

# Proyecto\_RNAseq

## Análisis de Expresión Diferencial

Karla Ximena González Platas

2025-02-05

## Contents

|  |    |
|--|----|
| Introducción . . . . .                     | 1  |
| Pregunta de investigación . . . . .        | 1  |
| Hipótesis . . . . .                        | 2  |
| Instalación y carga de paquetes . . . . .  | 2  |
| Selección de Proyecto . . . . .            | 2  |
| Preparación de los datos . . . . .         | 6  |
| Filtrar genes de baja expresión . . . . .  | 8  |
| Normalización de los datos . . . . .       | 10 |
| Determinar el modelo estadístico . . . . . | 11 |
| Visualizar matriz . . . . .                | 13 |
| Expresión diferencial . . . . .            | 14 |
| Visualizar genes DE . . . . .              | 19 |
| Conclusión . . . . .                       | 22 |

## Introducción

El conjunto de datos analizado en este proyecto se deriva del estudio de Muluhngwi y Klinge (2021), que explora las interacciones regulatorias entre los miembros de la familia miR-29 y los lncRNAs (ARN largos no codificantes) en el contexto de la resistencia a la terapia endocrina. En este estudio, se aplicó el análisis de RNA-seq para investigar la expresión de lncRNAs regulados por miR-29b-1-3p y miR-29a-3p en células de cáncer de mama sensibles a la terapia endocrina (MCF-7) y resistentes a la terapia endocrina (LCC9). Estas líneas celulares fueron empleadas para estudiar los efectos de la regulación de lncRNAs por los miR-29b-1-3p y miR-29a-3p en la resistencia endocrina del cáncer de mama. Para ello, se realizaron transfecciones en las células MCF-7 y LCC9 con pre-miR-29b-1-3p y pre-miR-29a-3p para evaluar los efectos de su sobreexpresión en la proliferación celular y en la regulación de lncRNAs. Mientras tanto, el uso de anti-miR-29 y un control negativo permitió analizar la inhibición de estos miRNAs.

## Pregunta de investigación

¿Cuáles son los genes diferencialmente expresados entre las células de cáncer de mama sensibles (MCF-7) y resistentes (LCC9) a la terapia endocrina, y cómo se ven afectados por la sobreexpresión e inhibición de miR-29a y miR-29b-1?

## Hipótesis

La sobreexpresión de miR-29a y miR-29b-1 en las células de cáncer de mama sensibles (MCF-7) y resistentes (LCC9) a la terapia endocrina altera la expresión de genes involucrados en la resistencia a los tratamientos.

## Instalación y carga de paquetes

```
# Instalar BiocManager si no está instalado
#if (!requireNamespace("BiocManager", quietly = TRUE)) {
#   install.packages("BiocManager")
#}

# Instalar paquetes de Bioconductor
#BiocManager::install(
#  c(
#    "edgeR",
#    "ExploreModelMatrix",
#    "limma",
#    "recount3",
#    "SummarizedExperiment",
#    "GenomicRanges"
#  )
#)

# Instalar paquetes de CRAN
#install.packages(c(
#  "pheatmap",
#  "patchwork",
#  "RColorBrewer",
#  "cowplot"
#))

## Cargar los paquetes
library("recount3")
library("SummarizedExperiment")
library("GenomicRanges")
library("limma")
library("edgeR")
library("ExploreModelMatrix")
library("cowplot")
library("RColorBrewer")
library("pheatmap")
```

## Selección de Proyecto

```
# Obtener la lista de proyectos disponibles
human_projects <- available_projects()

## 2025-02-07 23:35:41.491777 caching file sra.recount_project.MD.gz.
```

```
## 2025-02-07 23:35:43.297466 caching file gtex.recount_project.MD.gz.
```

```
## 2025-02-07 23:35:43.951757 caching file tcga.recount_project.MD.gz.
```

```
# Ver los proyectos disponibles  
dim(human_projects)
```

```
## [1] 8742      6
```

```
# Esto nos indica cuántos proyectos están disponibles (número de filas)  
# y cuántas columnas de información se proporcionan para cada proyecto.
```

```
# Mostrar las primeras filas para inspeccionar su estructura y contenido  
head(human_projects)
```

```
##   project organism file_source    project_home project_type n_samples  
## 1 SRP107565     human     sra data_sources/sra data_sources      216  
## 2 SRP149665     human     sra data_sources/sra data_sources       4  
## 3 SRP017465     human     sra data_sources/sra data_sources      23  
## 4 SRP119165     human     sra data_sources/sra data_sources       6  
## 5 SRP133965     human     sra data_sources/sra data_sources      12  
## 6 SRP096765     human     sra data_sources/sra data_sources       7
```

```
# Seleccionar un estudio de interés  
human_projects[709, ]
```

```
##   project organism file_source    project_home project_type n_samples  
## 709 SRP075398     human     sra data_sources/sra data_sources      18
```

```
# Filtrar el dataframe para seleccionar un proyecto específico basado en su ID y tipo  
project_info <- subset(  
  human_projects,  
  project == "SRP075398" & project_type == "data_sources"  
)
```

```
# Mostrar la información del proyecto seleccionado para confirmar que se ha  
# filtrado correctamente  
project_info
```

```
##   project organism file_source    project_home project_type n_samples  
## 709 SRP075398     human     sra data_sources/sra data_sources      18
```

```
# Crear un objeto de tipo RangedSummarizedExperiment (RSE) con la información  
# a nivel de genes  
rse_gene_SRP075398 <- create_rse(project_info)
```

```
## 2025-02-07 23:35:52.11554 downloading and reading the metadata.
```

```
## 2025-02-07 23:35:52.918874 caching file sra.sra.SRP075398.MD.gz.
```

```

## 2025-02-07 23:35:53.721565 caching file sra.recount_project.SRP075398.MD.gz.

## 2025-02-07 23:35:54.535165 caching file sra.recount_qc.SRP075398.MD.gz.

## 2025-02-07 23:35:55.332418 caching file sra.recount_seq_qc.SRP075398.MD.gz.

## 2025-02-07 23:35:56.259312 caching file sra.recount_pred.SRP075398.MD.gz.

## 2025-02-07 23:35:56.536538 downloading and reading the feature information.

## 2025-02-07 23:35:57.030372 caching file human.gene_sums.G026.gtf.gz.

## 2025-02-07 23:35:58.159632 downloading and reading the counts: 18 samples across 63856 features.

## 2025-02-07 23:35:58.844525 caching file sra.gene_sums.SRP075398.G026.gz.

## 2025-02-07 23:35:59.423482 constructing the RangedSummarizedExperiment (rse) object.

# Explorar el objeto RSE
rse_gene_SRP075398

## class: RangedSummarizedExperiment
## dim: 63856 18
## metadata(8): time_created recount3_version ... annotation recount3_url
## assays(1): raw_counts
## rownames(63856): ENSG00000278704.1 ENSG00000277400.1 ...
##   ENSG00000182484.15_PAR_Y ENSG00000227159.8_PAR_Y
## rowData names(10): source type ... havana_gene tag
## colnames(18): SRR3544525 SRR3544526 ... SRR3544537 SRR3544540
## colData names(175): rail_id external_id ...
##   recount_pred.curated.cell_line BigWigURL

## Información sobre el RSE creado
metadata(rse_gene_SRP075398)

## $time_created
## [1] "2025-02-07 23:35:59 CST"
##
## $recount3_version
##           package ondiskversion loadedversion
## recount3 recount3      1.16.0      1.16.0
##                               path
## recount3 /usr/local/lib/R/site-library/recount3
##           loadedpath attached is_base      date
## recount3 /usr/local/lib/R/site-library/recount3     TRUE    FALSE 2024-10-29
##                               source md5ok          library
## recount3 Bioconductor 3.20 (R 4.4.2)    NA /usr/local/lib/R/site-library
##
## $project
## [1] "SRP075398"

```

```

## 
## $project_home
## [1] "data_sources/sra"
##
## $type
## [1] "gene"
##
## $organism
## [1] "human"
##
## $annotation
## [1] "gencode_v26"
##
## $recount3_url
## [1] "http://duffel.rail.bio/recount3"

## Número de genes y número de muestras
dim(rse_gene_SRP075398)

```

```
## [1] 63856    18
```

El estudio **SRP075398** se compuso de **18 muestras**, para las cuales tenemos **63,856 genes** en GENCODE v26. La información específica de la anotación está disponible rowRanges() como se muestra a continuación con la columna gene\_id utilizada para identificar genes en cada una de las anotaciones.

```
# Información sobre los genes
rowRanges(rse_gene_SRP075398)
```

```

## GRanges object with 63856 ranges and 10 metadata columns:
##           seqnames      ranges strand |  source
##           <Rle>        <IRanges>  <Rle> | <factor>
## ENSG00000278704.1 GL000009.2    56140-58376   - | ENSEMBL
## ENSG00000277400.1 GL000194.1   53590-115018   - | ENSEMBL
## ENSG00000274847.1 GL000194.1   53594-115055   - | ENSEMBL
## ENSG00000277428.1 GL000195.1   37434-37534   - | ENSEMBL
## ENSG00000276256.1 GL000195.1   42939-49164   - | ENSEMBL
##           ...
##           ...       ...     ...   ...
## ENSG0000124334.17_PAR_Y     chrY 57184101-57197337   + | HAVANA
## ENSG0000185203.12_PAR_Y     chrY 57201143-57203357   - | HAVANA
## ENSG00000270726.6_PAR_Y     chrY 57190738-57208756   + | HAVANA
## ENSG00000182484.15_PAR_Y     chrY 57207346-57212230   + | HAVANA
## ENSG00000227159.8_PAR_Y     chrY 57212184-57214397   - | HAVANA
##           type bp_length phase          gene_id
##           <factor> <numeric> <integer> <character>
## ENSG00000278704.1   gene    2237    <NA>  ENSG00000278704.1
## ENSG00000277400.1   gene    2179    <NA>  ENSG00000277400.1
## ENSG00000274847.1   gene    1599    <NA>  ENSG00000274847.1
## ENSG00000277428.1   gene     101    <NA>  ENSG00000277428.1
## ENSG00000276256.1   gene    2195    <NA>  ENSG00000276256.1
##           ...
##           ...       ...     ...   ...
## ENSG0000124334.17_PAR_Y   gene    2504    <NA>  ENSG0000124334.17_P..
## ENSG0000185203.12_PAR_Y   gene    1054    <NA>  ENSG0000185203.12_P..
## ENSG00000270726.6_PAR_Y   gene     773    <NA>  ENSG00000270726.6_PA..
```

```

##   ENSG00000182484.15_PAR_Y      gene      4618      <NA> ENSG00000182484.15_P..
##   ENSG00000227159.8_PAR_Y      gene      1306      <NA> ENSG00000227159.8_PA..
##                                     gene_type    gene_name      level
##                                     <character> <character> <character>
##   ENSG00000278704.1      protein_coding  BX004987.1      3
##   ENSG00000277400.1      protein_coding  AC145212.2      3
##   ENSG00000274847.1      protein_coding  AC145212.1      3
##   ENSG00000277428.1      misc_RNA       Y_RNA      3
##   ENSG00000276256.1      protein_coding  AC011043.1      3
##                                     ...
##                                     ...      ...      ...
##   ENSG00000124334.17_PAR_Y  protein_coding  IL9R      2
##   ENSG00000185203.12_PAR_Y  antisense      WASIR1      2
##   ENSG00000270726.6_PAR_Y  processed_transcript AJ271736.10      2
##   ENSG00000182484.15_PAR_Y transcribed_unproces..      WASH6P      2
##   ENSG00000227159.8_PAR_Y unprocessed_pseudogene      DDX11L16      2
##                                     havana_gene      tag
##                                     <character> <character>
##   ENSG00000278704.1          <NA>      <NA>
##   ENSG00000277400.1          <NA>      <NA>
##   ENSG00000274847.1          <NA>      <NA>
##   ENSG00000277428.1          <NA>      <NA>
##   ENSG00000276256.1          <NA>      <NA>
##                                     ...
##                                     ...      ...
##   ENSG00000124334.17_PAR_Y  OTTHUMG00000022720.1      PAR
##   ENSG00000185203.12_PAR_Y  OTTHUMG00000022676.3      PAR
##   ENSG00000270726.6_PAR_Y  OTTHUMG00000184987.2      PAR
##   ENSG00000182484.15_PAR_Y OTTHUMG00000022677.5      PAR
##   ENSG00000227159.8_PAR_Y  OTTHUMG00000022678.1      PAR
##   -----
## seqinfo: 374 sequences from an unspecified genome; no seqlengths

```

## Preparación de los datos

```
# Convertir las cuentas por nucleotido a cuentas por lectura usando compute_read_counts()
assay(rse_gene_SRP075398, "counts") <- compute_read_counts(rse_gene_SRP075398)
```

```
# Inspeccionar la información experimental de cada muestra
rse_gene_SRP075398$sra.sample_attributes[]
```

```

## [1] "cell line;;LCC9|source_name;;LCC9 cell line pre-miR-29b-1 transfected|transfection;;Pre-miR-29b-1"
## [2] "cell line;;LCC9|source_name;;LCC9 cell line Anti-miR-29a transfected|transfection;;Anti-miR-29a"
## [3] "cell line;;LCC9|source_name;;LCC9 cell line Anti-miR-29a transfected|transfection;;Anti-miR-29a"
## [4] "cell line;;LCC9|source_name;;LCC9 cell line Anti-miR-29a transfected|transfection;;Anti-miR-29a"
## [5] "cell line;;LCC9|source_name;;LCC9 cell line Pre-miR-29a transfected|transfection;;Pre-miR-29a"
## [6] "cell line;;LCC9|source_name;;LCC9 cell line Pre-miR-29a transfected|transfection;;Pre-miR-29a"
## [7] "cell line;;LCC9|source_name;;LCC9 cell line Pre-miR-29a transfected|transfection;;Pre-miR-29a"
## [8] "cell line;;MCF-7|source_name;;MCF-7 cell line pre-miR-29b-1 transfected|transfection;;Pre-miR-29b-1"
## [9] "cell line;;MCF-7|source_name;;MCF-7 cell line pre-miR-29b-1 transfected|transfection;;Pre-miR-29b-1"
## [10] "cell line;;MCF-7|source_name;;MCF-7 cell line Pre-miR-29a transfected|transfection;;Pre-miR-29a"
## [11] "cell line;;MCF-7|source_name;;MCF-7 cell line Pre-miR-29a transfected|transfection;;Pre-miR-29a"
## [12] "cell line;;LCC9|source_name;;LCC9 cell line pre-miR-29b-1 transfected|transfection;;Pre-miR-29b-1"
## [13] "cell line;;LCC9|source_name;;LCC9 cell line pre-miR-29b-1 transfected|transfection;;Pre-miR-29b-1"

```

```

## [14] "cell line;;MCF-7|source_name;;MCF-7 cell line pre-miR-29b-1 transfected|transfection;;Pre-miR-29b-1"
## [15] "cell line;;MCF-7|source_name;;MCF-7 cell line Anti-miR-29a transfected|transfection;;Anti-miR-29a"
## [16] "cell line;;MCF-7|source_name;;MCF-7 cell line Anti-miR-29a transfected|transfection;;Anti-miR-29a"
## [17] "cell line;;MCF-7|source_name;;MCF-7 cell line Anti-miR-29a transfected|transfection;;Anti-miR-29a"
## [18] "cell line;;MCF-7|source_name;;MCF-7 cell line Pre-miR-29a transfected|transfection;;Pre-miR-29a"

# Expandir los atributos en columnas separadas para facilitar su uso
rse_gene_SRP075398 <- expand_sra_attributes(rse_gene_SRP075398)

# Extraer y mostrar las columnas que contienen atributos
colData(rse_gene_SRP075398) [
  ,
  grepl("^sra_attribute", colnames(colData(rse_gene_SRP075398)))
]

## DataFrame with 18 rows and 3 columns
##           sra_attribute.cell_line sra_attribute.source_name
##           <character>          <character>
## SRR3544525          LCC9    LCC9 cell line pre-m..
## SRR3544526          LCC9    LCC9 cell line Anti-..
## SRR3544527          LCC9    LCC9 cell line Anti-..
## SRR3544528          LCC9    LCC9 cell line Anti-..
## SRR3544529          LCC9    LCC9 cell line Pre-m..
## ...
## SRR3544534          MCF-7   MCF-7 cell line pre-..
## SRR3544535          MCF-7   MCF-7 cell line Anti-..
## SRR3544536          MCF-7   MCF-7 cell line Anti-..
## SRR3544537          MCF-7   MCF-7 cell line Anti-..
## SRR3544540          MCF-7   MCF-7 cell line Pre-..

##           sra_attribute.transfection
##           <character>
## SRR3544525      Pre-miR-29b-1
## SRR3544526      Anti-miR-29a
## SRR3544527      Anti-miR-29a
## SRR3544528      Anti-miR-29a
## SRR3544529      Pre-miR-29a
## ...
## SRR3544534      Pre-miR-29b-1
## SRR3544535      Anti-miR-29a
## SRR3544536      Anti-miR-29a
## SRR3544537      Anti-miR-29a
## SRR3544540      Pre-miR-29a

# Ajustar el tipo de dato de las variables categóricas

rse_gene_SRP075398$sra_attribute.cell_line <-
  factor(rse_gene_SRP075398$sra_attribute.cell_line)

rse_gene_SRP075398$sra_attribute.source_name <-
  factor(tolower(rse_gene_SRP075398$sra_attribute.source_name))

rse_gene_SRP075398$sra_attribute.transfection <-
  factor(rse_gene_SRP075398$sra_attribute.transfection)

```

```

# Resumen estadístico de las variables seleccionadas
summary(as.data.frame(colData(rse_gene_SRP075398) [
  ,
  grep("sra_attribute.[cell_line|source_name|transfection]",
    colnames(colData(rse_gene_SRP075398)))
]))
```

```

##   sra_attribute.cell_line           sra_attribute.source_name
## LCC9 :9                         lcc9 cell line anti-mir-29a transfected :3
## MCF-7:9                         lcc9 cell line pre-mir-29a transfected :3
##                               lcc9 cell line pre-mir-29b-1 transfected :3
##                               mcf-7 cell line anti-mir-29a transfected :3
##                               mcf-7 cell line pre-mir-29a transfected :3
##                               mcf-7 cell line pre-mir-29b-1 transfected:3
##   sra_attribute.transfection
## Anti-miR-29a :6
## Pre-miR-29a  :6
## Pre-miR-29b-1:6
##
##
```

```

# Calcular la proporción de lecturas asignadas a genes para evaluar la calidad
# de las muestras
```

```

rse_gene_SRP075398$assigned_gene_prop <-
  rse_gene_SRP075398$recount_qc.gene_fc_count_all.assigned /
  rse_gene_SRP075398$recount_qc.gene_fc_count_all.total
```

```

# Resumen de la nueva variable para identificar si las muestras tienen una
# asignación adecuada de lecturas (valores cercanos a 1 son indicadores de buena calidad)
```

```

summary(rse_gene_SRP075398$assigned_gene_prop)
```

```

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## 0.6076  0.6405  0.6603  0.6585  0.6696  0.7017
```

## Filtrar genes de baja expresión

```

# Guardar el objeto original
rse_gene_SRP075398_unfiltered <- rse_gene_SRP075398
```

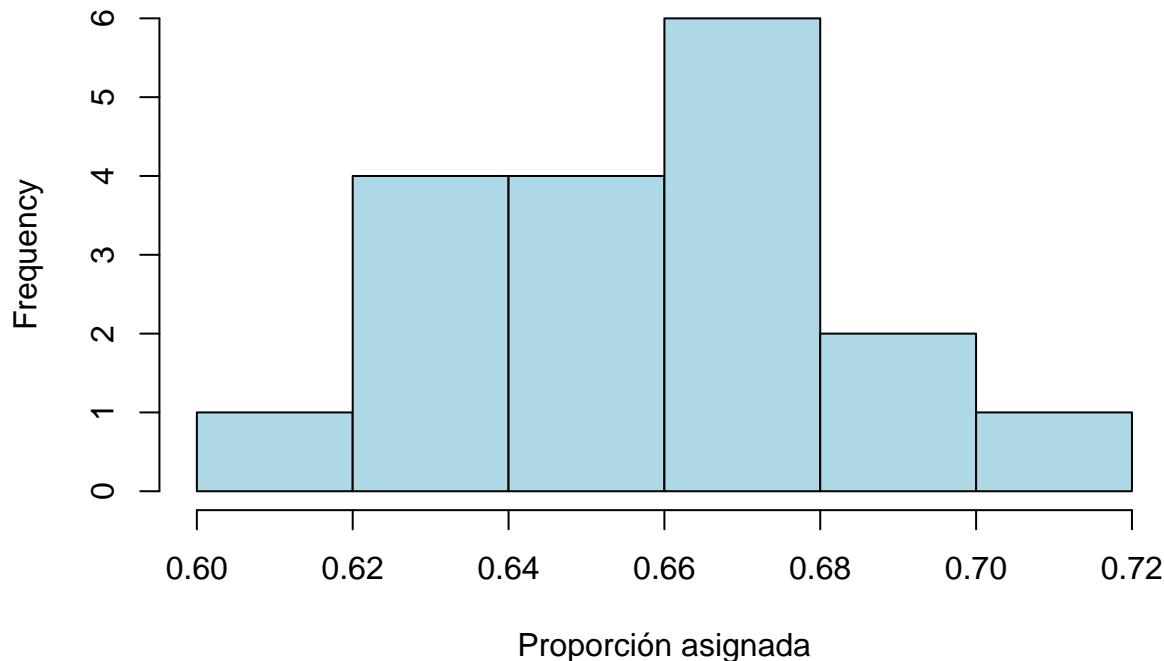
```

# Visualizar la distribución de la proporción de lecturas asignadas a genes en
# cada muestra
```

```

hist(rse_gene_SRP075398$assigned_gene_prop,
  main = "Proporción de lecturas asignadas a genes",
  xlab = "Proporción asignada", col = "lightblue")
```

## Proporción de lecturas asignadas a genes



```
# Verificar si existen muestras de baja calidad antes del filtrado
table(rse_gene_SRP075398$assigned_gene_prop < 0.3)

##
## FALSE
##      18

# Filtrar las muestras con proporción de lecturas asignadas superior a 0.3
rse_gene_SRP075398 <- rse_gene_SRP075398[, rse_gene_SRP075398$assigned_gene_prop > 0.3]

# Crear un objeto DGEList, para el análisis diferencial usando edgeR
dge <- DGEList(counts = assay(rse_gene_SRP075398, "counts"))

# Filtrar genes de baja expresión considerando combinaciones de transfección y
# línea celular
keep <- filterByExpr(dge, group = interaction(
  rse_gene_SRP075398$sra_attribute.transfection,
  rse_gene_SRP075398$sra_attribute.cell_line
))
rse_gene_SRP075398 <- rse_gene_SRP075398[keep, ]

# Dimensiones finales
dim(rse_gene_SRP075398)

## [1] 23741     18
```

```

# Porcentaje de genes retenidos
round(nrow(rse_gene_SRP075398) / nrow(rse_gene_SRP075398_unfiltered) * 100, 2)

## [1] 37.18

```

Se descartó los genes de baja expresión porque no contribuyen significativamente a las conclusiones biológicas. De modo que, después del filtrado se obtuvieron **23,741** lo cual representa el **37.18%** de genes retenidos.

## Normalización de los datos

```

# Crear un objeto DGEList para normalización
dge <- DGEList(
  counts = assay(rse_gene_SRP075398, "counts"),
  genes = rowData(rse_gene_SRP075398)
)

# Normalización TMM
dge <- calcNormFactors(dge)

dge

## An object of class "DGEList"
## $counts
##          SRR3544525 SRR3544526 SRR3544527 SRR3544528 SRR3544529
## ENSG00000223972.5     44      31      54      44      62
## ENSG00000227232.5    297     264     405     352     242
## ENSG00000238009.6     32      27      19      32      20
## ENSG00000233750.3     10      15       6       1       9
## ENSG00000268903.1     17       7       9       6      17
##          SRR3544530 SRR3544531 SRR3544532 SRR3544533 SRR3544538
## ENSG00000223972.5     37      51      13      18      16
## ENSG00000227232.5    215     277     200     204     245
## ENSG00000238009.6     15      25      19      30      48
## ENSG00000233750.3     7       9      13       9      14
## ENSG00000268903.1     6      13       8       3       4
##          SRR3544539 SRR3544523 SRR3544524 SRR3544534 SRR3544535
## ENSG00000223972.5     13      71      52      10      11
## ENSG00000227232.5    105     509     353     217     123
## ENSG00000238009.6     23      45      28      26      31
## ENSG00000233750.3     3       16       7       8       3
## ENSG00000268903.1     1      35      24       0       4
##          SRR3544536 SRR3544537 SRR3544540
## ENSG00000223972.5     7       9      25
## ENSG00000227232.5    74     163     183
## ENSG00000238009.6     32      27      45
## ENSG00000233750.3     8       3       6
## ENSG00000268903.1     0       2       3
## 23736 more rows ...
##
## $samples

```

```

##          group lib.size norm.factors
## SRR3544525     1 40711468    1.0516143
## SRR3544526     1 44717701    1.0300052
## SRR3544527     1 55798559    0.9927744
## SRR3544528     1 61298574    0.9995922
## SRR3544529     1 34513836    1.0385196
## 13 more rows ...
##
## $genes
##           source type bp_length phase      gene_id
## ENSG00000223972.5 HAVANA gene      1735     NA ENSG00000223972.5
## ENSG00000227232.5 HAVANA gene      1351     NA ENSG00000227232.5
## ENSG00000238009.6 HAVANA gene      3726     NA ENSG00000238009.6
## ENSG00000233750.3 HAVANA gene      3812     NA ENSG00000233750.3
## ENSG00000268903.1 HAVANA gene      755      NA ENSG00000268903.1
##           gene_type      gene_name level
## ENSG00000223972.5 transcribed_unprocessed_pseudogene      DDX11L1    2
## ENSG00000227232.5 unprocessed_pseudogene      WASH7P    2
## ENSG00000238009.6 lincRNA      RP11-34P13.7    2
## ENSG00000233750.3 processed_pseudogene      CICP27    1
## ENSG00000268903.1 processed_pseudogene      RP11-34P13.15   2
##           havana_gene      tag
## ENSG00000223972.5 OTTHUMG00000000961.2      <NA>
## ENSG00000227232.5 OTTHUMG00000000958.1      <NA>
## ENSG00000238009.6 OTTHUMG00000001096.2 overlapping_locus
## ENSG00000233750.3 OTTHUMG00000001257.3 pseudo_consens
## ENSG00000268903.1 OTTHUMG00000182518.2      <NA>
## 23736 more rows ...

```

## Determinar el modelo estadístico

```

# Construcción de la matriz de diseño para el modelo lineal.
mod <- model.matrix(
  ~ sra_attribute.cell_line + sra_attribute.transfection + assigned_gene_prop,
  data = colData(rse_gene_SRP075398)
)

# Cada columna representa un coeficiente del modelo
colnames(mod)

## [1] "(Intercept)"
## [2] "sra_attribute.cell_lineMCF-7"
## [3] "sra_attribute.transfectionPre-miR-29a"
## [4] "sra_attribute.transfectionPre-miR-29b-1"
## [5] "assigned_gene_prop"

# Visualizar la matriz de diseño completa
# Las filas representan muestras, mientras que las columnas son las variables del modelo
mod

## (Intercept) sra_attribute.cell_lineMCF-7

```

```

## SRR3544525      1          0
## SRR3544526      1          0
## SRR3544527      1          0
## SRR3544528      1          0
## SRR3544529      1          0
## SRR3544530      1          0
## SRR3544531      1          0
## SRR3544532      1          1
## SRR3544533      1          1
## SRR3544538      1          1
## SRR3544539      1          1
## SRR3544523      1          0
## SRR3544524      1          0
## SRR3544534      1          1
## SRR3544535      1          1
## SRR3544536      1          1
## SRR3544537      1          1
## SRR3544540      1          1
##           sra_attribute.transfectionPre-miR-29a
## SRR3544525      0
## SRR3544526      0
## SRR3544527      0
## SRR3544528      0
## SRR3544529      1
## SRR3544530      1
## SRR3544531      1
## SRR3544532      0
## SRR3544533      0
## SRR3544538      1
## SRR3544539      1
## SRR3544523      0
## SRR3544524      0
## SRR3544534      0
## SRR3544535      0
## SRR3544536      0
## SRR3544537      0
## SRR3544540      1
##           sra_attribute.transfectionPre-miR-29b-1 assigned_gene_prop
## SRR3544525      1      0.6075824
## SRR3544526      0      0.6587443
## SRR3544527      0      0.6946650
## SRR3544528      0      0.7017052
## SRR3544529      0      0.6617664
## SRR3544530      0      0.6664334
## SRR3544531      0      0.6554926
## SRR3544532      1      0.6359671
## SRR3544533      1      0.6320064
## SRR3544538      0      0.6623243
## SRR3544539      0      0.6705981
## SRR3544523      1      0.6586547
## SRR3544524      1      0.6286454
## SRR3544534      1      0.6352978
## SRR3544535      0      0.6539348
## SRR3544536      0      0.6907278

```

```

## SRR3544537          0      0.66651556
## SRR3544540          0      0.6725733
## attr(,"assign")
## [1] 0 1 2 2 3
## attr(,"contrasts")
## attr(,"contrasts")$sra_attribute.cell_line
## [1] "contr.treatment"
##
## attr(,"contrasts")$sra_attribute.transfection
## [1] "contr.treatment"

```

Es importante mencionar que la línea celular o el tipo de transfección que acompañan a cada variable del modelo son a los que se les otorga el valor de 1, debido a un orden alfabético. Por lo tanto, si la primera columna es “sra\_attribute.cell\_lineMCF-7”, la línea celular correspondiente a MCF-7 tendrá valor de 1 y la línea celular LCC9 de 0.

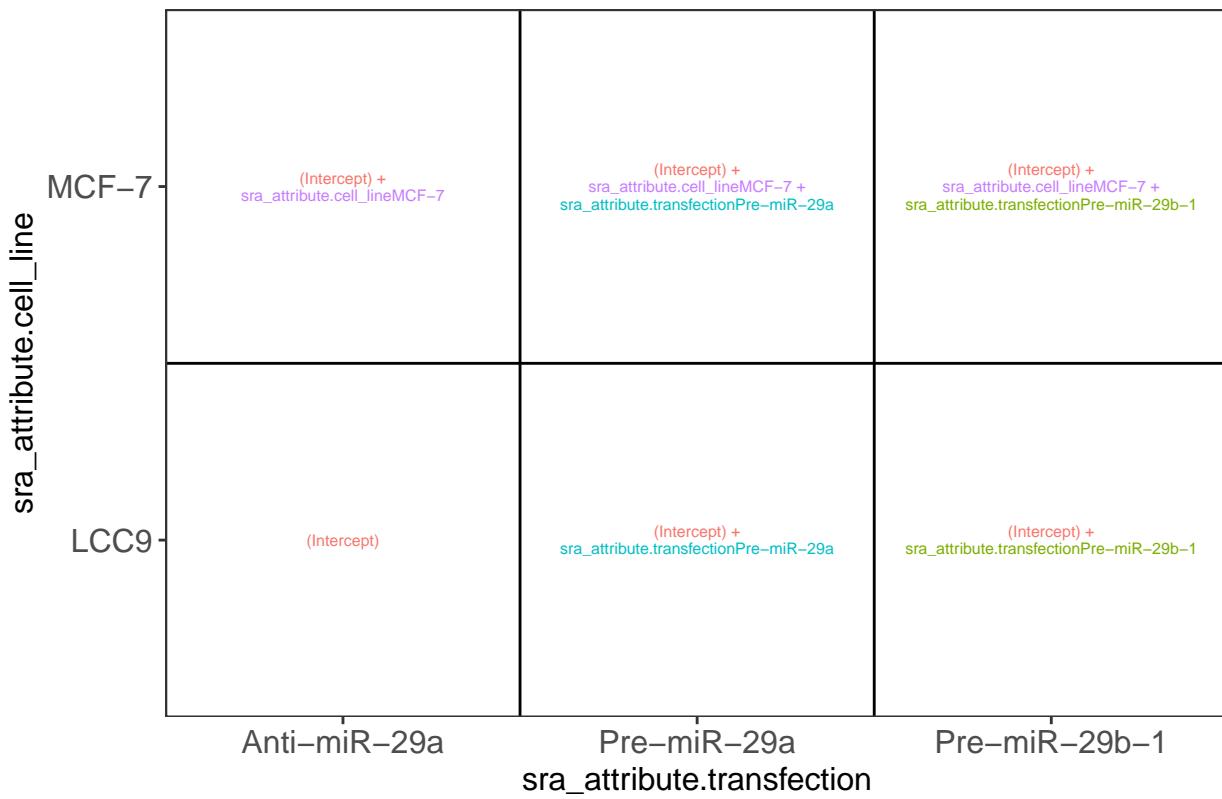
## Visualizar matriz

```

## Crear las visualizaciones
vd <- ExploreModelMatrix::VisualizeDesign(
  sampleData = colData(rse_gene_SRP075398), # Metadatos de las muestras
  designFormula = ~ sra_attribute.cell_line + sra_attribute.transfection,
  textSizeFitted = 2
)

cowplot::plot_grid(plotlist = vd$plotlist)

```



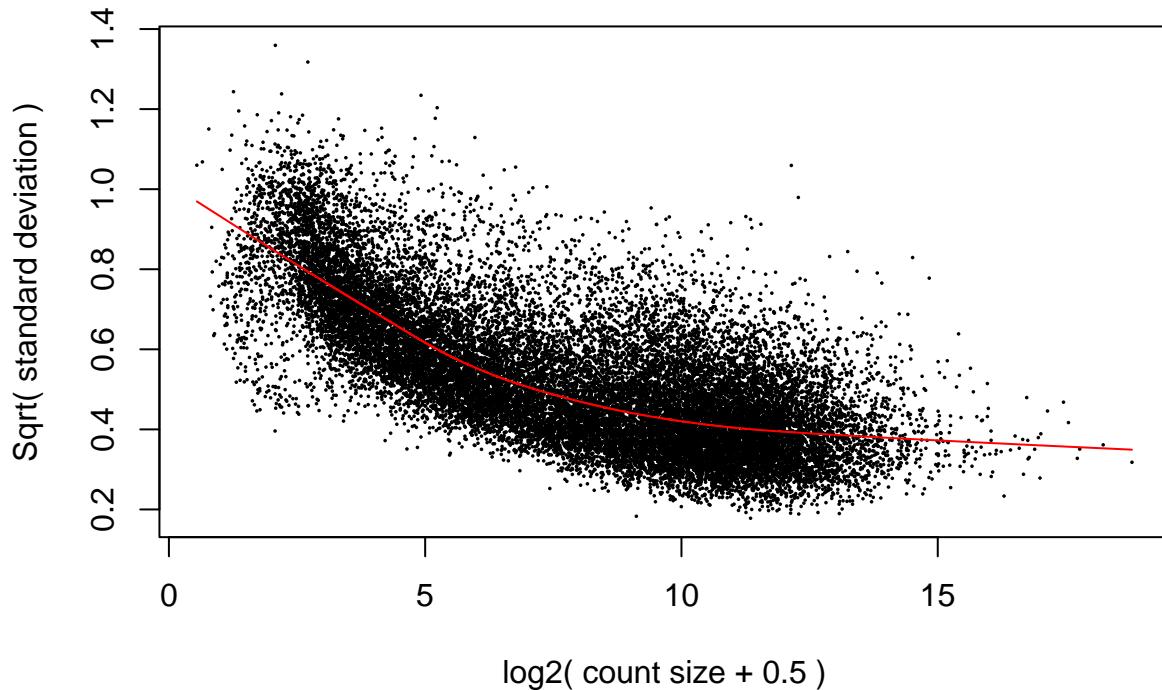
`sra_attribute.cell_lineMCF-7`: Este coeficiente es la diferencia entre la línea celular sensible al tratamiento (MCF-7), contra la línea celular resistente al tratamiento (LCC9), cuando la transfección se mantiene constante. De modo que, representa la diferencia en la respuesta entre la línea celular MCF-7 y la línea celular LCC9.

## Expresión diferencial

```
# Convertir los datos de conteo a valores log2 y ajusta las varianzas para
# hacerlos aptos para un análisis lineal
```

```
vGene <- voom(dge, mod, plot = TRUE)
```

## voom: Mean–variance trend



El método voom estima la relación media-varianza de los recuentos logarítmicos, genera un peso de precisión para cada observación y los ingresa en el flujo de trabajo del análisis bayesiano empírico de Limma. Por lo tanto, este gráfico muestra la relación entre la media y la varianza de los datos de expresión génica en escala log2. En el eje X se observa la expresión promedio de los genes, mientras que el eje Y representa la raíz cuadrada de la desviación estándar. Los genes con baja expresión presentan mayor dispersión, mientras que a niveles altos de expresión la varianza disminuye y se estabiliza. [https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-2-r29]

```
# Ajuste del modelo lineal y cálculo de estadísticas empíricas de Bayes
eb_results <- eBayes(lmFit(vGene))

# Extraer la tabla de genes diferencialmente expresados.
de_results <- topTable(
  eb_results,
  coef = 2, # Se refiere al coeficiente del segundo término en el modelo
  number = nrow(rse_gene_SRP075398),
  sort.by = "none"
)

# Dimensiones y vista preliminar de los resultados
dim(de_results)
```

```
## [1] 23741    16
```

```
head(de_results)
```

```
##           source type bp_length phase      gene_id
## ENSG00000223972.5 HAVANA gene     1735     NA ENSG00000223972.5
## ENSG00000227232.5 HAVANA gene    1351     NA ENSG00000227232.5
## ENSG00000238009.6 HAVANA gene    3726     NA ENSG00000238009.6
## ENSG00000233750.3 HAVANA gene    3812     NA ENSG00000233750.3
## ENSG00000268903.1 HAVANA gene    755      NA ENSG00000268903.1
## ENSG00000269981.1 HAVANA gene    284      NA ENSG00000269981.1
##                      gene_type      gene_name level
## ENSG00000223972.5 transcribed_unprocessed_pseudogene DDX11L1    2
## ENSG00000227232.5 unprocessed_pseudogene WASH7P    2
## ENSG00000238009.6 lincRNA RP11-34P13.7    2
## ENSG00000233750.3 processed_pseudogene CICP27    1
## ENSG00000268903.1 processed_pseudogene RP11-34P13.15   2
## ENSG00000269981.1 processed_pseudogene RP11-34P13.16   2
##          havana_gene      tag      logFC AveExpr
## ENSG00000223972.5 OTTHUMG00000000961.2 <NA> -1.57964773 -0.7349755
## ENSG00000227232.5 OTTHUMG00000000958.1 <NA> -0.65730720  2.4021094
## ENSG00000238009.6 OTTHUMG00000001096.2 overlapping_locus 0.50892797 -0.5758187
## ENSG00000233750.3 OTTHUMG00000001257.3 pseudo_consens 0.07505824 -2.5024602
## ENSG00000268903.1 OTTHUMG00000182518.2 <NA> -2.26677866 -2.9392186
## ENSG00000269981.1 OTTHUMG00000182738.2 <NA> -0.20655991 -1.1726322
##            t      P.Value adj.P.Val      B
## ENSG00000223972.5 -9.0583130 5.502604e-08 1.724585e-07 8.293369
## ENSG00000227232.5 -4.2492998 5.203366e-04 8.224575e-04 -1.699547
## ENSG00000238009.6  3.2989311 4.153389e-03 5.833271e-03 -3.230385
## ENSG00000233750.3  0.2097501 8.363087e-01 8.516989e-01 -7.250773
## ENSG00000268903.1 -4.8120051 1.545050e-04 2.624197e-04  0.559485
## ENSG00000269981.1 -0.8707765 3.957782e-01 4.289509e-01 -7.156075
```

El coeficiente 2 (sra\_attribute.cell\_lineMCF-7) evalúa la expresión diferencial entre las líneas celulares MCF-7 y LCC9, lo que podría permitir identificar genes relacionados con la resistencia a la terapia endocrina.

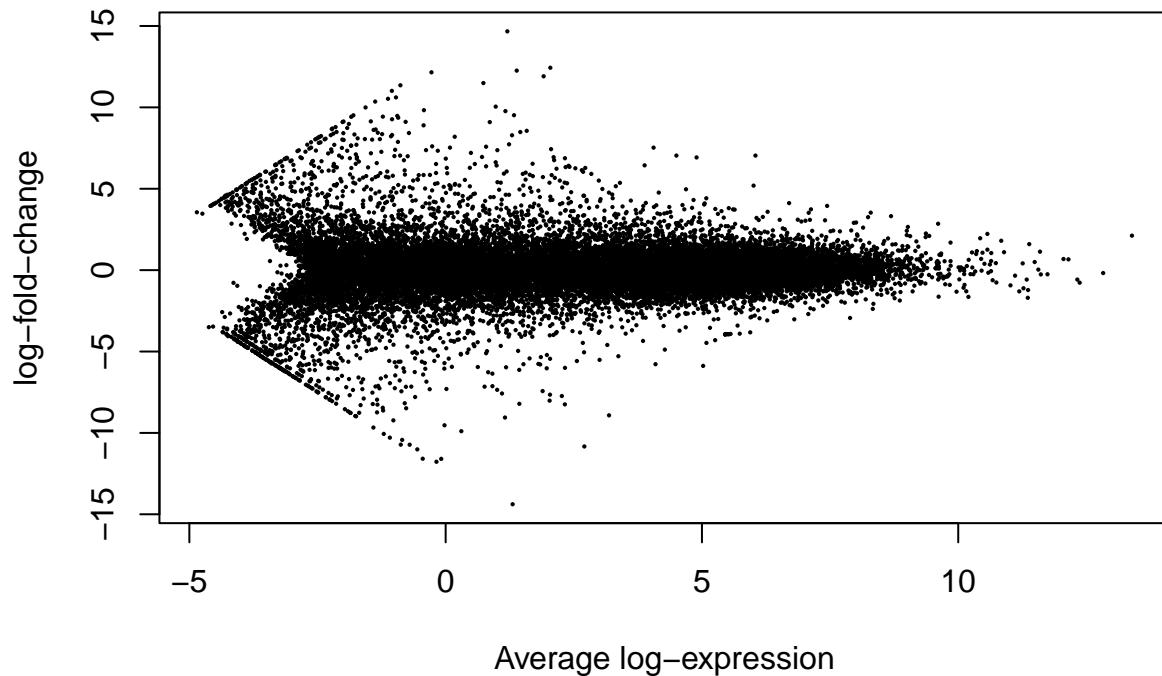
```
# Genes diferencialmente expresados con FDR < 5%
table(de_results$adj.P.Val < 0.05)
```

```
##
## FALSE  TRUE
## 4642 19099
```

```
# Visualizar los resultados estadísticos

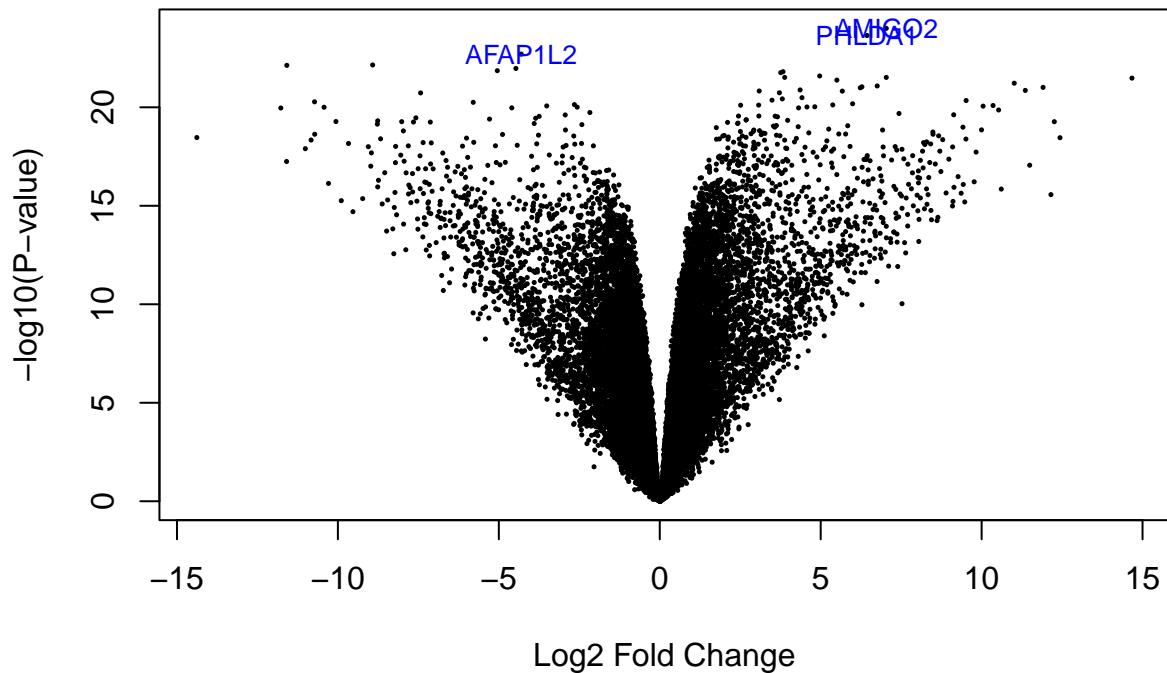
plotMA(eb_results, coef = 2)
```

## sra\_attribute.cell\_lineMCF-7



El gráfico generado por la función `plotMA()` en el análisis de expresión diferencial muestra los resultados estadísticos del contraste entre las líneas celulares MCF-7 y LCC9. En este tipo de gráfico, el eje X representa la expresión promedio en escala logarítmica para cada gen, mientras que el eje Y muestra el cambio logarítmico en la expresión ( $\log_2$  fold-change) asociado al coeficiente 2 del modelo. La línea horizontal en  $y = 0$  representa genes sin cambios significativos en su expresión. <https://www.rdocumentation.org/packages/DESeq2/versions/1.12.3/topics/plotMA>

```
volcanoplot(eb_results, coef = 2, highlight = 3, names = de_results$gene_name)
```



```
# Información de los 3 genes más significativos
de_results[de_results$gene_name %in% c("AMIGO2", "AFAP1L2", "PHLDA1"), ]
```

```
##           source type bp_length phase      gene_id
## ENSG00000169129.14 HAVANA gene     5789    NA ENSG00000169129.14
## ENSG00000139211.6 HAVANA gene     3956    NA ENSG00000139211.6
## ENSG00000139289.13 HAVANA gene     8069    NA ENSG00000139289.13
##           gene_type gene_name level      havana_gene tag
## ENSG00000169129.14 protein_coding AFAP1L2      2 OTTHUMG00000019086.3 <NA>
## ENSG00000139211.6 protein_coding AMIGO2      2 OTTHUMG00000169616.1 <NA>
## ENSG00000139289.13 protein_coding PHLDA1      2 OTTHUMG00000169783.2 <NA>
##           logFC AveExpr      t   P.Value adj.P.Val
## ENSG00000169129.14 -4.302232 5.257918 -77.07393 1.991303e-23 1.575851e-19
## ENSG00000139211.6    7.038259 6.044177  91.73069 9.814316e-25 2.330017e-20
## ENSG00000139289.13   6.438081 3.881338  87.38040 2.274059e-24 2.699422e-20
##           B
## ENSG00000169129.14 43.77924
## ENSG00000139211.6  46.45332
## ENSG00000139289.13 45.06009
```

AMIGO2 en GeneCards

AFAP1L2 en GeneCards

PHLDA1 en GeneCards

En el volcanoplot creado los genes más significativos se muestran en la parte superior. Cada punto del gráfico representa un gen. Las diferencias de log 2 entre los grupos se representan en el eje x y las diferencias de log 10 en el valor p se representan en el eje y. Los genes cuya expresión disminuye se ubican a la izquierda del cero en el eje x, mientras que los genes cuya expresión aumenta se ilustran a la derecha del cero. Entonces, el gráfico muestra genes con aumento de expresión en MCF-7 (sensible) a la derecha (en comparación con LCC9), y genes con disminución de expresión en MCF-7 a la izquierda (en comparación con LCC9). <https://www.htgmolecular.com/blog/2022-08-25/understanding-volcano-plots>

## Visualizar genes DE

```
# Revisar los top 50 genes diferencialmente expresados

# Extraer valores de los genes de interés
exprs_heatmap <- vGene$E[rank(de_results$adj.P.Val) <= 50, ]

# Crear una tabla con información de las muestras y con nombres de columnas más amigables
df <- as.data.frame(colData(rse_gene_SRP075398)[, c("sra_attribute.cell_line",
                                                    "sra_attribute.transfection")])

colnames(df) <- c("Cell_line", "Transfection")

## Guardemos los IDs de nuestros 50 genes
nombres_originales <- rownames(exprs_heatmap)

## Con match() podemos encontrar cual es cual
rownames(exprs_heatmap) <- rowRanges(rse_gene_SRP075398)$gene_name[
  match(rownames(exprs_heatmap), rowRanges(rse_gene_SRP075398)$gene_id)
]

## Guardar la imagen en un PDF largo para poder ver los nombres de los genes
pdf("pheatmap_con_nombres.pdf", height = 16, useDingbats = FALSE)
pheatmap(
  exprs_heatmap,
  cluster_rows = TRUE,
  cluster_cols = TRUE,
  show_rownames = TRUE,
  show_colnames = FALSE,
  annotation_col = df,
  fontsize_row = 6,
)
```



```
dev.off()
```

```
## pdf
## 3
```

El heatmap muestra patrones de expresión génica distintos entre las líneas celulares LCC9 (resistente al tratamiento) y MCF-7 (sensible al tratamiento) en respuesta a diferentes transfecciones. En general, se observa una diferencia en la expresión de genes entre ambas líneas celulares. Con respecto a, genes específicos como AMIGO2 se observa una expresión más alta en las células MCF-7 en comparación con las células LCC9, con todos los tipos de miRNAs usados para la transfección. Otros genes como AFAP1L2, PLS3, STAT6, FKBP10 también muestran patrones de expresión interesantes, con niveles más altos en las células LCC9 en comparación con las células MCF-7, especialmente en las muestras transfectadas con Pre-miR-29a y Pre-miR-29b-1.

Por otro lado, PHLDA1 no muestra diferencias tan marcadas en su expresión entre las dos líneas celulares, aunque parece tener una expresión ligeramente más alta en las células MCF-7 en comparación con las células LCC9.

```
# MDS (multidimensional scaling)

## Convertir los grupos de Cell_line a colores
col.group <- df$Cell_line
levels(col.group) <- brewer.pal(nlevels(col.group), "Set2")
```

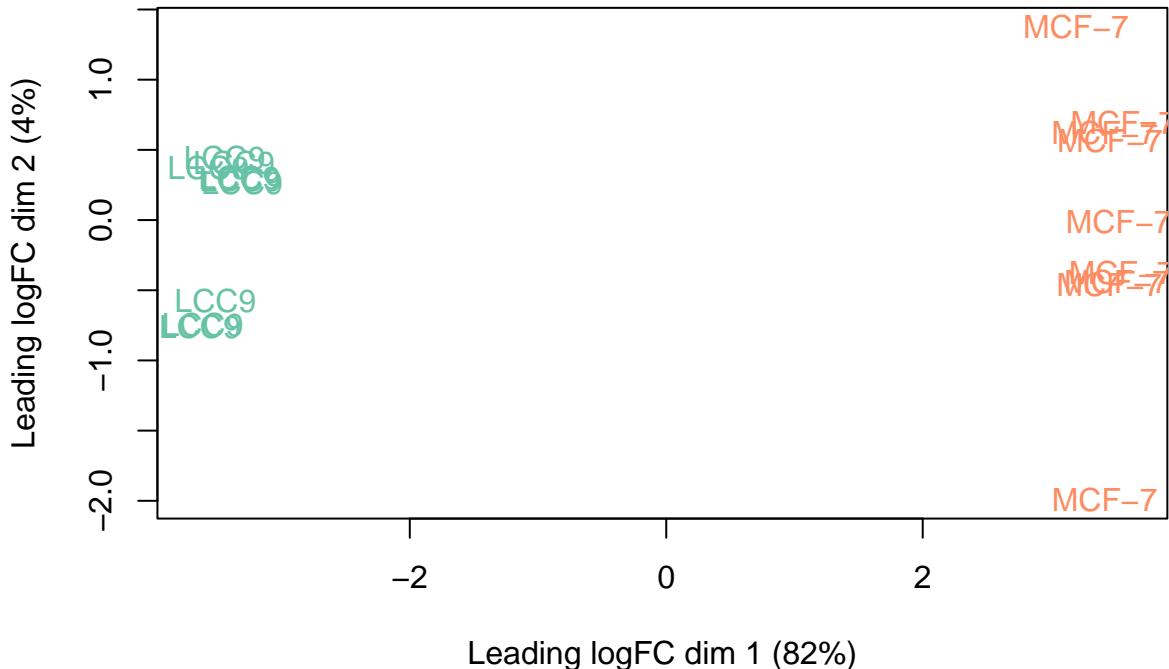
```
## Warning in brewer.pal(nlevels(col.group), "Set2"): minimal value for n is 3, returning requested pal
```

```

col.group <- as.character(col.group)

## MDS por grupos de Cell_line
plotMDS(vGene$E, labels = df$Cell_line, col = col.group)

```



Este gráfico evalúa si las muestras se agrupan según la variable Cell\_line, la separación clara entre MCF-7 y LCC9 indica que hay diferencias notables en la expresión génica entre estas líneas celulares. Esto sugiere que LCC9 y MCF-7 tienen perfiles de expresión distintos, lo que podría estar influenciado por la transfección con miR-29a/b. Por lo tanto, es posible que los genes regulados por miR-29a/b sean clave en la resistencia a la terapia endocrina.

```

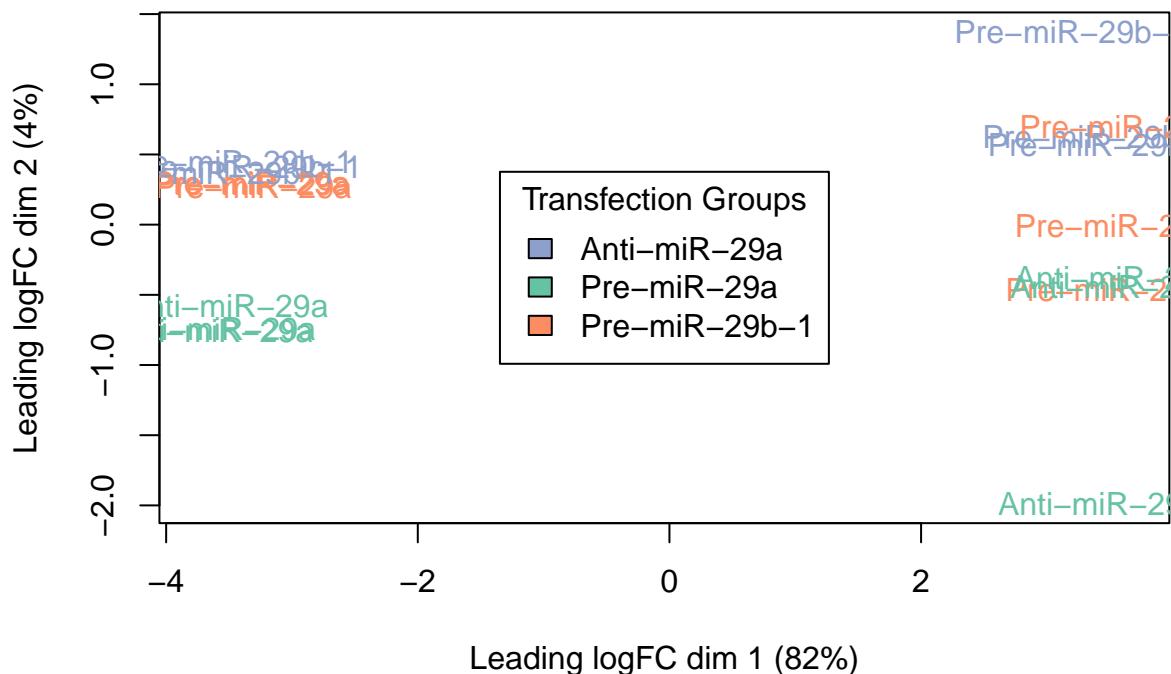
## Convertir Transfection a colores
col.group <- df$Transfection
df$Transfection <- as.factor(df$Transfection) # Asegúrate de que sea un factor
colors <- brewer.pal(nlevels(df$Transfection), "Set2") # Generar paleta de colores
levels(col.group) <- colors # Asignar colores a los niveles
col.group <- as.character(col.group) # Convertir a vector de caracteres

## MDS por grupos de Transfection
plotMDS(vGene$E, labels = df$Transfection, col = col.group,
         main = "MDS Plot by Transfection", pch = 16)

## Agregar leyenda
legend("center", legend = levels(df$Transfection), fill = unique(col.group),
       title = "Transfection Groups")

```

## MDS Plot by Transfection



El gráfico MDS muestra una cierta separación entre los grupos de transfección, donde Anti-miR-29a presenta un perfil distinto en comparación con Pre-miR-29a y Pre-miR-29b-1, que comparten similitudes pero aún muestran cierta dispersión. La dimensión principal (logFC dim 1) captura la mayor variabilidad en la expresión génica, sugiriendo que la transfección tiene un impacto significativo.

## Conclusión

El análisis de expresión diferencial realizado en este proyecto permitió identificar genes relacionados con la resistencia a la terapia endocrina en cáncer de mama. A través de diversas visualizaciones, se evidencianaron diferencias significativas en los perfiles de expresión entre las líneas celulares MCF-7 (sensibles) y LCC9 (resistentes). El heatmap mostró patrones distintivos de expresión génica entre ambas líneas celulares, destacando genes como AMIGO2 con mayor expresión en MCF-7, mientras que AFAP1L2, PLS3, STAT6 y FKBP10 presentaron niveles más altos en LCC9.

Además, los paquetes empleados en este estudio, son herramientas bioinformáticas muy útiles que permiten llevar a cabo un protocolo estándar para el análisis de expresión diferencial que incluye: procesamiento de datos, filtrado de genes, normalización, construcción de modelos y pruebas estadísticas. Por lo que, el conocimiento de este flujo de trabajo es fundamental para cualquier investigador que busca comprender mecanismos moleculares o desarrollar nuevas estrategias terapéuticas.