : 5.173   Median : 5.174
##  Mean   : 5.716   Mean   : 5.734   Mean   : 5.720   Mean   : 5.710
##  3rd Qu.: 7.103   3rd Qu.: 7.143   3rd Qu.: 7.130   3rd Qu.: 7.061
##  Max.   :14.386   Max.   :14.307   Max.   :14.523   Max.   :14.385
##    GSM4171550       GSM4171551       GSM4171552       GSM4171553
##  Min.   : 2.642   Min.   : 2.687   Min.   : 2.635   Min.   : 2.684
##  1st Qu.: 4.051   1st Qu.: 4.030   1st Qu.: 4.049   1st Qu.: 4.037
##  Median : 5.151   Median : 5.164   Median : 5.168   Median : 5.160
##  Mean   : 5.716   Mean   : 5.739   Mean   : 5.736   Mean   : 5.722
##  3rd Qu.: 7.111   3rd Qu.: 7.178   3rd Qu.: 7.166   3rd Qu.: 7.137
##  Max.   :14.483   Max.   :14.517   Max.   :14.446   Max.   :14.512
##    GSM4171554       GSM4171555       GSM4171556       GSM4171557
##  Min.   : 2.587   Min.   : 2.636   Min.   : 2.632   Min.   : 2.659
##  1st Qu.: 4.061   1st Qu.: 4.033   1st Qu.: 4.057   1st Qu.: 4.042
##  Median : 5.185   Median : 5.151   Median : 5.158   Median : 5.168
##  Mean   : 5.724   Mean   : 5.707   Mean   : 5.719   Mean   : 5.730
##  3rd Qu.: 7.111   3rd Qu.: 7.114   3rd Qu.: 7.094   3rd Qu.: 7.148
##  Max.   :14.506   Max.   :14.586   Max.   :14.432   Max.   :14.318
##    GSM4171558       GSM4171559       GSM4171560       GSM4171561
##  Min.   : 2.685   Min.   : 2.690   Min.   : 2.717   Min.   : 2.659
##  1st Qu.: 4.040   1st Qu.: 4.066   1st Qu.: 4.038   1st Qu.: 4.050
##  Median : 5.145   Median : 5.158   Median : 5.144   Median : 5.163
##  Mean   : 5.729   Mean   : 5.720   Mean   : 5.716   Mean   : 5.719
##  3rd Qu.: 7.155   3rd Qu.: 7.077   3rd Qu.: 7.124   3rd Qu.: 7.083
##  Max.   :14.590   Max.   :14.540   Max.   :14.473   Max.   :14.497
```

```
##    GSM4171562       GSM4171563       GSM4171564       GSM4171565
##  Min.   : 2.689   Min.   : 2.685   Min.   : 2.632   Min.   : 2.717
##  1st Qu.: 4.071   1st Qu.: 4.053   1st Qu.: 4.048   1st Qu.: 4.038
##  Median : 5.154   Median : 5.150   Median : 5.163   Median : 5.164
##  Mean   : 5.715   Mean   : 5.704   Mean   : 5.730   Mean   : 5.725
##  3rd Qu.: 7.056   3rd Qu.: 7.077   3rd Qu.: 7.157   3rd Qu.: 7.136
##  Max.   :14.326   Max.   :14.398   Max.   :14.370   Max.   :14.386
##    GSM4171566       GSM4171567       GSM4171568       GSM4171569
##  Min.   : 2.693   Min.   : 2.588   Min.   : 2.662   Min.   : 2.680
##  1st Qu.: 4.028   1st Qu.: 4.048   1st Qu.: 4.030   1st Qu.: 4.052
##  Median : 5.131   Median : 5.182   Median : 5.173   Median : 5.129
##  Mean   : 5.727   Mean   : 5.740   Mean   : 5.735   Mean   : 5.715
##  3rd Qu.: 7.157   3rd Qu.: 7.176   3rd Qu.: 7.173   3rd Qu.: 7.102
##  Max.   :14.380   Max.   :14.410   Max.   :14.453   Max.   :14.525
##    GSM4171570       GSM4171571       GSM4171572       GSM4171573
##  Min.   : 2.665   Min.   : 2.705   Min.   : 2.644   Min.   : 2.702
##  1st Qu.: 4.028   1st Qu.: 4.039   1st Qu.: 4.048   1st Qu.: 4.045
##  Median : 5.167   Median : 5.161   Median : 5.143   Median : 5.162
##  Mean   : 5.729   Mean   : 5.734   Mean   : 5.718   Mean   : 5.729
##  3rd Qu.: 7.166   3rd Qu.: 7.176   3rd Qu.: 7.101   3rd Qu.: 7.156
##  Max.   :14.598   Max.   :14.375   Max.   :14.531   Max.   :14.397
##    GSM4171574       GSM4171575       GSM4171576       GSM4171577
##  Min.   : 2.674   Min.   : 2.658   Min.   : 2.682   Min.   : 2.684
##  1st Qu.: 4.043   1st Qu.: 4.050   1st Qu.: 4.054   1st Qu.: 4.038
##  Median : 5.143   Median : 5.164   Median : 5.152   Median : 5.140
##  Mean   : 5.722   Mean   : 5.724   Mean   : 5.713   Mean   : 5.732
##  3rd Qu.: 7.137   3rd Qu.: 7.138   3rd Qu.: 7.092   3rd Qu.: 7.163
##  Max.   :14.684   Max.   :14.388   Max.   :14.375   Max.   :14.503
##    GSM4171578       GSM4171579       GSM4171580       GSM4171581
##  Min.   : 2.662   Min.   : 2.694   Min.   : 2.664   Min.   : 2.686
##  1st Qu.: 4.052   1st Qu.: 4.064   1st Qu.: 4.038   1st Qu.: 4.019
##  Median : 5.147   Median : 5.137   Median : 5.136   Median : 5.152
##  Mean   : 5.711   Mean   : 5.716   Mean   : 5.721   Mean   : 5.742
##  3rd Qu.: 7.092   3rd Qu.: 7.092   3rd Qu.: 7.149   3rd Qu.: 7.223
##  Max.   :14.520   Max.   :14.701   Max.   :14.574   Max.   :14.403
##    GSM4171582       GSM4171583       GSM4171584       GSM4171585
##  Min.   : 2.651   Min.   : 2.700   Min.   : 2.695   Min.   : 2.719
##  1st Qu.: 4.049   1st Qu.: 4.046   1st Qu.: 4.055   1st Qu.: 4.048
##  Median : 5.145   Median : 5.156   Median : 5.139   Median : 5.153
##  Mean   : 5.718   Mean   : 5.727   Mean   : 5.709   Mean   : 5.724
##  3rd Qu.: 7.120   3rd Qu.: 7.140   3rd Qu.: 7.099   3rd Qu.: 7.139
##  Max.   :14.303   Max.   :14.416   Max.   :14.275   Max.   :14.457
```

```r
# Obtenir dimensions de la matriu
dimensions <- dim(matriu_expressio_genetica_1)
# Mostrar el nombre de files y columnas
cat("La matriu d'expressions genètiques té ", dimensions[1], "files i", dimensions[2], "columnes.\n")
```

```
## La matriu d'expressions genètiques té  54675 files i 92 columnes.
```

4.2 1.2 Normalització de valors conjunt de dades 1

Per tal de tenir totes les dades (les d'aquest conjunt de dades i les del segon) en la mateixa escala, es sotmet la matriu d'expressions genètiques a un **procés de normalització** que deixa tots els seus valors entre **l'interval 0-1**.

Es fa servir la funció genèrica creada en el punt 0: normalize_matrix

```
cat("DATASET 1, PAS 5: Normalització de dades")
```

```
## DATASET 1, PAS 5: Normalització de dades
```

```
## Per tal de tenir tots els valors en la mateixa escala a l'hora d'unir els dos conjunts de dades, real
## Cridem la funció creada en 'general_functions' anomenada 'normalize_matrix'
# Passem els valors numèrics (no s'inclou la primera columna dels identificadors)
matriu_expressio_genetica_1_norm <- normalize_matrix(matriu_expressio_genetica_1[, 2:ncol(matriu_express

# Reafegim la columna original dels identificadors
matriu_expressio_genetica_1_norm <- cbind(matriu_expressio_genetica_1[, 1, drop = FALSE], matriu_express

cat("La matriu d'expressions genètiques ja ha estat normalitzada amb valors entre 0 i 1:\n")
```

```
## La matriu d'expressions genètiques ja ha estat normalitzada amb valors entre 0 i 1:
```

```
head(matriu_expressio_genetica_1_norm[1:10],n=5)
```

```
##      ID_REF GSM4171495 GSM4171496 GSM4171497 GSM4171498 GSM4171499 GSM4171500
## 1 1007_s_at 0.57370824 0.54391787 0.52447888 0.62361636  0.5888146  0.5557507
## 2   1053_at 0.38863518 0.45384949 0.46649077 0.41808939  0.4290664  0.4382088
## 3    117_at 0.32102844 0.30244912 0.29661488 0.28359081  0.3838621  0.2989768
## 4    121_at 0.41335383 0.41269897 0.43855523 0.41630876  0.4245738  0.4172467
## 5 1255_g_at 0.08643846 0.08376874 0.06934137 0.07682278  0.1041081  0.1082788
##   GSM4171501 GSM4171502 GSM4171503
## 1 0.57198526 0.62764442 0.61400243
## 2 0.37587734 0.37925092 0.38594108
## 3 0.26667801 0.29262347 0.25842092
## 4 0.40194426 0.45670203 0.42561219
## 5 0.07053131 0.08655906 0.07321753
```

```
tail(matriu_expressio_genetica_1_norm[1:10],n=5)
```

```
##                ID_REF GSM4171495 GSM4171496 GSM4171497 GSM4171498 GSM4171499
## 54671  AFFX-ThrX-5_at 0.36984646 0.42472287 0.39871608 0.42146214 0.47013050
## 54672  AFFX-ThrX-M_at 0.41844192 0.44447870 0.43280412 0.45491007 0.49486253
## 54673 AFFX-TrpnX-3_at 0.02808937 0.03338444 0.02447978 0.02930659 0.04100585
## 54674 AFFX-TrpnX-5_at 0.06724816 0.07573205 0.07614964 0.06483242 0.07253882
## 54675 AFFX-TrpnX-M_at 0.03856339 0.04250918 0.06261943 0.05052048 0.05827746
##       GSM4171500 GSM4171501 GSM4171502 GSM4171503
## 54671 0.40605409 0.42285172 0.36180125 0.44073501
## 54672 0.43385170 0.46227643 0.38733872 0.47974899
## 54673 0.03174374 0.04408614 0.03848265 0.02276585
## 54674 0.07052744 0.06809061 0.06676919 0.06494127
## 54675 0.05186887 0.05813985 0.05398796 0.04337632
```

4.3 1.3 Càrrega i processament conjunt de dades 2 (data_part_2)

A continuació es mostra tot el procés de càrrega, filtratge de variables i processament del conjunt de dades 2

```
#### FILE 2 PROCESSMENT ####

### CÀRREGA DADES ###
# Ruta al fitxer origen
file_path_2 <- paste0(ruta_input,"data_part_2.txt")
```

```r
# Lectura de línies
lines_file_2 <- readLines(file_path_2)

cat("DATASET 2, PAS 1: Filtratge i captura de camps")
```

```
## DATASET 2, PAS 1: Filtratge i captura de camps
```

```r
####### FILTERING #########
# Filtrem els camps que ens interessen en cada fitxer.
# Id del pacient ve a !Sample_title
sample_title_2 <- grep("^!Sample_title", lines_file_2, value = TRUE)
# Capturem informació sobre el grau histològic del tumor en '!Sample_characteristics_ch1' (només el que
sample_characteristics_2 <- grep("^!Sample_characteristics_ch1.*grade histo:", lines_file_2, value = TRU
# 'geo_accession' coné un identificador únic (en aquest cas GSM, per tant corresponent a una mostra indi
sample_geo_accession_2 <- grep("^!Sample_geo_accession", lines_file_2, value = TRUE)
# Es captura el tipus d'expressió genètica que es sempre RNA
sample_type_2 <- grep("^!Sample_type", lines_file_2, value = TRUE)
# Es captura el país de residència de la pacient en 'Sample_contact_country'
sample_country_2 <- grep("^!Sample_contact_country", lines_file_2, value = TRUE)
# En 'sample_source_name_ch1' es captura la font biològica d'origen de la mostra (tumor de mama, biòpsia
sample_name_2 <- grep("^!Sample_source_name_ch1", lines_file_2, value = TRUE)


cat("DATASET 2, PAS 2: Transformació de camps")
```

```
## DATASET 2, PAS 2: Transformació de camps
```

```r
###### TRANSFORMATION ######
# Separar els elements pel delimitador "\t"
sample_title_split_2 <- strsplit(sample_title_2, "\t")[[1]]
sample_characteristics_split_2 <- strsplit(sample_characteristics_2, "\t")[[1]]
sample_geo_accession_split_2 <- strsplit(sample_geo_accession_2, "\t")[[1]]
sample_type_split_2 <- strsplit(sample_type_2, "\t")[[1]]
sample_country_split_2 <- strsplit(sample_country_2, "\t")[[1]]
sample_name_split_2 <- strsplit(sample_name_2, "\t")[[1]]

# Eliminar el primer element corresponent al nom de la columna o del camp
sample_title_clean_2 <- sample_title_split_2[-1]
sample_characteristics_clean_2 <- sample_characteristics_split_2[-1]
sample_geo_accession_clean_2 <- sample_geo_accession_split_2[-1]
sample_type_clean_2 <- sample_type_split_2[-1]
sample_country_clean_2 <- sample_country_split_2[-1]
sample_name_clean_2 <- sample_name_split_2[-1]

# Eliminació de les cometes entre els elements
sample_title_clean_2 <- gsub("\"", "", sample_title_clean_2)
sample_characteristics_clean_2 <- gsub("\"", "", sample_characteristics_clean_2)
sample_geo_accession_clean_2 <- gsub("\"", "", sample_geo_accession_clean_2)
sample_type_clean_2 <- gsub("\"", "", sample_type_clean_2)
sample_country_clean_2 <- gsub("\"", "", sample_country_clean_2)
sample_name_clean_2 <- gsub("\"", "", sample_name_clean_2)

# En el cas dels graus histològics, capturem només l'element numèric que ens interessa que es troba desp
tumor_grades_2 <- as.numeric(sub(".*histo: ", "", sample_characteristics_clean_2))
```

```r
cat("DATASET 2, PAS 3: Generació del dataset parcial sense la informació de l'expressió genètica")
```

```
## DATASET 2, PAS 3: Generació del dataset parcial sense la informació de l'expressió genètica
```

```r
# Generem el dataset2 parcial sense la informació de l'expressió genètica del pacient amb les variables
# - TITLE (identificador del pacient amb tumor de mama, obligatori)
# - GEO_ACCESSION (identificador de la mostra del pacient, obligatori)
# - COUNTRY (país de residència del pacient, opcional)
# - TYPE (tipus estrucutral de la mostra: ARN, obligatori)
# - NAME (font biològica d'origen de la mostra, obligatori)
# - HISTOLOGICAL_GRADE (grau histològic del tumor que serà la variable objectiu obligatòria i podrà ten
dataset2_parcial <- data.frame(
  TITLE = sample_title_clean_2,
  GEO_ACCESSION = sample_geo_accession_clean_2,
  COUNTRY = sample_country_clean_2,
  TYPE = sample_type_clean_2,
  NAME = sample_name_clean_2,
  HISTOLOGICAL_GRADE = tumor_grades_2
)

head(dataset2_parcial)
```

```
##         TITLE GEO_ACCESSION COUNTRY TYPE
## 1 PK-09-26449    GSM1061032     USA  RNA
## 2 PK-09-26710    GSM1061033     USA  RNA
## 3 PK-09-26716    GSM1061034     USA  RNA
## 4 PK-09-26718    GSM1061035     USA  RNA
## 5 PK-09-26722    GSM1061036     USA  RNA
## 6 PK-09-26723    GSM1061037     USA  RNA
##                                         NAME HISTOLOGICAL_GRADE
## 1 primary breast tumor - fresh surgical samples                2
## 2 primary breast tumor - fresh surgical samples                2
## 3 primary breast tumor - fresh surgical samples                2
## 4 primary breast tumor - fresh surgical samples                2
## 5 primary breast tumor - fresh surgical samples                3
## 6 primary breast tumor - fresh surgical samples                1
```

```r
tail(dataset2_parcial)
```

```
##           TITLE GEO_ACCESSION COUNTRY TYPE
## 106 PK-09-27531    GSM1061137     USA  RNA
## 107 PK-09-27864    GSM1061138     USA  RNA
## 108 PK-09-27865    GSM1061139     USA  RNA
## 109 PK-09-27866    GSM1061140     USA  RNA
## 110 PK-09-27869    GSM1061141     USA  RNA
## 111 PK-09-27870    GSM1061142     USA  RNA
##                                           NAME HISTOLOGICAL_GRADE
## 106 primary breast tumor - fresh surgical samples                3
## 107 primary breast tumor - fresh surgical samples                3
## 108 primary breast tumor - fresh surgical samples                1
## 109 primary breast tumor - fresh surgical samples                3
## 110 primary breast tumor - fresh surgical samples                1
## 111 primary breast tumor - fresh surgical samples                2
```

Anàlogament al conjunt de dades 1 de la secció anterior, ja es té una part del conjunt de dades filtrat i processat amb la següent informació:

- TITLE (identificador del pacient amb tumor de mama, obligatori)
- GEO_ACCESSION (identificador de la mostra del pacient, obligatori)
- COUNTRY (país de residència del pacient, opcional)
- TYPE (tipus estrucutral de la mostra: ARN, obligatori)
- NAME (font biològica d'origen de la mostra, obligatori)
- HISTOLOGICAL_GRADE (grau histològic del tumor que serà la variable objectiu obligatòria i podrà tenir tres valors: 1, 2 o 3)

El següent pas és, per tant,**l'obtenció i processament de la matriu d'expressions genètiques** extreta per microarrays aplicant els mateixos criteris.

```
cat("DATASET 2, PAS 4: Captura i càrrega de la matriu d'expressions genètiques extreta per microarrays")
```

```
## DATASET 2, PAS 4: Captura i càrrega de la matriu d'expressions genètiques extreta per microarrays
```

```
### MATRIU DE DADES DE L'EXPRESSIÓ GENÈTICA ###
# En aquest fitxer els ID_REF apunten a la taula de la plataforma 'GPL570-55999'

# Ens situem en la zona de la matriu de dades, que s'inicia en '!series_matrix_table_begin' i conclou en
inici_matriu_2 <- grep("!series_matrix_table_begin", lines_file_2)
fi_matriu_2 <- grep("!series_matrix_table_end", lines_file_2)

# Capturem el contingut entre inici i fi
# Extraiem la part desitjada
matriu_expressio_genetica_2 <- lines_file_2[(inici_matriu_2 + 1):(fi_matriu_2 - 1)]

# Ho convertim a un dataframe on l'espai '\t' serà el delimitador de columna
matriu_expressio_genetica_2 <- read.table(text = matriu_expressio_genetica_2, header = TRUE, sep = "\t"

# Mostrem estadístiques bàsiques
cat("** DADES : **\n")
```

```
## ** DADES : **
```

```
head(matriu_expressio_genetica_2[1:10],n=5)
```

```
##      ID_REF GSM1061032  GSM1061033 GSM1061034 GSM1061035 GSM1061036 GSM1061037
## 1 1007_s_at 5351.82927 10674.46947 2799.49522 2366.85808 4000.68026 3332.72475
## 2   1053_at  290.55460   766.17108  569.83552  394.32404  745.86748  443.38110
## 3    117_at  263.76875   260.25713  307.37530  560.69762  263.24342  326.65987
## 4    121_at  523.70470   307.00543  461.55857  379.45614  690.92321  445.10054
## 5 1255_g_at   85.62612    36.14115   53.99714   25.09399   56.66522   47.81017
##   GSM1061038 GSM1061039  GSM1061040
## 1 2159.45859 4402.35401 2829.155711
## 2  404.43983  654.55475  352.142683
## 3  279.86931  160.18542  221.093597
## 4  389.21866  397.78377  378.543538
## 5   35.11095   17.40934    4.632572
```

```
tail(matriu_expressio_genetica_2[1:10],n=5)
```

```
##                ID_REF  GSM1061032 GSM1061033 GSM1061034 GSM1061035 GSM1061036
## 54671  AFFX-ThrX-5_at  370.463629  108.37852 135.473788 355.048985  46.815268
## 54672  AFFX-ThrX-M_at 1035.112099  306.64380 317.729110 646.476423 209.176557
## 54673 AFFX-TrpnX-3_at   37.881275   15.66693  25.818021   2.166817   2.179472
## 54674 AFFX-TrpnX-5_at   25.786976   19.26360  47.688931  12.769100  10.603362
## 54675 AFFX-TrpnX-M_at    7.668208   12.60433   4.766902   3.824329   4.302800
```

```
##       GSM1061037 GSM1061038 GSM1061039 GSM1061040
## 54671 476.447492  428.97583 175.402201  71.387854
## 54672 739.928175  638.11899 394.934343 265.801970
## 54673   4.193657   17.66680  19.414596   2.776103
## 54674  19.759899   21.79114  33.588351  25.514416
## 54675   4.604787   30.24339   1.903061   2.671613
```

```
cat("** ESTRUCTURA:**\n ")
```

```
## ** ESTRUCTURA:**
##
```

```
str(matriu_expressio_genetica_2)
```

```
## 'data.frame':    54675 obs. of  112 variables:
##  $ ID_REF    : chr  "1007_s_at" "1053_at" "117_at" "121_at" ...
##  $ GSM1061032: num  5351.8 290.6 263.8 523.7 85.6 ...
##  $ GSM1061033: num  10674.5 766.2 260.3 307 36.1 ...
##  $ GSM1061034: num  2799 570 307 462 54 ...
##  $ GSM1061035: num  2366.9 394.3 560.7 379.5 25.1 ...
##  $ GSM1061036: num  4000.7 745.9 263.2 690.9 56.7 ...
##  $ GSM1061037: num  3332.7 443.4 326.7 445.1 47.8 ...
##  $ GSM1061038: num  2159.5 404.4 279.9 389.2 35.1 ...
##  $ GSM1061039: num  4402.4 654.6 160.2 397.8 17.4 ...
##  $ GSM1061040: num  2829.16 352.14 221.09 378.54 4.63 ...
##  $ GSM1061041: num  5460.7 446 238.8 364 60.1 ...
##  $ GSM1061042: num  4353.5 530.6 139.9 245.4 49.5 ...
##  $ GSM1061043: num  2831 662.2 208.1 298.1 63.7 ...
##  $ GSM1061044: num  6436.3 602.6 87.8 443.7 24.7 ...
##  $ GSM1061045: num  4362.6 512.4 227 288.5 7.7 ...
##  $ GSM1061046: num  3293.04 477.07 788.76 317.96 6.07 ...
##  $ GSM1061047: num  6731.3 870.9 749.4 373.4 21.8 ...
##  $ GSM1061048: num  2976.6 598.1 262.5 560.9 44.9 ...
##  $ GSM1061049: num  2187.2 422.5 358.7 373.2 12.6 ...
##  $ GSM1061050: num  3514 500 162 419 34 ...
##  $ GSM1061051: num  4478.71 351.49 265.09 325.44 7.02 ...
##  $ GSM1061052: num  2685.1 466.7 140.2 237.5 47.8 ...
##  $ GSM1061053: num  4620.6 707.6 314.3 382.6 49.9 ...
##  $ GSM1061054: num  2358.33 291.04 282.78 374.17 2.22 ...
##  $ GSM1061055: num  3294.9 730.9 381.4 376.4 23.1 ...
##  $ GSM1061056: num  3706.6 603.2 133 337.6 13.4 ...
##  $ GSM1061057: num  3151.1 619.9 245.2 331.7 32.9 ...
##  $ GSM1061058: num  4652 506.8 391.2 530.2 17.6 ...
##  $ GSM1061059: num  3555.5 471.8 427.1 428 36.6 ...
##  $ GSM1061060: num  3707.9 1083.4 330.8 228.7 24.4 ...
##  $ GSM1061061: num  1762.47 539.52 658.55 141.65 4.41 ...
##  $ GSM1061062: num  3021 423 360 444 51 ...
##  $ GSM1061063: num  3429.4 744.6 207.5 495.3 16.2 ...
##  $ GSM1061064: num  4755.3 385.5 370.9 565.1 32.3 ...
##  $ GSM1061065: num  5052 923.4 188 382.3 30.3 ...
##  $ GSM1061066: num  2241 477.2 308.5 437.9 38.2 ...
##  $ GSM1061067: num  3990.9 313 795.7 373.9 53.8 ...
##  $ GSM1061068: num  2559.7 529.8 397.8 407.4 37.3 ...
##  $ GSM1061069: num  4244.3 406.5 283.1 635 78.6 ...
##  $ GSM1061070: num  4970 422 180 587 24 ...
```

```
##  $ GSM1061071: num  6032.9 480.5 272.3 322 10.7 ...
##  $ GSM1061072: num  6366.77 783.72 233.25 355.85 3.45 ...
##  $ GSM1061073: num  7664.9 465.4 1635.3 469.4 37.5 ...
##  $ GSM1061074: num  3610.4 435.3 204 165.3 30.5 ...
##  $ GSM1061075: num  4371 580.5 372.3 374.1 39.8 ...
##  $ GSM1061076: num  5930.07 393.02 193.67 269.76 5.93 ...
##  $ GSM1061077: num  4552 646 158 328 12 ...
##  $ GSM1061078: num  4466.4 465.8 180.5 754.7 64.1 ...
##  $ GSM1061079: num  3705.6 569.3 292.5 500.7 50.7 ...
##  $ GSM1061080: num  2604.9 334.3 209.1 293.4 66.9 ...
##  $ GSM1061081: num  3003.5 570.3 394.9 365.5 27.4 ...
##  $ GSM1061082: num  4227.7 1114.9 128.7 479.2 22.2 ...
##  $ GSM1061083: num  4206.6 412.5 201.3 524.4 51.5 ...
##  $ GSM1061084: num  2973.55 473.51 202.63 495.54 6.31 ...
##  $ GSM1061085: num  6089.4 683.1 234.5 272.8 37.7 ...
##  $ GSM1061086: num  3663.6 555.3 174.5 527.7 93.3 ...
##  $ GSM1061087: num  2952.85 552.72 235.75 395.68 6.45 ...
##  $ GSM1061088: num  2962 608 178 296 30 ...
##  $ GSM1061089: num  4242.1 392.3 288.4 275.7 43.9 ...
##  $ GSM1061090: num  3506.5 394.2 342.6 372.4 33.5 ...
##  $ GSM1061091: num  2683 506.5 226.9 301.2 50.1 ...
##  $ GSM1061092: num  2269.8 453.4 150.7 392.6 16.7 ...
##  $ GSM1061093: num  4090.5 511.1 137.7 263.8 40.6 ...
##  $ GSM1061094: num  4422 526 356 306 35 ...
##  $ GSM1061095: num  7108.7 459.6 211.9 705 65.8 ...
##  $ GSM1061096: num  1829 358 361 302 60 ...
##  $ GSM1061097: num  4442.2 520 321.9 544.6 27.2 ...
##  $ GSM1061098: num  3913.95 583.99 145.1 434.26 6.05 ...
##  $ GSM1061099: num  2296.9 475.4 146 283.9 23.5 ...
##  $ GSM1061100: num  3708 451 210 465 33 ...
##  $ GSM1061101: num  5265.2 505.4 334.7 596 17.6 ...
##  $ GSM1061102: num  5789.3 509 127.6 378.5 29.4 ...
##  $ GSM1061103: num  3784.1 377.4 306.1 541 43.5 ...
##  $ GSM1061104: num  3524.08 429.05 284.71 363.39 8.51 ...
##  $ GSM1061105: num  4802.61 409.52 199.49 273.67 4.02 ...
##  $ GSM1061106: num  4939.4 505.4 265.3 374.4 28.8 ...
##  $ GSM1061107: num  5105.8 675.7 517 349.9 27.9 ...
##  $ GSM1061108: num  4348.3 373.3 125.6 315.4 48.4 ...
##  $ GSM1061109: num  2516.4 381 183.8 434 14.2 ...
##  $ GSM1061110: num  3948.7 527.4 104.9 289 5.5 ...
##  $ GSM1061111: num  6524.5 473.5 485.4 273 27.8 ...
##  $ GSM1061112: num  3860.8 648.6 158.5 408.5 4.5 ...
##  $ GSM1061113: num  4718.6 808.6 168.7 252.7 30.5 ...
##  $ GSM1061114: num  2328.7 522.4 330.2 408.8 62.7 ...
##  $ GSM1061115: num  5734.6 297 144.7 267 30.6 ...
##  $ GSM1061116: num  4036.39 721.45 171.08 647.98 4.22 ...
##  $ GSM1061117: num  2571.03 406.39 267.51 376.36 8.25 ...
##  $ GSM1061118: num  3800.6 917.1 289.3 384.1 94.9 ...
##  $ GSM1061119: num  6731.3 443.3 173.2 688.3 11.2 ...
##  $ GSM1061120: num  3150.2 555.6 151.7 583 66.9 ...
##  $ GSM1061121: num  3527.19 517.91 191.13 345.4 1.43 ...
##  $ GSM1061122: num  4473.2 655.2 168.3 238.8 11.6 ...
##  $ GSM1061123: num  4728.4 488.2 170.4 527.3 44.9 ...
##  $ GSM1061124: num  2848 595 213 471 28 ...
```

```
##  $ GSM1061125: num  7321.9 493 271.9 302.5 22.5 ...
##  $ GSM1061126: num  3368.8 328.2 185.4 253.1 60.8 ...
##  $ GSM1061127: num  4703.7 561.7 162.1 299.2 23.2 ...
##  $ GSM1061128: num  6602.49 618.22 279.82 648.12 6.71 ...
##  $ GSM1061129: num  6237.6 371.5 228.6 304.5 23.3 ...
##   [list output truncated]
```

```
cat("** ESTADÍSTIQUES BÀSIQUES: **\n")
```

```
## ** ESTADÍSTIQUES BÀSIQUES: **
```

```
summary(matriu_expressio_genetica_2)
```

```
##     ID_REF             GSM1061032          GSM1061033          GSM1061034
##  Length:54675       Min.   :     0.38   Min.   :    0.10   Min.   :    0.23
##  Class :character   1st Qu.:    32.82   1st Qu.:   23.75   1st Qu.:   35.89
##  Mode  :character   Median :   110.22   Median :   91.72   Median :  122.72
##                     Mean   :   893.96   Mean   :  760.54   Mean   :  849.34
##                     3rd Qu.:   460.21   3rd Qu.:  474.59   3rd Qu.:  525.77
##                     Max.   :100277.34   Max.   :41715.29   Max.   :65088.85
##    GSM1061035          GSM1061036          GSM1061037          GSM1061038
##  Min.   :    0.29   Min.   :    0.27   Min.   :    0.38   Min.   :     0.27
##  1st Qu.:   31.32   1st Qu.:   33.39   1st Qu.:   37.82   1st Qu.:    39.82
##  Median :  114.82   Median :  110.72   Median :  122.36   Median :   130.55
##  Mean   :  805.05   Mean   :  839.10   Mean   :  895.65   Mean   :   914.36
##  3rd Qu.:  505.47   3rd Qu.:  485.01   3rd Qu.:  516.93   3rd Qu.:   515.84
##  Max.   :53365.64   Max.   :62172.32   Max.   :86183.67   Max.   :101248.91
##    GSM1061039          GSM1061040          GSM1061041          GSM1061042
##  Min.   :    0.27   Min.   :    0.24   Min.   :    0.14   Min.   :    0.18
##  1st Qu.:   24.09   1st Qu.:   29.76   1st Qu.:   29.25   1st Qu.:   30.16
##  Median :   85.76   Median :  104.61   Median :   99.01   Median :  108.72
##  Mean   :  817.33   Mean   :  838.54   Mean   :  827.89   Mean   :  799.37
##  3rd Qu.:  426.48   3rd Qu.:  479.24   3rd Qu.:  481.59   3rd Qu.:  500.35
##  Max.   :41877.51   Max.   :59516.46   Max.   :68165.68   Max.   :54705.72
##    GSM1061043          GSM1061044          GSM1061045          GSM1061046
##  Min.   :    0.18   Min.   :    0.13   Min.   :    0.18   Min.   :    0.13
##  1st Qu.:   31.34   1st Qu.:   29.11   1st Qu.:   34.24   1st Qu.:   29.16
##  Median :  106.85   Median :  100.62   Median :  116.09   Median :  104.09
##  Mean   :  901.33   Mean   :  873.21   Mean   :  802.01   Mean   :  806.08
##  3rd Qu.:  467.17   3rd Qu.:  458.03   3rd Qu.:  491.15   3rd Qu.:  497.87
##  Max.   :79737.63   Max.   :77374.81   Max.   :56217.40   Max.   :48804.93
##    GSM1061047          GSM1061048          GSM1061049          GSM1061050
##  Min.   :    0.07   Min.   :    0.19   Min.   :    0.3   Min.   :    0.14
##  1st Qu.:   27.79   1st Qu.:   32.25   1st Qu.:   33.9   1st Qu.:   36.81
##  Median :   95.21   Median :  113.98   Median :  113.1   Median :  128.12
##  Mean   :  795.47   Mean   :  795.06   Mean   :  915.3   Mean   :  817.01
##  3rd Qu.:  457.08   3rd Qu.:  526.14   3rd Qu.:  485.5   3rd Qu.:  520.89
##  Max.   :44529.99   Max.   :56264.97   Max.   :90301.3   Max.   :61299.95
##    GSM1061051          GSM1061052          GSM1061053          GSM1061054
##  Min.   :    0.13   Min.   :    0.33   Min.   :    0.10   Min.   :    0.27
##  1st Qu.:   34.62   1st Qu.:   28.83   1st Qu.:   34.24   1st Qu.:   32.20
##  Median :  118.98   Median :   97.82   Median :  120.02   Median :  123.36
##  Mean   :  819.66   Mean   :  835.11   Mean   :  835.54   Mean   :  828.09
##  3rd Qu.:  519.03   3rd Qu.:  483.72   3rd Qu.:  504.43   3rd Qu.:  531.12
##  Max.   :67708.87   Max.   :68060.88   Max.   :71484.95   Max.   :65271.45
```

```
##    GSM1061055         GSM1061056         GSM1061057         GSM1061058
##  Min.   :    0.20   Min.   :    0.24   Min.   :    0.28   Min.   :    0.17
##  1st Qu.:   36.26   1st Qu.:   30.01   1st Qu.:   33.23   1st Qu.:   35.22
##  Median :  123.09   Median :  100.22   Median :  112.85   Median :  122.08
##  Mean   :  822.61   Mean   :  832.95   Mean   :  816.37   Mean   :  789.28
##  3rd Qu.:  524.24   3rd Qu.:  479.54   3rd Qu.:  490.53   3rd Qu.:  512.62
##  Max.   :64192.71   Max.   :58810.99   Max.   :63583.74   Max.   :57280.00
##    GSM1061059         GSM1061060         GSM1061061         GSM1061062
##  Min.   :    0.16   Min.   :    0.18   Min.   :    0.28   Min.   :    0.13
##  1st Qu.:   29.87   1st Qu.:   27.90   1st Qu.:   29.30   1st Qu.:   29.26
##  Median :  111.53   Median :  105.06   Median :   97.21   Median :   99.96
##  Mean   :  843.88   Mean   :  839.15   Mean   :  854.46   Mean   :  792.50
##  3rd Qu.:  521.63   3rd Qu.:  503.48   3rd Qu.:  471.83   3rd Qu.:  472.73
##  Max.   :69039.31   Max.   :54085.56   Max.   :60955.44   Max.   :42212.10
##    GSM1061063         GSM1061064         GSM1061065         GSM1061066
##  Min.   :    0.35   Min.   :    0.07   Min.   :    0.24   Min.   :    0.14
##  1st Qu.:   37.56   1st Qu.:   30.71   1st Qu.:   29.32   1st Qu.:   35.01
##  Median :  122.76   Median :  115.82   Median :  107.75   Median :  112.59
##  Mean   :  903.82   Mean   :  788.39   Mean   :  815.69   Mean   :  853.16
##  3rd Qu.:  502.40   3rd Qu.:  525.12   3rd Qu.:  503.55   3rd Qu.:  503.06
##  Max.   :91534.40   Max.   :48940.27   Max.   :58457.12   Max.   :64206.63
##    GSM1061067         GSM1061068         GSM1061069         GSM1061070
##  Min.   :    0.22   Min.   :    0.12   Min.   :    0.25   Min.   :    0.10
##  1st Qu.:   34.30   1st Qu.:   32.85   1st Qu.:   37.90   1st Qu.:   23.54
##  Median :  107.98   Median :  117.07   Median :  120.49   Median :   91.40
##  Mean   :  831.12   Mean   :  810.56   Mean   :  850.51   Mean   :  808.80
##  3rd Qu.:  487.54   3rd Qu.:  505.87   3rd Qu.:  495.52   3rd Qu.:  485.77
##  Max.   :61352.33   Max.   :56017.27   Max.   :71456.62   Max.   :52147.42
##    GSM1061071         GSM1061072         GSM1061073         GSM1061074
##  Min.   :    0.27   Min.   :    0.05   Min.   :    0.25   Min.   :    0.11
##  1st Qu.:   30.37   1st Qu.:   24.48   1st Qu.:   35.00   1st Qu.:   26.79
##  Median :  108.40   Median :   93.46   Median :  119.61   Median :  108.63
##  Mean   :  861.68   Mean   :  768.01   Mean   :  871.52   Mean   :  778.54
##  3rd Qu.:  495.92   3rd Qu.:  483.66   3rd Qu.:  499.96   3rd Qu.:  525.63
##  Max.   :97916.93   Max.   :44088.77   Max.   :85223.06   Max.   :47434.50
##    GSM1061075         GSM1061076         GSM1061077         GSM1061078
##  Min.   :    0.13   Min.   :    0.17   Min.   :    0.18   Min.   :     0.34
##  1st Qu.:   30.59   1st Qu.:   27.87   1st Qu.:   31.73   1st Qu.:    44.29
##  Median :  112.73   Median :  102.19   Median :  116.90   Median :   156.10
##  Mean   :  824.85   Mean   :  840.73   Mean   :  795.65   Mean   :   993.39
##  3rd Qu.:  513.39   3rd Qu.:  485.70   3rd Qu.:  519.81   3rd Qu.:   512.07
##  Max.   :61983.88   Max.   :58249.79   Max.   :49332.23   Max.   :234788.19
##    GSM1061079         GSM1061080         GSM1061081         GSM1061082
##  Min.   :    0.36   Min.   :    0.18   Min.   :    0.11   Min.   :    0.08
##  1st Qu.:   33.28   1st Qu.:   31.77   1st Qu.:   32.54   1st Qu.:   24.90
##  Median :  113.50   Median :  116.45   Median :  114.68   Median :   91.12
##  Mean   :  825.06   Mean   :  819.98   Mean   :  820.22   Mean   :  835.11
##  3rd Qu.:  508.52   3rd Qu.:  511.76   3rd Qu.:  503.48   3rd Qu.:  472.85
##  Max.   :62989.49   Max.   :62962.32   Max.   :55012.33   Max.   :61668.06
##    GSM1061083         GSM1061084         GSM1061085         GSM1061086
##  Min.   :    0.26   Min.   :    0.18   Min.   :    0.17   Min.   :     0.30
##  1st Qu.:   33.91   1st Qu.:   29.68   1st Qu.:   31.82   1st Qu.:    36.41
##  Median :  118.89   Median :  108.38   Median :  112.28   Median :   118.76
##  Mean   :  848.66   Mean   :  849.88   Mean   :  834.93   Mean   :   900.56
```

```
##  3rd Qu.:  523.38   3rd Qu.:  492.85   3rd Qu.:  492.78   3rd Qu.:   502.62
##  Max.   :70209.56   Max.   :66202.01   Max.   :61511.35   Max.   :104254.68
##    GSM1061087         GSM1061088         GSM1061089         GSM1061090
##  Min.   :    0.14   Min.   :    0.25   Min.   :    0.22   Min.   :    0.18
##  1st Qu.:   27.49   1st Qu.:   35.51   1st Qu.:   28.78   1st Qu.:   26.74
##  Median :  102.32   Median :  124.34   Median :  102.44   Median :   95.13
##  Mean   :  852.08   Mean   :  841.72   Mean   :  888.39   Mean   :  812.16
##  3rd Qu.:  502.71   3rd Qu.:  531.32   3rd Qu.:  483.33   3rd Qu.:  485.04
##  Max.   :65076.97   Max.   :64987.08   Max.   :71811.29   Max.   :60321.28
##    GSM1061091         GSM1061092         GSM1061093         GSM1061094
##  Min.   :    0.28   Min.   :    0.14   Min.   :    0.15   Min.   :    0.14
##  1st Qu.:   33.36   1st Qu.:   33.11   1st Qu.:   27.10   1st Qu.:   29.56
##  Median :  106.15   Median :  109.30   Median :  104.70   Median :  104.71
##  Mean   :  885.37   Mean   :  922.88   Mean   :  783.21   Mean   :  839.24
##  3rd Qu.:  486.75   3rd Qu.:  494.37   3rd Qu.:  506.83   3rd Qu.:  496.10
##  Max.   :75503.56   Max.   :88406.98   Max.   :51857.03   Max.   :60908.69
##    GSM1061095         GSM1061096         GSM1061097         GSM1061098
##  Min.   :    0.29   Min.   :    0.20   Min.   :    0.09   Min.   :    0.09
##  1st Qu.:   32.19   1st Qu.:   29.16   1st Qu.:   31.04   1st Qu.:   31.77
##  Median :  109.07   Median :  109.47   Median :  118.01   Median :  109.40
##  Mean   :  838.12   Mean   :  816.66   Mean   :  813.61   Mean   :  806.97
##  3rd Qu.:  506.86   3rd Qu.:  510.53   3rd Qu.:  527.90   3rd Qu.:  487.28
##  Max.   :69271.25   Max.   :57909.67   Max.   :52495.49   Max.   :46792.76
##    GSM1061099         GSM1061100         GSM1061101         GSM1061102
##  Min.   :    0.30   Min.   :    0.20   Min.   :    0.16   Min.   :    0.21
##  1st Qu.:   33.53   1st Qu.:   30.46   1st Qu.:   20.46   1st Qu.:   23.24
##  Median :  125.65   Median :  102.58   Median :   81.77   Median :   87.06
##  Mean   :  813.91   Mean   :  839.75   Mean   :  771.31   Mean   :  812.46
##  3rd Qu.:  528.60   3rd Qu.:  466.30   3rd Qu.:  473.44   3rd Qu.:  462.98
##  Max.   :59215.68   Max.   :59196.59   Max.   :42002.11   Max.   :45538.81
##    GSM1061103         GSM1061104         GSM1061105         GSM1061106
##  Min.   :    0.13   Min.   :    0.20   Min.   :    0.11   Min.   :    0.34
##  1st Qu.:   34.84   1st Qu.:   30.64   1st Qu.:   25.45   1st Qu.:   35.66
##  Median :  111.58   Median :  113.41   Median :   99.84   Median :  115.02
##  Mean   :  857.98   Mean   :  842.61   Mean   :  772.31   Mean   :  863.27
##  3rd Qu.:  490.20   3rd Qu.:  500.27   3rd Qu.:  498.87   3rd Qu.:  495.81
##  Max.   :79810.14   Max.   :66051.85   Max.   :43582.46   Max.   :65147.41
##    GSM1061107         GSM1061108         GSM1061109         GSM1061110
##  Min.   :    0.13   Min.   :    0.17   Min.   :    0.13   Min.   :    0.11
##  1st Qu.:   27.11   1st Qu.:   28.28   1st Qu.:   35.28   1st Qu.:   26.01
##  Median :   97.96   Median :   95.23   Median :  120.46   Median :   98.77
##  Mean   :  846.06   Mean   :  907.12   Mean   :  821.28   Mean   :  807.10
##  3rd Qu.:  482.11   3rd Qu.:  455.75   3rd Qu.:  515.51   3rd Qu.:  505.46
##  Max.   :65166.57   Max.   :69181.63   Max.   :58431.41   Max.   :53760.88
##    GSM1061111         GSM1061112         GSM1061113         GSM1061114
##  Min.   :    0.10   Min.   :    0.28   Min.   :    0.17   Min.   :    0.27
##  1st Qu.:   27.12   1st Qu.:   31.91   1st Qu.:   27.10   1st Qu.:   34.70
##  Median :   99.84   Median :  108.98   Median :   99.83   Median :  117.33
##  Mean   :  811.40   Mean   :  904.07   Mean   :  782.27   Mean   :  900.70
##  3rd Qu.:  479.61   3rd Qu.:  494.33   3rd Qu.:  484.56   3rd Qu.:  517.66
##  Max.   :61857.23   Max.   :70244.12   Max.   :44482.13   Max.   :77820.54
##    GSM1061115         GSM1061116         GSM1061117         GSM1061118
##  Min.   :    0.24   Min.   :     0.39   Min.   :    0.15   Min.   :    0.25
##  1st Qu.:   23.58   1st Qu.:    39.77   1st Qu.:   28.53   1st Qu.:   35.74
```

```
##  Median :   87.14   Median :   122.53   Median :  110.30   Median :  118.50
##  Mean   :  840.40   Mean   :   932.91   Mean   :  806.23   Mean   :  922.65
##  3rd Qu.:  475.73   3rd Qu.:   512.03   3rd Qu.:  509.34   3rd Qu.:  501.40
##  Max.   :58957.12   Max.   :144796.32   Max.   :58932.83   Max.   :85552.07
##    GSM1061119         GSM1061120         GSM1061121         GSM1061122
##  Min.   :     0.6   Min.   :    0.4   Min.   :    0.24   Min.   :    0.08
##  1st Qu.:    39.5   1st Qu.:   33.0   1st Qu.:   27.39   1st Qu.:   19.63
##  Median :   128.1   Median :  106.8   Median :  102.82   Median :   82.96
##  Mean   :   885.4   Mean   :  889.8   Mean   :  820.19   Mean   :  796.31
##  3rd Qu.:   532.8   3rd Qu.:  482.8   3rd Qu.:  495.97   3rd Qu.:  457.78
##  Max.   :156241.0   Max.   :78634.1   Max.   :51163.21   Max.   :41593.08
##    GSM1061123         GSM1061124         GSM1061125         GSM1061126
##  Min.   :    0.25   Min.   :    0.18   Min.   :    0.07   Min.   :    0.32
##  1st Qu.:   29.01   1st Qu.:   34.84   1st Qu.:   32.06   1st Qu.:   31.91
##  Median :  116.03   Median :  125.05   Median :  108.57   Median :  118.85
##  Mean   :  836.30   Mean   :  812.70   Mean   :  815.07   Mean   :  813.08
##  3rd Qu.:  540.74   3rd Qu.:  524.59   3rd Qu.:  499.23   3rd Qu.:  530.31
##  Max.   :66424.84   Max.   :66730.17   Max.   :57488.07   Max.   :56278.94
##    GSM1061127         GSM1061128         GSM1061129         GSM1061130
##  Min.   :    0.14   Min.   :    0.07   Min.   :    0.12   Min.   :    0.06
##  1st Qu.:   26.41   1st Qu.:   28.93   1st Qu.:   26.64   1st Qu.:   27.66
##  Median :  102.73   Median :  110.56   Median :  108.39   Median :  115.51
##  Mean   :  802.40   Mean   :  786.14   Mean   :  745.99   Mean   :  816.51
##  3rd Qu.:  495.51   3rd Qu.:  497.02   3rd Qu.:  512.28   3rd Qu.:  518.16
##  Max.   :51487.16   Max.   :44217.28   Max.   :39321.20   Max.   :62031.15
##    GSM1061131         GSM1061132         GSM1061133         GSM1061134
##  Min.   :    0.20   Min.   :    0.77   Min.   :    0.19   Min.   :    0.08
##  1st Qu.:   26.76   1st Qu.:   45.82   1st Qu.:   33.20   1st Qu.:   26.46
##  Median :  102.45   Median :  148.04   Median :  109.87   Median :  100.75
##  Mean   :  772.91   Mean   :  929.31   Mean   :  964.10   Mean   :  800.81
##  3rd Qu.:  490.45   3rd Qu.:  514.39   3rd Qu.:  445.52   3rd Qu.:  485.64
##  Max.   :48053.03   Max.   :184212.20   Max.   :119789.24   Max.   :46084.93
##    GSM1061135         GSM1061136         GSM1061137         GSM1061138
##  Min.   :    0.14   Min.   :    0.19   Min.   :    0.28   Min.   :    0.25
##  1st Qu.:   27.61   1st Qu.:   28.35   1st Qu.:   24.59   1st Qu.:   26.21
##  Median :  113.12   Median :  106.82   Median :   95.10   Median :  109.80
##  Mean   :  783.83   Mean   :  807.25   Mean   :  814.67   Mean   :  793.51
##  3rd Qu.:  521.15   3rd Qu.:  483.08   3rd Qu.:  469.73   3rd Qu.:  523.65
##  Max.   :51292.65   Max.   :49678.32   Max.   :51130.83   Max.   :47172.52
##    GSM1061139         GSM1061140         GSM1061141         GSM1061142
##  Min.   :    0.38   Min.   :    0.05   Min.   :    0.05   Min.   :     0.33
##  1st Qu.:   35.58   1st Qu.:   26.50   1st Qu.:   24.61   1st Qu.:    43.04
##  Median :  130.68   Median :  104.69   Median :  106.95   Median :   146.58
##  Mean   :  812.43   Mean   :  780.59   Mean   :  761.48   Mean   :  1011.14
##  3rd Qu.:  537.69   3rd Qu.:  507.57   3rd Qu.:  511.77   3rd Qu.:   528.29
##  Max.   :64208.59   Max.   :47771.81   Max.   :40392.57   Max.   :172078.50
```

```r
# Obtenir dimensions de la matriu
dimensions <- dim(matriu_expressio_genetica_2)
# Mostrar el nombre de files y columnas
cat("La matriu d'expressions genètiques té ", dimensions[1], "files i", dimensions[2], "columnes.\n")
```

```
## La matriu d'expressions genètiques té  54675 files i 112 columnes.
```

4.4 1.4 Normalització de valors conjunt de dades 2

Finalment, tal i com s'ha fet també en el conjunt de dades anterior, s'aplica una **normalització dels valors genètics (entre 0 i 1)** ja que en els dos conjunts de dades estan en escales amb intervals completament diferents.

```
cat("DATASET 2, PAS 5: Normalització de dades")
```

```
## DATASET 2, PAS 5: Normalització de dades
```

```
## Per tal de tenir tots els valors en la mateixa escala a l'hora d'unir els dos conjunts de dades, rea
## Cridem la funció creada en 'general_functions' anomenada 'normalize_matrix'
# Passem els valors numèrics (no s'inclou la primera columna dels identificadors)
matriu_expressio_genetica_2_norm <- normalize_matrix(matriu_expressio_genetica_2[, 2:ncol(matriu_expres

# Reafegim la columna original dels identificadors
matriu_expressio_genetica_2_norm <- cbind(matriu_expressio_genetica_2[, 1, drop = FALSE], matriu_expres
cat("La matriu d'expressions genètiques ja ha estat normalitzada amb valors entre 0 i 1:\n")
```

```
## La matriu d'expressions genètiques ja ha estat normalitzada amb valors entre 0 i 1:
```

```
head(matriu_expressio_genetica_2_norm[1:10],n=5)
```

```
##      ID_REF   GSM1061032   GSM1061033   GSM1061034   GSM1061035   GSM1061036
## 1 1007_s_at 0.0227940909 0.0454640613 0.0119232945 0.0100806242 0.0170393325
## 2   1053_at 0.0012373177 0.0032630439 0.0024268195 0.0016792883 0.0031765676
## 3    117_at 0.0011232325 0.0011082760 0.0013089598 0.0023878998 0.0011209951
## 4    121_at 0.0022303411 0.0013073844 0.0019656508 0.0016159635 0.0029425512
## 5 1255_g_at 0.0003644948 0.0001537304 0.0002297819 0.0001066788 0.0002411457
##     GSM1061037   GSM1061038   GSM1061039   GSM1061040
## 1 0.0141944038 0.0091972768 1.875012e-02 1.204962e-02
## 2 0.0018882301 0.0017223730 2.787652e-03 1.499631e-03
## 3 0.0013910958 0.0011918074 6.820547e-04 9.414723e-04
## 4 0.0018955535 0.0016575436 1.694024e-03 1.612077e-03
## 5 0.0002034306 0.0001493426 7.394866e-05 1.953036e-05
```

```
tail(matriu_expressio_genetica_2_norm[1:10],n=5)
```

```
##                ID_REF   GSM1061032   GSM1061033   GSM1061034   GSM1061035
## 54671  AFFX-ThrX-5_at 1.577663e-03 4.614008e-04 0.0005768039 1.512010e-03
## 54672  AFFX-ThrX-M_at 4.408506e-03 1.305844e-03 0.0013530583 2.753245e-03
## 54673 AFFX-TrpnX-3_at 1.611419e-04 6.652744e-05 0.0001097626 9.028321e-06
## 54674 AFFX-TrpnX-5_at 1.096303e-04 8.184624e-05 0.0002029142 5.418513e-05
## 54675 AFFX-TrpnX-M_at 3.245962e-05 5.348336e-05 0.0000201025 1.608793e-05
##          GSM1061036   GSM1061037   GSM1061038   GSM1061039   GSM1061040
## 54671 1.991932e-04 2.029065e-03 1.826876e-03 7.468654e-04 3.038517e-04
## 54672 8.907157e-04 3.151271e-03 2.717650e-03 1.681888e-03 1.131892e-03
## 54673 9.082220e-06 1.766096e-05 7.504521e-05 8.248935e-05 1.162337e-05
## 54674 4.496091e-05 8.396005e-05 9.261142e-05 1.428576e-04 1.084695e-04
## 54675 1.812581e-05 1.941202e-05 1.286109e-04 7.904942e-06 1.117833e-05
```

4.5 1.5 Obtenció d'informació de gens de la matriu a partir de la plataforma i unió dels conjunts de dades

Un cop generades i normalitzades les matrius d'expressions genètiques d'ambdós conjunts de dades, cal **accedir a la taula de la plataforma** corresponent de l'extracció per microarrays i obtenir la **informació sobre els símbols dels gens** que necessitem.

Les plataformes dels microarrays pertanyen a l'empresa Affymetrix. El nom dels gens el conté la columna 'GENE_SYMBOL'i l'element relacionant entre les matrius d'expressió és 'ID_REF.'

Per tant, s'afegeix a les dues matrius **la unió amb la nova columna informativa** sobre el símbols dels gens de cada mostra a partir de la taula de la plataforma.

```
# Ruta a la taula de la plataforma
file_to_platform_affymetrix <- paste0(ruta_input,"platform_microarrays/")
file_platofrm_name <- "GPL570-55999.txt"

cat("PROCESSAT PAS 1: Càrrega, processat i unió de les dades de la matriu d'expressió genètica obtinguda
```

```
## PROCESSAT PAS 1: Càrrega, processat i unió de les dades de la matriu d'expressió genètica obtinguda
```

```
# 1. Cridem la funció 'process_platform_table' de 'general_functions' que realitza tota la carrega, proc
# de la matriu d'expressió genètica obtinguda del conjunt de dades amb la informació proporcionada per
# El nexe d'unió és el ID_REF
# Per a la matriu del conjunt de dades 1
matriu_expressio_genetica_1_final <- process_platform_table(file_to_platform_affymetrix, file_platofrm_
# Per a la matriu del conjunt de dades 2
matriu_expressio_genetica_2_final <- process_platform_table(file_to_platform_affymetrix, file_platofrm_
```

Seguidament, cal **netejar i processar els resultats** de les matrius d'expressions genètiques unides amb la informació dels noms dels gens de la plataforma. Al realitzar aquesta unió ens trobem amb el següent:

- 1) Files amb **ID_REF que no existeix a la taula de la plataforma** i, en conseqüència, no se'n pot extreure el nom del gen o conjunt de gens. En cas de trobar-se, aquestes files **s'eliminen** de la matriu ja que es consideren sondes de control d'experiment sense correspondència.
- 2) Files que el seu **ID_REF apunta als mateixos símbols de gens** (GENE_SYMBOL igual) ja que es troba situat en diferents llocs. En aquest cas s'agrupen totes les files que tenen el mateix símbol de gen/s i com a valor d'expressió final s'assigna la seva **mitjana aritmètica**.

```
cat("PROCESSAT PAS 2: Procés de neteja de les dades d'expressió genètica unides amb la informació de la
```

```
## PROCESSAT PAS 2: Procés de neteja de les dades d'expressió genètica unides amb la informació de la pl
```

```
# 2. Cridem la funció 'merge_platform_with_gene_expression' que realitza el procés de neteja de les dade
# informació de la plataforma que s'han obtingut en el pas anterior

# Per a la matriu unificada del conjunt de dades 1
cat("Matriu del conjunt de dades 1:")
```

```
## Matriu del conjunt de dades 1:
```

```
matriu_expressio_genetica_1_final_proc <- process_merged_gene_expression(matriu_expressio_genetica_1_fi
```

```
## S'han trobat símbols de gens repetits en la columna GENE_SYMBOL, procedim a realitzar les mitjanes ar
## Tenim 8893  registres NO associats a cap identificador de la plataforma que seran eliminats.Els regis
```

```
# Per a la matriu unificada del conjunt de dades 2
cat("Matriu del conjunt de dades 2:")
```

```
## Matriu del conjunt de dades 2:
```

```
matriu_expressio_genetica_2_final_proc <- process_merged_gene_expression(matriu_expressio_genetica_2_fi
```

```
## S'han trobat símbols de gens repetits en la columna GENE_SYMBOL, procedim a realitzar les mitjanes ar
## Tenim 8893  registres NO associats a cap identificador de la plataforma que seran eliminats.Els regis
```

```r
# Visualizem els resultats
head(matriu_expressio_genetica_1_final_proc[1:10],n=5)
```

```
## # A tibble: 5 x 10
##   GENE_SYMBOL GSM4171495 GSM4171496 GSM4171497 GSM4171498 GSM4171499 GSM4171500
##   <chr>            <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1 A1BG             0.313      0.319      0.248      0.329      0.239      0.344
## 2 A1BG-AS1         0.158      0.218      0.168      0.201      0.176      0.169
## 3 A1CF             0.113      0.116      0.12       0.114      0.134      0.109
## 4 A2M              0.394      0.332      0.363      0.438      0.396      0.385
## 5 A2M-AS1          0.159      0.224      0.132      0.216      0.282      0.142
## # i 3 more variables: GSM4171501 <dbl>, GSM4171502 <dbl>, GSM4171503 <dbl>
```

```r
tail(matriu_expressio_genetica_1_final_proc[1:10],n=5)
```

```
## # A tibble: 5 x 10
##   GENE_SYMBOL  GSM4171495 GSM4171496 GSM4171497 GSM4171498 GSM4171499 GSM4171500
##   <chr>             <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1 abParts ,  ~      0.226      0.555      0.477      0.498      0.5        0.31
## 2 av27s1 ,  T~      0.084      0.117      0.098      0.132      0.082      0.088
## 3 hsa-let-7a-~      0.233      0.252      0.277      0.229      0.25       0.26
## 4 hsa-let-7a-~      0.212      0.222      0.18       0.143      0.157      0.151
## 5 mir-223           0.146      0.146      0.14       0.246      0.226      0.143
## # i 3 more variables: GSM4171501 <dbl>, GSM4171502 <dbl>, GSM4171503 <dbl>
```

```r
head(matriu_expressio_genetica_2_final_proc[1:10],n=5)
```

```
## # A tibble: 5 x 10
##   GENE_SYMBOL GSM1061032 GSM1061033 GSM1061034 GSM1061035 GSM1061036 GSM1061037
##   <chr>            <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1 A1BG             0.002      0          0.001      0.001      0.001      0.001
## 2 A1BG-AS1         0          0          0          0          0.001      0.001
## 3 A1CF             0          0          0          0          0          0
## 4 A2M              0.024      0.007      0.019      0.037      0.028      0.051
## 5 A2M-AS1          0.002      0.001      0.002      0.002      0.001      0.001
## # i 3 more variables: GSM1061038 <dbl>, GSM1061039 <dbl>, GSM1061040 <dbl>
```

```r
tail(matriu_expressio_genetica_2_final_proc[1:10],n=5)
```

```
## # A tibble: 5 x 10
##   GENE_SYMBOL  GSM1061032 GSM1061033 GSM1061034 GSM1061035 GSM1061036 GSM1061037
##   <chr>             <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1 abParts ,  ~      0          0          0.011      0.002      0.005      0
## 2 av27s1 ,  T~      0          0          0          0          0          0
## 3 hsa-let-7a-~      0          0          0          0          0          0
## 4 hsa-let-7a-~      0.001      0.001      0          0.001      0          0.001
## 5 mir-223           0          0          0          0.001      0          0
## # i 3 more variables: GSM1061038 <dbl>, GSM1061039 <dbl>, GSM1061040 <dbl>
```

```r
# Obtenir dimensions de la matrius
dimensions_1 <- dim(matriu_expressio_genetica_1_final_proc)
# Calcular el número de mostres excloent la primera columna
num_mostres_1 <- dimensions_1[2] - 1  # Resta 1 per a excloure la columna GENE_SYMBOL
# Mostrar el nombre de files y columnas
cat("La matriu d'expressió genètica processada 1 té ", dimensions_1[1], "sondes genètiques i", num_mostr
```

```
## La matriu d'expressió genètica processada 1 té  23520 sondes genètiques i 91 mostres
```

```
dimensions_2 <- dim(matriu_expressio_genetica_2_final_proc)
# Calcular el número de mostres excloent la primera columna
num_mostres_2 <- dimensions_2[2] - 1  # Resta 1 per a excloure la columna GENE_SYMBOL
# Mostrar el nombre de files y columnas
cat("La matriu d'expressió genètica processada 2 té ", dimensions_2[1], "sondes genètiques i", num_mostr
```

```
## La matriu d'expressió genètica processada 2 té  23520 sondes genètiques i 111 mostres
```

Realitzat el pas anterior, **unifiquem les dues matrius** d'expressions genètiques dels dos conjunts de dades en una de sola a través del seu 'GENE_SYMBOL'.

```
cat("PROCESSAT PAS 3: Unió de les matrius resultants en una de sola")
```

```
## PROCESSAT PAS 3: Unió de les matrius resultants en una de sola
```

```
# Unim les dues matrius resultants de les expressions genètiques dels diferents conjunts de dades per a
# El nexe d'unió serà GENE_SYMBOL

matriu_expressio_genetica_final_unificada <- merge(
  matriu_expressio_genetica_1_final_proc,
  matriu_expressio_genetica_2_final_proc,
  by = c("GENE_SYMBOL"),
  all = TRUE
)
```

Elidim els 5 últims registres del conjunt de dades que no aporten cap tipus d'informació gènica (no es consideren símbols de gens vàlids) i visualitzem una mostra dels resultats:

```
cat("PROCESSAT PAS 4: Neteja de dades")
```

```
## PROCESSAT PAS 4: Neteja de dades
```

```
n <- nrow(matriu_expressio_genetica_final_unificada)
matriu_expressio_genetica_final_unificada <- matriu_expressio_genetica_final_unificada[1:(n - 5), ]

# Visualizem els resultats
head(matriu_expressio_genetica_final_unificada[1:10], n=5)
```

```
##   GENE_SYMBOL GSM4171495 GSM4171496 GSM4171497 GSM4171498 GSM4171499 GSM4171500
## 1        A1BG      0.313      0.319      0.248      0.329      0.239      0.344
## 2    A1BG-AS1      0.158      0.218      0.168      0.201      0.176      0.169
## 3        A1CF      0.113      0.116      0.120      0.114      0.134      0.109
## 4         A2M      0.394      0.332      0.363      0.438      0.396      0.385
## 5     A2M-AS1      0.159      0.224      0.132      0.216      0.282      0.142
##   GSM4171501 GSM4171502 GSM4171503
## 1      0.320      0.358      0.364
## 2      0.202      0.206      0.197
## 3      0.111      0.118      0.121
## 4      0.443      0.364      0.424
## 5      0.263      0.167      0.191
```

```
tail(matriu_expressio_genetica_final_unificada[1:10], n=5)
```

```
##        GENE_SYMBOL GSM4171495 GSM4171496 GSM4171497 GSM4171498 GSM4171499
## 23511        ZWINT      0.573      0.647      0.627      0.654      0.678
## 23512         ZXDA      0.310      0.271      0.272      0.269      0.416
## 23513 ZXDA ,  ZXDB      0.198      0.167      0.172      0.150      0.265
## 23514         ZXDB      0.237      0.222      0.223      0.198      0.259
```

```
## 23515         ZXDC      0.301      0.276      0.267      0.269      0.262
##       GSM4171500 GSM4171501 GSM4171502 GSM4171503
## 23511      0.684      0.459      0.558      0.445
## 23512      0.312      0.311      0.378      0.274
## 23513      0.169      0.182      0.234      0.165
## 23514      0.226      0.239      0.316      0.220
## 23515      0.256      0.288      0.303      0.254
```

```
# Obtenir dimensions de la matrius
dimensions_finals <- dim(matriu_expressio_genetica_final_unificada)
# Calcular el número de mostres excloent la primera columna
num_mostres_finals <- dimensions_finals[2] - 1  # Resta 1 per a excloure la columna GENE_SYMBOL
# Mostrar el nombre de files y columnas
cat("La matriu d'expressió genètica processada 1 té ", dimensions_finals[1], "sondes genètiques i", num
```

```
## La matriu d'expressió genètica processada 1 té  23515 sondes genètiques i 202 mostres
```

Finalment, **s'unifiquen també els dos conjunts de dades parcials** generats en les seccions 1.1 i 1.2 que contenien informació de les pacients, els tumors i els graus histològics d'aquests. D'aquesta manera es tenen els dos conjunts de dades unificats tant per la part informativa com per les matrius d'expressions genètiques.

```
## UNIÓ DATASETS parcials sense informació genètica
cat("PROCESSAT PAS 5: Unió dels datasets 1 i 2 parcials (que no incluen informació genètica dels pacien
```

```
## PROCESSAT PAS 5: Unió dels datasets 1 i 2 parcials (que no incluen informació genètica dels pacients
```

```
dataset_final <- rbind(dataset1_parcial, dataset2_parcial)

head(dataset_final, n=5)
```

```
##                       TITLE GEO_ACCESSION COUNTRY TYPE                NAME
## 1 BC_Patient1007_Genearray1    GSM4171495 Germany  RNA Pretreatment biopsy
## 2 BC_Patient1008_Genearray1    GSM4171496 Germany  RNA Pretreatment biopsy
## 3 BC_Patient1010_Genearray1    GSM4171497 Germany  RNA Pretreatment biopsy
## 4 BC_Patient1011_Genearray1    GSM4171498 Germany  RNA Pretreatment biopsy
## 5 BC_Patient1013_Genearray1    GSM4171499 Germany  RNA Pretreatment biopsy
##   HISTOLOGICAL_GRADE
## 1                  2
## 2                  2
## 3                  3
## 4                  3
## 5                  2
```

```
tail(dataset_final, n=5)
```

```
##           TITLE GEO_ACCESSION COUNTRY TYPE
## 198 PK-09-27864    GSM1061138     USA  RNA
## 199 PK-09-27865    GSM1061139     USA  RNA
## 200 PK-09-27866    GSM1061140     USA  RNA
## 201 PK-09-27869    GSM1061141     USA  RNA
## 202 PK-09-27870    GSM1061142     USA  RNA
##                                                NAME HISTOLOGICAL_GRADE
## 198 primary breast tumor - fresh surgical samples                   3
## 199 primary breast tumor - fresh surgical samples                   1
## 200 primary breast tumor - fresh surgical samples                   3
## 201 primary breast tumor - fresh surgical samples                   1
## 202 primary breast tumor - fresh surgical samples                   2
```

5 2. Selecció de característiques

Després de tot el procés de generació de la matriu d'expressions genètiques final unificada i normalitzada, cal sotmetre tot el conjunt de files amb informació genètica a un **procés de selecció per a que triï aquells gens o conjunt de gens que tenen més rellevància**. Aquest procés serà útil per reduir la dimensionalitat de les variables a analitzar pels posterors models de Machine-Learning i reduir soroll, costos temporals i computacionals

Com que encara no es disposa d'una variable objectiu, una manera de seleccionar les característiques (gens) més importants és a partir de la **desviació estàndard** de cada una de les files a partir dels valors de totes les mostres.

La desviació estàndard és un bon **indicador de variabilitat** de les mostres: les files (gens) amb valors molt baixos de desviació estàndard denotaran una variabilitat molt baixa i per tant, la majoria dels seus valors seran similars i sense diferències significatives (propers a la mitjana aritmètica). El que ens interessa es veure quins gens són capaços de classificar i per tant, els que variïn poc no tindran aquesta capacitat ja que seran iguals per a tots els graus histològics. Pel contrari, els que tinguin valors més alts o més baixos són els que aportaran més informació.

```
## SELECCIÓ I REDUCCIÓ DE CARACTERÍSTIQUES ##

cat("SELECCIÓ DE CARACTERÍSTIQUES a partir de la desviació estàndard\n")
```

```
## SELECCIÓ DE CARACTERÍSTIQUES a partir de la desviació estàndard
```

```
# 1) En primer lloc, es calcularà la mitjana aritmètica dels valors de totes les mostres per a cada una

cat("Càlculs desviació estàndard de cada fila:\n")
```

```
## Càlculs desviació estàndard de cada fila:
```

```
# Capturem en una variable la matriu d'expressió genètica sense la columna del GENE_SYMBOL
gene_expression_matrix <- matriu_expressio_genetica_final_unificada[, -1]
# Realitzem el càlcul de la mitjana aritmètica de cada fila a partir dels valors de les mostres i ho em
matriu_expressio_genetica_final_unificada$SD <- apply(gene_expression_matrix, 1, sd, na.rm = TRUE)
head(matriu_expressio_genetica_final_unificada[1:10], n=5)
```

```
##   GENE_SYMBOL GSM4171495 GSM4171496 GSM4171497 GSM4171498 GSM4171499 GSM4171500
## 1        A1BG      0.313      0.319      0.248      0.329      0.239      0.344
## 2    A1BG-AS1      0.158      0.218      0.168      0.201      0.176      0.169
## 3        A1CF      0.113      0.116      0.120      0.114      0.134      0.109
## 4         A2M      0.394      0.332      0.363      0.438      0.396      0.385
## 5     A2M-AS1      0.159      0.224      0.132      0.216      0.282      0.142
##   GSM4171501 GSM4171502 GSM4171503
## 1      0.320      0.358      0.364
## 2      0.202      0.206      0.197
## 3      0.111      0.118      0.121
## 4      0.443      0.364      0.424
## 5      0.263      0.167      0.191
```

```
tail(matriu_expressio_genetica_final_unificada[1:10], n=5)
```

```
##        GENE_SYMBOL GSM4171495 GSM4171496 GSM4171497 GSM4171498 GSM4171499
## 23511        ZWINT      0.573      0.647      0.627      0.654      0.678
## 23512         ZXDA      0.310      0.271      0.272      0.269      0.416
## 23513 ZXDA ,  ZXDB      0.198      0.167      0.172      0.150      0.265
## 23514         ZXDB      0.237      0.222      0.223      0.198      0.259
## 23515         ZXDC      0.301      0.276      0.267      0.269      0.262
```

```
##       GSM4171500 GSM4171501 GSM4171502 GSM4171503
## 23511      0.684      0.459      0.558      0.445
## 23512      0.312      0.311      0.378      0.274
## 23513      0.169      0.182      0.234      0.165
## 23514      0.226      0.239      0.316      0.220
## 23515      0.256      0.288      0.303      0.254
```

```
# 2) Seguidament, s'estabilarà el llindar que servirà de filtre per determinar quines files (gens) tener

# Calculem la mitjana de totes les desviacions estàndard obtingudes per tenir una referència del promig
total_sd_average <- mean(matriu_expressio_genetica_final_unificada$SD, na.rm = TRUE)
# Establim el llindar com la mitjana aritmètica global de les desviacions estàndard
llindar <- total_sd_average
cat("El llindar del filtrat és ", llindar, "\n")
```

```
## El llindar del filtrat és  0.1342392
```

```
cat("Aplicació de filtre: \n")
```

```
## Aplicació de filtre:
```

```
# 3) Finalment, apliquem el filtre a partir del llindar ontingut. D'aquesta manera, totes les files (ge
matriu_expressio_genetica_final_unificada_filtrada <- matriu_expressio_genetica_final_unificada[matriu_

cat("Nombre de gens originalment:", nrow(matriu_expressio_genetica_final_unificada), "\n")
```

```
## Nombre de gens originalment: 23515
```

```
cat("Nombre de gens seleccionats:", nrow(matriu_expressio_genetica_final_unificada_filtrada), "\n")
```

```
## Nombre de gens seleccionats: 10640
```

```
cat("La matriu d'expressió genètica s'ha reduït de" ,nrow(matriu_expressio_genetica_final_unificada), "
```

```
## La matriu d'expressió genètica s'ha reduït de 23515  a  10640  característiques.
```

6 3. Generació i muntatge del dataset final

Un cop aplicada la selecció de característiques anterior, ens queda preparar el conjunt de dades per a que sigui compatible amb els models de machine-learning futurs. Per tant, cal realitzar les següents operacions:

- 1) Transposició de la matriu d'expressions genètiques: El conjunt de dades final haurà de tenir les dades gèniques en columnes i no en files, és a dir, cada fila representarà la informació d'un pacient diferent, els valors de les seves mostres per a cada un dels gens o conjunt de gens filtrats i finalment, el grau histològic del tumor. Per tant, per tal de poder tenir les dades de forma que cada fila representi un pacient caldrà transposar la matriu final d'expressions genètiques i capgirar les files per les columnes (els gens ara seran columnes i les mostres dels pacients seran files).
- 2) Unió de les dades de les expressions genètiques amb la informació dels pacients i el grau histològic del tumor. Això suposoarà la generació i muntatge del conjunt de dades supervisat final que podrà ser utilitzat en totes les seccions posteriors.

6.1 3.1 Operacions de transposició i unificació final

A continuació, realtizem tots els processos necessaris per arribar a la transposició de la matriu d'expressions genètiques del conjunt de dades final, tal com s'ha especificat en la secció anterior:

```
cat("Transposició matriu d'expressions genètiques\n")
```

Transposició matriu d'expressions genètiques

```
# Eliminem la columna temporal 'Average' necessària en el punt anterior
matriu_expressio_genetica_final_unificada_filtrada <- matriu_expressio_genetica_final_unificada_filtrada
# Canviar el nom de la columna GENE_SYMBOL i posar-los com a nomsn (identificadors) de columna GENE_SYMB
rownames(matriu_expressio_genetica_final_unificada_filtrada) <- matriu_expressio_genetica_final_unificad
# Eliminació de la columna 'GENE_SYMBOL' ja que ara hi haurà tantes columnes com diferents GENE_SYMBOLS
matriu_expressio_genetica_final_unificada_filtrada <- matriu_expressio_genetica_final_unificada_filtrada
# Transpoció de la matriu per a que les files siguin mostres de pacients i les columnes siguin els gens
matriu_expressio_genetica_final_unificada_filtrada_t <- t(matriu_expressio_genetica_final_unificada_filt
# Conversió a dataframe
matriu_expressio_genetica_final_unificada_filtrada_t <- as.data.frame(matriu_expressio_genetica_final_un
# Finalment, creem la columna 'GEO_ACCESSION' que es correspon amb l'identificador de les mostres dels p
matriu_expressio_genetica_final_unificada_filtrada_t$GEO_ACCESSION <- rownames(matriu_expressio_genetica
```

Seguidament, es realitza el procés d'unificació final que comporta la generació i muntatge del conjunt de dades final definitiu i l'eliminació dels registres que tenen el camp corresponent al grau histològic desinformat:

```
cat("Unió de les expressions genètiques amb la informació dels pacients i el grau histològic del tumor\n
```

Unió de les expressions genètiques amb la informació dels pacients i el grau histològic del tumor

```
# 2) Unió de la matriu d'expressions genètiques unificada i transposada, amb el dataset que conté la in
# del GEO_ACCESSION com a variable relacionant.
dataset_final_unificat <- merge(dataset_final, matriu_expressio_genetica_final_unificada_filtrada_t, by

# Eliminem els registres que no tenen la informació mínima obligatòria requerida: grau histològic
dataset_final_unificat <- dataset_final_unificat[!is.na(dataset_final_unificat$HISTOLOGICAL_GRADE), ]

head(dataset_final_unificat[1:10],n=5)
```

```
##   GEO_ACCESSION       TITLE COUNTRY TYPE
## 1    GSM1061032 PK-09-26449     USA  RNA
## 2    GSM1061033 PK-09-26710     USA  RNA
## 3    GSM1061034 PK-09-26716     USA  RNA
## 4    GSM1061035 PK-09-26718     USA  RNA
## 5    GSM1061036 PK-09-26722     USA  RNA
##                                          NAME HISTOLOGICAL_GRADE  A1BG   A2M
## 1 primary breast tumor - fresh surgical samples                 2 0.002 0.024
## 2 primary breast tumor - fresh surgical samples                 2 0.000 0.007
## 3 primary breast tumor - fresh surgical samples                 2 0.001 0.019
## 4 primary breast tumor - fresh surgical samples                 2 0.001 0.037
## 5 primary breast tumor - fresh surgical samples                 3 0.001 0.028
##    AAAS  AACS
## 1 0.000 0.005
## 2 0.000 0.008
## 3 0.001 0.005
## 4 0.000 0.003
## 5 0.000 0.004
```

```
tail(dataset_final_unificat[1:10],n=5)
```

```
##     GEO_ACCESSION                      TITLE COUNTRY TYPE                NAME
## 197    GSM4171580 BC_Patient4005_Genearray5 Germany  RNA Pretreatment biopsy
## 199    GSM4171582 BC_Patient6001_Genearray5 Germany  RNA Pretreatment biopsy
## 200    GSM4171583 BC_Patient4012_Genearray5 Germany  RNA Pretreatment biopsy
## 201    GSM4171584 BC_Patient4003_Genearray5 Germany  RNA Pretreatment biopsy
```

```
## 202    GSM4171585 BC_Patient9021_Genearray5 Germany  RNA Pretreatment biopsy
##     HISTOLOGICAL_GRADE  A1BG   A2M  AAAS  AACS
## 197                  2 0.248 0.458 0.360 0.478
## 199                  2 0.291 0.462 0.355 0.503
## 200                  2 0.285 0.417 0.359 0.490
## 201                  3 0.299 0.466 0.361 0.551
## 202                  2 0.326 0.468 0.355 0.450
```

```
# Obtenir dimensions finals del dataset
dimensions_finals_dataset <- dim(dataset_final_unificat)
# Mostrar el nombre de files y columnas
cat("El conjunt de dades final i definitiu té ", dimensions_finals_dataset[1], "files (pacients) i", di
```

```
## El conjunt de dades final i definitiu té  200 files (pacients) i 10646 columnes
```

```
# Finalment, escribim els resultats finals i definitius del conjunt de dades en un .csv
write_to_csv(dataset_final_unificat, ruta_output, "dataset_final.csv")
cat("Conjunt de dades final definitiu guardat i generat com a .csv a la ruta: ", paste0(ruta_output,"da
```

```
## Conjunt de dades final definitiu guardat i generat com a .csv a la ruta:  data/output/dataset_final.
```

6.2 3.2 Estadístiques bàsiques

Amb el conjunt de dades final generat, es mostren una sèrie d'estadístiques bàsiques per a mostrar el seu contingut, les dades i la seva estructura per ajudar en el seu anàlisi.

A més a més, es mostren també estadístiques i gràfiques del repartiment dels registres (pacients) segons el grau histològic del tumor (variable objectiu). D'aquesta manera es podrà observar el balanceig de les dades per a les 3 classes possibles.

```
dataset <- dataset_final_unificat

# Estadístiques descriptives i resums de columnes
# Per sintetitzar, excloem totes les columnes referents a gens o conjunt de gens
cat("Estadístiques \n")
```

```
## Estadístiques
```

```
summary(dataset[1:7])
```

```
##  GEO_ACCESSION         TITLE             COUNTRY             TYPE
##  Length:200         Length:200         Length:200         Length:200
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      NAME            HISTOLOGICAL_GRADE      A1BG
##  Length:200         Min.   :1.00       Min.   :0.0000
##  Class :character   1st Qu.:2.00       1st Qu.:0.0010
##  Mode  :character   Median :2.00       Median :0.0020
##                     Mean   :2.28       Mean   :0.1407
##                     3rd Qu.:3.00       3rd Qu.:0.2993
##                     Max.   :3.00       Max.   :0.4760
```

```
cat("Estructura interna: \n")
```

```
## Estructura interna:
```

```r
str(dataset[1:7])
```

```
## 'data.frame':    200 obs. of  7 variables:
##  $ GEO_ACCESSION     : chr  "GSM1061032" "GSM1061033" "GSM1061034" "GSM1061035" ...
##  $ TITLE             : chr  "PK-09-26449" "PK-09-26710" "PK-09-26716" "PK-09-26718" ...
##  $ COUNTRY           : chr  "USA" "USA" "USA" "USA" ...
##  $ TYPE              : chr  "RNA" "RNA" "RNA" "RNA" ...
##  $ NAME              : chr  "primary breast tumor - fresh surgical samples" "primary breast tumor - f
##  $ HISTOLOGICAL_GRADE: num  2 2 2 2 3 1 2 3 2 2 ...
##  $ A1BG              : num  0.002 0 0.001 0.001 0.001 0.001 0.001 0.001 0.002 0.002 ...
```

```r
cat(" ESTADÍSTIQUES PER GRAU HISTOLÒGIC, VARIABLE OBJECTIU: \n")
```

```
##  ESTADÍSTIQUES PER GRAU HISTOLÒGIC, VARIABLE OBJECTIU:
```

```r
# Pas 1: Comptar freqüències de cada categoria en la columna histologic_grade
counts <- table(dataset$HISTOLOGICAL_GRADE)
cat("Recompte de registres per a cada grau histològic: \n")
```

```
## Recompte de registres per a cada grau histològic:
```

```r
# Convertim els resultats a un data frame
counts_df <- as.data.frame(counts)
# Renombrament columnes
colnames(counts_df) <- c("GRAU HISTOLOGIC", "NUM. REGISTRES")
# Visualització taula
print(counts_df)
```

```
##   GRAU HISTOLOGIC NUM. REGISTRES
## 1               1             24
## 2               2             96
## 3               3             80
```

```r
# Pas 2: Càlcul de percentatges
total_rows = nrow(dataset)
percentatges <- (counts / total_rows) * 100

cat("Percentatges de representació de registres per a cada grau histològic: \n")
```

```
## Percentatges de representació de registres per a cada grau histològic:
```

```r
# Convertim els resultats a un data frame
percentatges_df <- as.data.frame(percentatges)
# Renombrament columnes
colnames(percentatges_df) <- c("GRAU HISTOLOGIC", "% REGISTRES")
# Visualització taula
print(percentatges_df)
```

```
##   GRAU HISTOLOGIC % REGISTRES
## 1               1          12
## 2               2          48
## 3               3          40
```

```r
# Loop que recorre les files de la taula de percentatges
for(i in 1:nrow(percentatges_df)) {
  cat("Les pacients amb tumor de mama de grau histològic ",
      percentatges_df$`GRAU HISTOLOGIC`[i],
      " representen el ",
```

```
      percentatges_df$`% REGISTRES`[i],
      "% del total\n")
}
```

```
## Les pacients amb tumor de mama de grau histològic  1  representen el  12 % del total
## Les pacients amb tumor de mama de grau histològic  2  representen el  48 % del total
## Les pacients amb tumor de mama de grau histològic  3  representen el  40 % del total
```

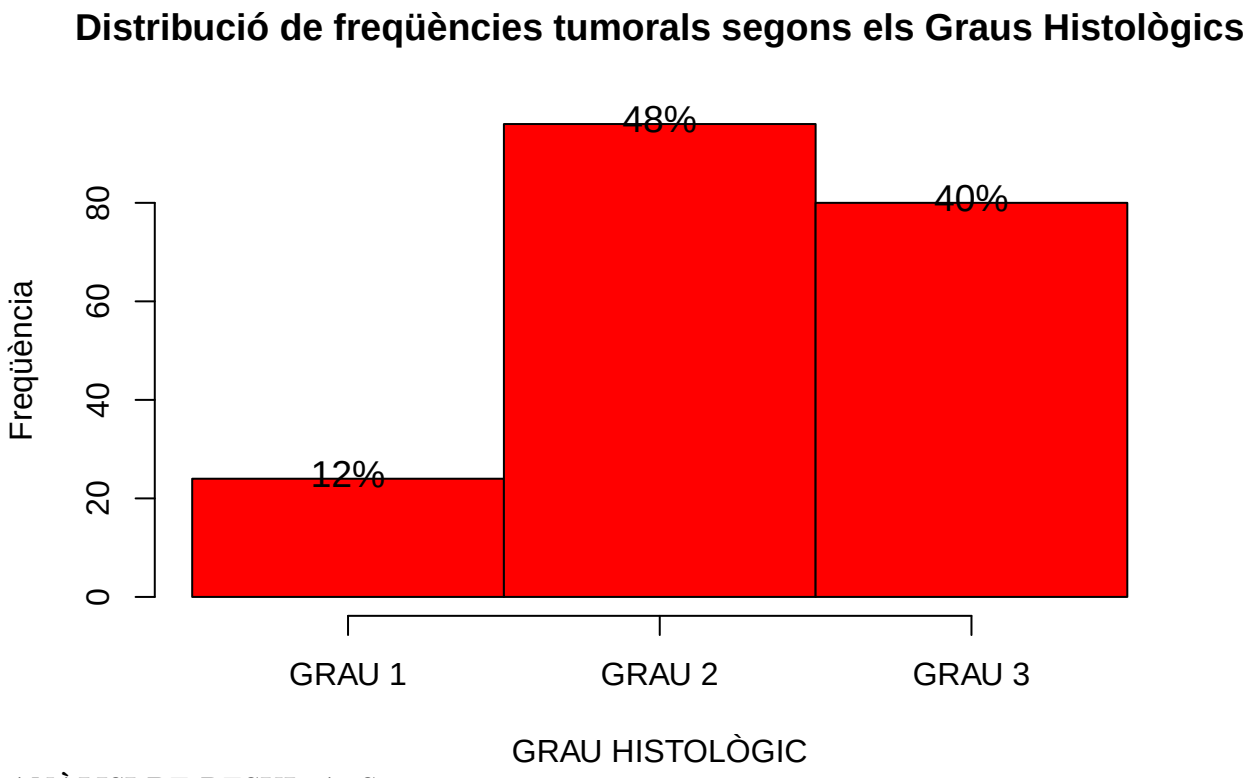
```
# Pas 3: Mostra d'histogrames segons variable objectiu: GRAU HISTOLOGIC
cat("Generació d'histograma que mostra les freqüències de pacients de tumors de mama en cada un dels tr
```

```
## Generació d'histograma que mostra les freqüències de pacients de tumors de mama en cada un dels tres
```

```
# Creació histograma de freqüencies de pacients segons els diferents graus histològics
hist_obj  <- hist(dataset$HISTOLOGICAL_GRADE,
                  breaks = seq(0.5, 3.5, by = 1),  # Punts de tall
                  main = "Distribució de freqüències tumorals segons els Graus Histològics",
                  xlab = "GRAU HISTOLÒGIC",
                  ylab = "Freqüència",
                  col = "red",
                  border = "black",
                  xaxt = "n")  # Evita los valors per defecte en eix X

# Inclusió de labels en l'eix X
axis(1, at = 1:3, labels = c("GRAU 1", "GRAU 2", "GRAU 3"))

# Inclusió informació percentatges en les barres
text(x = hist_obj$mids,  # Posicions X
     y = hist_obj$counts + 1,
     labels = paste0(round(percentatges, 1), "%"),  # Format de percentatges
     cex = 1.2,  # Mida text
     col = "black")  # Color text
```



ANÀLISI DE RESULTATS:

Els casos de tumors de grau histològic 2 són els més freqüents en el nostre conjunt de dades destacant amb un 48% i 96 registres.El segueix de prop els tumors de grau histològic 3 amb un 40% (80 registres) i, en últim lloc, i amb una diferència considerablement pronunciada, tenim els casos de tumors de grau histològic 1 amb un 12% i 24 registres.

S'aprecia un **desbalanceig considerable del nombre de tumors de grau histològic 1** (12%) respecte als de grau 2 (48%) i 3 (40%). Aquest fet pot afectar a l'hora de definir amb precisió els gens o conjunts de gens més involucrats, ja que, a excepció de la resta, es posseeixen pocs casos per provar. Pel que respecta a la diferència entre el nombre d'elements de grau histològic 2 amb el d'elements de grau histològic 3, aquesta no és considerable ni molt significativa. Per tant, en aquests dos casos no hi ha un desbalanceig molt pronunciat i els resultats són bastant equitatius.El nombre de casos és més nombrós i, en conseqüència, els resultats poden ser més precisos i estar més ben ajustats.

Tot i els resultats, podem obviar el desbalanceig ja que els graus histològics que més interessa definir amb precisió són el 2 i el 3 que són més crítics que els de grau 1.

6.3 3.3 Resum de característiques

Finalment, es mostra una taula resum de les característiques principals del conjunt de dades final generat per tal de tenir clares les idees abans d'aplicar totes les tècniques de machine-learning supervisat i de regressió:

```
tabla_info <- data.frame(
  "CARACTERÍSTICA" = c("TIPUS CONJUNT DE DADES", "TIPUS D'ALGORISME", "VARIABLE OBJECTIU", "TIPUS VARIA
                       "VARIABLES INDEPENDENTS", "TIPUS VARIABLES INDEPENDENTS"),
  "DESCRIPCIÓ" = c("Supervisat", "Regressió", "HISTOLOGICAL_GRADE",
                   "Quantitativa categòrica ordinal de 3 classes (1,2 i 3)",
                   "Tot el conjunt de columnes que representen gens o conjunts de gens",
                   "Quantitatives contínues amb valors normalitzats compresos entre 0 i 1")
)
```

```r
# Mostrem la taula
#kable(tabla_info, col.names = c("CARACTERÍSTICA", "DESCRIPCIÓ"), align = "l")

tabla_info %>%
  kable(col.names = c("CARACTERÍSTICA", "DESCRIPCIÓ"), align = "l") %>%
  kable_styling(full_width = FALSE, position = "center", latex_options = "HOLD_position") %>%
  row_spec(0, bold = TRUE, color = "white", background = "#40E0D0") %>%  # Capçalera amb estil
  row_spec(1:nrow(tabla_info), hline_after = TRUE)  # Línies entre files
```

CARACTERÍSTICA	DESCRIPCIÓ
TIPUS CONJUNT DE DADES	Supervisat
TIPUS D'ALGORISME	Regressió
VARIABLE OBJECTIU	HISTOLOGICAL_GRADE
TIPUS VARIABLE OBJECTIU	Quantitativa categòrica ordinal de 3 classes (1,2 i 3)
VARIABLES INDEPENDENTS	Tot el conjunt de columnes que representen gens o conjunts de gens
TIPUS VARIABLES INDEPENDENTS	Quantitatives contínues amb valors normalitzats compresos entre 0 i 1

7 4. Tècniques de machine-learning supervisat de regressió

7.1 4.0 Preparació de dades

Com a pas previ a l'aplicació de les tècniques d'aprenentatge automàtic supervisat de regressió, cal **preprar les dades** per tal d'adaptar-les als requeriments exigits.

Per a poder treballar sobre el conjunt de dades supervisat que tenim, els models d'aprenentatge automàtic de regressió necessiten:

- **Variable objectiu**: Es correspon amb el grau histològic del tumor, que és sobre el que volem determinar els coeficients o índexs d'importàncies genètiques. Es tracta d'una variable numèrica ordinal de tres categories: 1, 2 i 3. Per tal de poder treballar bé sobre aquesta variable, i donat que el que ens interessa són tres resultats d'importàncies diferents per grau histològic (rellevàncies de gens o conjunts de gens en grau histològic 1, en el 2 i en el 3) el que es fa es **afegir tres columnes noves amb contingut binari que serveixin d'indicadors de pertinença del tumor de la pacient a un dels graus histològics**. És a dir, es tindran tres indicadors de pertinença per cada un dels graus histològics:
  - **is_grade_1**: Si el pacient té un tumor que està classificat amb grau histològic 1, el valor serà 1 i el dels altres indicadors, 0
  - **is_grade_2**: Si el pacient té un tumor que està classificat amb grau histològic 2, el valor serà 1 i el dels altres indicadors, 0
  - **is_grade_3**: Si el pacient té un tumor que està classificat amb grau histològic 3, el valor serà 1 i el dels altres indicadors, 0

  D'aquesta manera es podran aplicar els models d'aprenentatge automàtic per a cada un dels diferents graus histològics com a variable objectiu. Aquesta, serà de tipus binari (on l'1 indica la pertinença i el 0 la no pertinença).
- **Variables independents (X)** : Conjunt de columnes informatives de les expressions genètiques de les pacients amb tumor de mama. Aquestes característiques es corresponen a les que hi ha a partir de la

columna 7 en amunt (les que tenen un nom de gen o conjunt de gens). Els seus valors estan escalats entre 0 i 1 i són quantitatius continus.

```
# Variable objectiu--> grau histològic (categòrica ordinal)
# Variables independents --> conjunt de 10.600 columnes referents a gens (valors continus normalitzats

# Crear variables indicadores binàries per a cada registre sobre la seva pertinença a un grau histològi
# D'aquesta manera es podran aplicar els algorismes de machine-learning de forma independent per
# a cada una de les possibles categories de la variable objectiu (pertinença a grau 1, grau 2 o grau 3)
dataset$is_grade_1 <- ifelse(dataset$HISTOLOGICAL_GRADE == 1, 1, 0)
dataset$is_grade_2 <- ifelse(dataset$HISTOLOGICAL_GRADE == 2, 1, 0)
dataset$is_grade_3 <- ifelse(dataset$HISTOLOGICAL_GRADE == 3, 1, 0)

# Capturem les variables independents que són les corresponents als gens o conjunts de gens. Aquestes va
# a la N (10647) però n'hem d'excloure les 3 últimes variables temporals afegides en el pas anterior d'
# Ho convertim en matriu per a que sigui compatible amb els algorismes.

X <- as.matrix(dataset[, 7:(ncol(dataset) - 3)])
```

Per les característiques del conjunt de dades que tenim, l'objectiu que es persegueix i la quantitat immensa de variables independents que es posseeix, les tècniques d'aprenentatge automàtic supervisat de regressió que més s'adequen són:

- Tècniques basades en arbres de decisió i/o descens del gradient: **Random Forest, XGBoost o LightGBM**
- Tècniques basades en regularitzacions: **Ridge Regression, Lasso Regression o Elastic Net**

7.2 4.1 Tècniques basades en arbres de decisió i/o descens del gradient

7.2.1 4.1.1 Random Forest

7.2.2 4.1.2 XGBoost

DEFINICIÓ:

XGBoost són les sigles de "Extreme Gradient Boosting". Es tracta d'un algorisme d'aprenentatge automàtic que permet manegar grans quantitats de conjunts de dades i té una elevada capacitat per a obtenir un bon rendiment en tasques de classificació i **regressió** sobre conjunts de dades **supervisats**. Porta integrat el processament en paral · lel.

Està basat en un **conjunt d'arbres de decisió potenciats per gradients**. Els arbres de decisió es creen seqüencialment i les variables independents s'incorporen en ells a través de pesos. Utilitza l'algorisme d'augment de gradient on cada predictor corregeix el seu predecessor.

AVANTATGES

- **Rendiment**: Té una producció de resultats d'alta qualitat en diverses tasques.
- **Escalabilitat**: És adequat per a grans conjunts de dades ja que el seu disseny el dota d'eficiència i escalabilitat en el seu entrenament.
- **Personalització**: Conté un ampli ventall d'hiperparàmetres que es poden ajustar per optimitzar els resultats
- Té suport integrat per a **manegar valors faltants** (resistent a dades que contenen valors buits o nuls)
- Permet obtenir les **importàncies de les característiques** que és, en gran mesura, l'objectiu que tenim per a cada valor del grau histològic.

INCONVENINENTS

- **Elevada complexitat computacional** derivada en un ús intensiu dels recursos computacionals especialment en l'entrenament de grans quantitats de dades.

- Propensió al **sobreentrenament** o sobreajustament si hi ha un conjunt de dades petit
- **Dificultat en l'ajustament dels seus hiperparàmetres** ja que el seu conjunt d'hiperparàmetres és molt ampli i pot resultat molt costós temporalment trobar-ne els òptims.
- **Elevat consum de memòria** si es treballa amb grans conjunts de dades.

CERCA HIPERPARÈMTRES

XGBoost té un gran conjunt d'hiperparàmetres que es poden ajustar i combinar de diferents maneres per tal de trobar el resultat òptim. El resultat òptim serà el que **maximitzi l'àrea sota la corba AUC** que és una mètrica de rendiment indicadora de la idoneïtat del nostre model. Per tant, el primer pas abans d'aplicar el model sobre el conjunt de dades, serà la cerca dels valors concrets combinats d'hiperparàmetres que donen més bon resultat en entrenar el model.

Els hiperparàmetres més rellevants en XGBoost són els següents: - **max_depth**: Profunditat màxima dels arbres de decisió utilitzats en l'entrenament. Una profunditat elevada pot retornar millors resultats però a la vegada pot ser pronpensa a provocar sobreentrenament. - **eta**: Taxa d'aprenentatge del model. Com més gran és el seu valor, el cost temporal i computacional es redueix ja que arriba abans a un "millor model". Pel contrari, per a valors petits es tardarà més en arribar a aquest millor model i el cost temporal i computacional serà més elevat.
- **subsample**: Percentatge de dades (files) utilitzades per l'arbre en cada iteració - **colsample_bytree**: Proporció de submostres de columnes al construir cada arbre - **objective**: Tipus de tasca de classificació que es realitzarà, en el nostre cas, serà de tipus binària pels indicadors de pertinença a un grau histològic determinat. - **nrounds**: Nombre d'iteracions a realitzar abans d'acabar amb el procés de cerca. A més iteracions, més bons solen ser els resultats però també incrementa el cost temporal.

A continuació creem la funció que s'encarregarà de la cerca dels hiperparàmetres òptims en l'algorisme XGBoost i maximitzarà la mètrica de rendiment AUC:

```
# Funció genèrica per a trobar els hiperpàrmatres òptims (aquells que maximitzen ) emprant la validació
# i la graella d'opcions de param_grid
# Paràmetres:
# - X :conjunt de variables independents)
# - y :variable objectiu binària: is_grade_n)
# - param_grid: matriu inicial de possibles hiperparàmetres amb possibles valors
hyperparams_xgboost_search <- function(X, y, param_grid) {
  cat("Cerca d'hiperparàmetres òptims: \n")

  # Inicialitzar les variables per guardar el millor AUC i els seus hiperparàmetres
  best_auc <- -Inf
  best_params <- NULL

  # Barra de progrés per a tenir una noció del temps
  pb <- progress_bar$new(
    format = "Avaluant hiperparàmetres :current/:total (:percent) Temps restant: \n",
    total = nrow(param_grid),
    clear = FALSE,
    width = 60
  )

  # Bucle sobre la graella
  for (i in 1:nrow(param_grid)) {

    # Actualitzar la barra de progrés
    pb$tick()

    params <- list(
```

```
      objective = "binary:logistic",
      max_depth = param_grid$max_depth[i],
      eta = param_grid$eta[i],
      subsample = param_grid$subsample[i],
      colsample_bytree = param_grid$colsample_bytree[i],
      nrounds = param_grid$nrounds[i]
    )

    # Imprimir els hiperparàmetres actuals
    cat("\n Avaluant hiperparàmetres en XGBoost ", y ," : ",
        paste(names(params), unlist(params), sep = "=", collapse = ", "), "\n")


    # Validació creuada emprant .cv
    cv_results <- xgb.cv(
      params = params,
      data = xgb.DMatrix(X, label = dataset[[y]]),
      nfold = 3,
      metrics = "auc", # corba error
      verbose = 0
    )

    # Obtenir el millor AUC per aquests hiperparàmetres
    current_auc <- max(cv_results$evaluation_log$test_auc_mean)

    # Si el AUC actual és millor que el millor fins ara, en el quedem. Així capturem el que maximitza b
    if (current_auc > best_auc) {
      best_auc <- current_auc
      best_params <- params
    }

  }

  # Retornar els millors hiperparàmetres
  return(list(best_params = best_params, best_auc = best_auc))

}
```

7.2.3 4.1.3 LightGBM

7.3 4.2 Tècniques basades en regularitzacions

7.3.1 4.2.1 Ridge Regression o L2

- cerca hiperparàmetres
- càlcul coeficients d'importàncies
- resultats
  - ranking
  - diagrames de barres horitzontals
- conclusions

7.3.2 4.2.2 Lasso Regression o L1

- cerca hiperparàmetres
- càlcul coeficients d'importàncies

- resultats
  - ranking
  - diagrames de barres horitzontals
- conclusions

7.3.3 4.2.3 Elastic Net

- cerca hiperparàmetres
- càlcul coeficients d'importàncies
- resultats
  - ranking
  - diagrames de barres horitzontals
- conclusions

8 Conclusions