

gse18842

Celeste Moya Valera

2024-04-03

Table of contents

| | |
|--|----------|
| Introduction | 1 |
| Packages | 1 |
| Experiment | 2 |
| Download of the data | 2 |
| Quality control | 2 |
| Representacion MA-plot | 2 |
| Comparacion de estimadores de densidad | 2 |
| Boxplots | 2 |
| Normalización de los datos: RMA | 3 |
| ExpressionSet | 3 |
| Anotación del ExpressionSet gse18842 | 5 |
| Guardado de datos | 5 |

Introduction

Packages

```
pacman::p_load(Biobase,BiocGenerics,GEOquery,affy,affyPLM,  
               org.Hs.eg.db,hgu133plus2.db,AnnotationDbi)
```

Experiment

1. We are going to use the data set [GSE18842](#).
2. Experimental data: *Homo sapiens*, 91 samples and platform [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array

Download of the data

```
options(timeout=2000)
GEOquery::getGEOSuppFiles("GSE18842")
system("tar -xvf GSE18842/GSE18842_RAW.tar")
gse18842 = ReadAffy()
save(gse18842raw,file = "gse18842raw.rda")
system("rm -fr GSE18842")
system("rm *CEL.gz")

load("gse18842raw.rda")
```

Quality control

Representacion MA-plot

```
affy::MAplot(gse18842raw)
```

Comparacion de estimadores de densidad

```
affy::hist(gse18842raw)
```

Boxplots

```
affy::boxplot(
  gse18842raw,
  col="red",
```

```
main="Boxplots before RMA",
ylab="Log2 Intensity",
xlab="samples",
names=FALSE)
```

Normalización de los datos: RMA

```
gse18842 = affy::rma(gse18842raw)
save(gse18842,file="gse18842.rda")
```

ExpressionSet

```
infoData = new("MIAME",
  name = "Ma Esther Farez-Vidal",
  lab = "Hospital Universitario San Cecilio",
  contact = "efarez@ugr.es",
  url = "https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18842",
  pubMedIds = "20878980",
  title = "Gene expression analysis of human lung cancer and control samples",
  abstract = "The development of reliable gene expression profiling technology
is having an increasing impact on our understanding of lung cancer biology.
The present study aims to determine whether the phenotypic heterogeneity and
genetic diversity of lung cancer are correlated.
```

PATIENTS AND METHODS

In this study, microarray analysis was performed in a set of 91 non-small cell lung cancer (NSCLC) samples in order to: establish gene signatures in primary adenocarcinomas and squamous-cell carcinomas; determine differentially expressed gene sequences at different stages of the disease; and identify sequences with biological significance for tumor progression. After microarray analysis, the expression level of 92 selected genes was validated by qPCR in an independent set of 70 samples.

RESULTS

Gene sequences were differentially expressed as a function of tumor type, stage, and differentiation grade. High upregulation was observed for KRT15 and PKP1, which may be good markers to distinguish squamous cell carcinoma samples. High downregulation was observed for DSG3 in stage IA adenocarcinomas.

CONCLUSION

Expression signatures in NSCLC distinguish tumor type, stage, and differentiation

```

grade.
Keywords: Tumor vs control comparative genomics study")

pd0 = read.csv("gse18842_pData.csv",header=TRUE)
rownames(pd0) = colnames(exprs(gse18842))
metadatos = data.frame(labelDescription =
                        colnames(pd0),row.names=colnames(pd0))
datosfenotipo = new("AnnotatedDataFrame", data = pd0, varMetadata = metadatos)
gse18842 = new("ExpressionSet", exprs=exprs(datos.rma), phenoData = datosfenotipo,
              experimentData = infoData, annotation="hgu133plus2.db")

```

```

colnames(8842_pData.csv"), package = "celestemoya") pd0 = read.csv("gse18842_pData.csv",header=TRUE)
rownames(pd0) = colnames(exprs(gse18842)) metadatos = data.frame(labelDescription = col-
names(pd0),row.names=colnames(pd0)) datosfenotipo = new("AnnotatedDataFrame", data
= pd0, varMetadata = metadatos) gse18842 = new("ExpressionSet", exprs=exprs(datos.rma),
phenoData = datosfenotipo, experimentData = infoData, annotation="hgu133plus2.db")

```

```

colnames(gse18842)

```

```

::: {.cell}

```

```

```{r .cell-code}
x = read.csv("gse18842_samples.csv",header=TRUE)
names(x)
sum(table(x$description) == 1)
dim(x)
summary(x)
View(x)
pd0 = read.csv("gse18842_pData.csv",header=TRUE)
sum(table(pd0$Sample) >1)
dim(pd0)
View(pd0)
dim(gse18842)
a = match(pd0[, "Sample"], x[, "description"])
data.frame(pd0[, "Sample"], x[a, "description"])

pd0[, "Sample"] == x[a, "description"]

y = data.frame(pd0, x[a,])
View(y)
y$file = paste0(y$file, ".CEL.gz")

b = match(y$file, colnames(gse18842))

```

y[b,]

...

## Anotación del ExpressionSet gse18842

Incluimos en el fData los identificadores ENTREZID, ENSEMBL y SYMBOL (Gene Symbol).

```
library(AnnotationDbi)
library(hgu133plus2.db)

ids = keys(hgu133plus2.db, keytype="PROBEID")
identif = select(hgu133plus2.db, keys=ids, columns=c("ENTREZID", "ENSEMBL", "SYMBOL"),
 keytype="PROBEID")
identif = identif[!is.na(identif[, "ENTREZID"]),]
indices = match(featureNames(gse18842), identif$PROBEID)
fData(gse18842) = identif[indices,]
fData(gse18842)
dim(gse18842)
dim(fData(gse18842))
```

---

## Guardado de datos

Finalmente, guardamos el resultado en un archivo con extensión RDA.

```
save(gse18842, file = "gse18842.rda")
```