

# Microarray-Construcción ExpressionSet

Fernando Gordillo

2024-04-03

## Table of contents

Análisis de sobre representación . . . . .	4
--	---

## Descarga de datos

### 1. Elegir un experimento con microarrays bien de GEO bien de ArrayExpress.

El estudio elegido para el análisis proviene de GEO, y su identificador es [GSE124593](#). En el se busca estudiar las diferencias transcriptómicas entre los tumores mamarios primarios y los recurrentes. Existe evidencia de que estos tipos de tumores existen diferencias que provocan unas respuestas a tratamiento distintas, por lo que con este estudio se pretende caracterizar las diferencias entre estas dos formas de tumor mamario, para así poder llegar a individualizar los tratamientos de estos dos tipos. Los resultados de este estudio se comentaron en un artículo publicado en Julio de 2020, cuyo PMID es [31988496](#).

Para este estudio se trabajo **modelos murinos MTB/TAN** (*Mus musculus*), que desarrollan cancer de mama por la via del Her2. Las muestras analizadas provienen de tumores de mama primarios y recurrentes, habiendo 2 muestras provenientes de tumores primario, con 2 replicas cada una y otras 2 de tumores recurrentes, con 2 replicas cada una, es decir, el estudio cuenta en total con **8 muestras**. El experimento pertenece a un array de Affymetrix, más concretamente de la plataforma *Affymetrix Mouse Genome 430A 2.0 Array*.

La descarga de los datos, así como su conversión a *Affy Batch*, se realiza de la siguiente manera:

```
GEOquery::getGEOSuppFiles("GSE124593")
system("tar xvf GSE124593/GSE124593_RAW.tar")
gse124593_raw = ReadAffy()
system("rm -fr GSE124593")
system("rm *CEL.gz")
```

```
#Guardar datos crudos
save(gse124593_raw, file="./gse124593_raw.rda")
```

El siguiente chunk permite cargar el archivo con los datos crudos en formato *Affy Batch* sin necesidad de volver a descargar los datos.

```
load("./gse124593_raw.rda")
```

## Analisis de calidad

**2. Realizar un estudio sobre la calidad de las muestras. Este estudio ha de incluir:**

- Los dibujos media-diferencia de Tukey o MA-plot en el contexto ómico.

```
affy::MAplot(gse124593_raw)
```

- Una comparación de los estimadores de densidad.

```
affy::hist(gse124593_raw)
```

- Una comparación entre los diagramas de cajas.

```
affy::boxplot(gse124593_raw,col=rainbow(14),las=2,ylab="Luminescence")
```

## Preprocesamiento

**3. Preprocesar los datos con el procedimiento RMA.**

```
gse124593 =rma(gse124593_raw)
```

**4. Repetir los tres primeros apartados del apartado 2 con los nuevos datos.**

- Los dibujos media-diferencia de Tukey o MA-plot en el contexto ómico.

```
affy::MAplot(gse124593)
```

- Una comparación de los estimadores de densidad.

```
affy::hist(gse124593)
```

Este grafico tambien se puede realizar mediante el paquete ggplot con el siguiente chunk

```
library(reshape);library(ggplot2)
df=data.frame(gene=featureNames(gse124593),exprs(gse124593))
df1=melt(df,id=c("gene"))
ggplot(df1,aes(x=value,colour=variable,group=variable))+geom_density(kernel="epanechnikov")
```

- Una comparación entre los diagramas de cajas.

```
affy::boxplot(gse124593,col=rainbow(14),las=2,ylab="Luminescence")
```

## Anotacion del ExpressionSet

### 5. En este ExpressionSet hay que incluir:

- Unos identificadores primarios que sean los identificadores ENTREZID.
- En el slot fData hay que incluir los identificadores Ensembl y alguna denominación habitual en los genes para el organismo de que se trate.

Toda esta anotacion se puede realizar a la vez mediante el siguiente chunk. Para ello, se precisa de la instalacion del paquete de la base de datos con la que se ha anotado el ExpressionSet ([mouse430a2.db](#)), esto se asegura cargando el paquete con pacman, ya que si no está instalado, lo instala automaticamente.

```
pacman::p_load("AnnotationDbi","mouse430a2.db")
conver = AnnotationDbi::select(mouse430a2.db::mouse430a2.db,
                               keys=featureNames(gse124593),
                               column=c("ENTREZID","ENSEMBL","SYMBOL","ALIAS"),keytype="PROBEID")

info = match(featureNames(gse124593),conver[, "PROBEID"])
fData(gse124593) = conver[info,]
```

## Variable fenotípica

```
load("../data/gse124593.rda")
pData(gse124593)$type = factor(rep(c(1,1,1,1,2,2,2,2)),levels=1:2,
                               labels=c("Primario","Recurrente"))
```

## Guardado del ExpressionSet

```
save(gse124593, file="./gse124593.rda")
```

## Análisis de sobre representación

Ambos análisis se realizarán comparando los datos con la base de datos de **Gene Ontology**, en concreto con los términos relacionados con **procesos biológicos**. Para ello, primero construimos esta colección GO para el **organismo Mus musculus** con el paquete **EnrichmentBrowser**.

```
mmu_go = EnrichmentBrowser::getGenesets(org = "mmu", db = "go",  
                                         onto = "BP", mode = "GO.db")  
names(mmu_go) = sapply(names(mmu_go),  
                        function(x) unlist(strsplit(x,split = "_"))[1])  
save(mmu_go, file = "mmu_go.rda")
```