geod19804

Carlos Gabriel Marrero Pérez

2021-04-10

Tarea 1

Cargamos las librerías

Los datos pueden encontrarse en

```
library(Biobase)
library(GEOquery)
library(affy)
library(hgu133plus2.db)
library(AnnotationDbi)
```

Elección del experimento

Para llevar a cabo esta tarea se ha elegido el DataSet GDS3837, que se corresponde con la serie **GSE19804** (que es la referencia que se utilizará en todo el documento).

Título: estudio de genoma completo de modulación transcriptómica en mujeres no fumadoras con cáncer de pulmón en Taiwán. Si bien el tabaquismo es el mayor factor de riesgo para desarrollar cáncer de pulmón, tan solo el 7% de las pacientes con cáncer de pulmón taiwanesas tienen un historial de consumo de tabaco, porcentaje mucho más básico que el correspondiente a mujeres caucásicas. Este estudio es un análisis exhaustivo de la "firma molecular" (conjunto de genes, tránscritos de mRNA, proteínas...) del sujeto de estudio descrito. Dado que las encuestas de gran escala sobre mecanismos moleculares en estas mujeres no eran comunes, se decidió llevar a cabo este experimento. Fueron extraídas 102 muestras de tumores pulmonares y sus respectivas contrapartes sanas. El RNA de los pares fue examinado mediante el microarray Affymetrix Human U133 plus2.0, mientras que el DNA fue analizado a través de arrays Affymetrix SNP 6.0, para buscar polimorfismos de un solo nucleótido.

Información previa del experimento

```
-Animal de estudio: Homo sapiens
-Número de muestras:  120
-Array: [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
```

Descarga y preprocesado de la información

```
files <- GEOquery::getGEOSuppFiles("GSE19804")
system("tar xvf GSE19804/GSE19804_RAW.tar")
# Procesamos los ficheros cel y creamos un objeto AffyBatch de nombre "datossinnorm"
datossinnorm = ReadAffy()
system("rm -fr GSE19804")
system("rm *CEL.gz")
```

En primer lugar, haciendo uso de la función

```
affy::MAplot
```

realizamos los **MA-plot** de cada array contra el pseudoarray de referencia antes de normalizar los datos. Estas gráficas permiten visualizar las diferencias entre las medidas tomadas en dos muestras, transformando los datos en dos escalas M (log ratio) y A (media aritmética) y luego dibujando esos valores. M y A vienen definidas por, en un ejemplo entre intensidades rojas contra verdes (R vs. G):

$$M = \log_2(R/G),$$

$$A = \frac{1}{2}\log_2(RG)$$

.

```
affy::MAplot(datossinnorm)
```

Ahora, obtenemos los **estimadores de densidad** y los comparamos en un histograma:

```
affy::hist(datossinnorm)
```

Por último, creamos los **diagramas de cajas** para cada muestra:

```
affy::boxplot(datossinnorm,col="blue",
        main="Diagrama de cajas preprocesamiento",
        ylab="Intensidad en log2",
        xlab="Muestras", names=FALSE)
```

Preprocesamiento de los datos

Para preprocesar la información hacemos uso del algoritmo RMA (*Robust Multi-array Average*), en el que los valores sin tratar de intensidad se "limpian" de ruido de fondo, se transforman a logaritmo en base 2 y se normalizan en cuantiles. Posteriormente, un modelo lineal se ajusta a los datos normalizados para obtener una medida de la expresión para cada conjunto de sondas de cada *array*.

```
datosnorm <- affy::rma(datossinnorm)
```

A partir de los datos normalizados, obtenemos la matriz de expresión.

```
matrizexpr = exprs(datosnorm)
```

Ahora que hemos tratado los datos, repetimos las tres fases del estudio cualitativo anterior.

MA-plot de los datos normalizados.

```
affy::MAplot(datosnorm)
```

Estimadores de densidad de los datos normalizados

```
affy::hist(datosnorm)
```

Diagramas de cajas de los datos normalizados

```
affy::boxplot(datosnorm,col="orange",
        main="Diagrama de cajas postprocesado",
        ylab="Intensidad en log2",
        xlab="Muestras", names=FALSE)
```

Elaboración del ExpressionSet

Rellenamos la información de los datos en formato MIAME

```
infoData = new('MIAME',
        name='Tzu-Pin Lu',
        lab='National Taiwan University, Taiwan',
        contact ='tplu@ntu.edu.tw',
                    url ="https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19804",
             pubMedIds =c("20802022","25889623"),
             title = "  Genome-wide screening of transcriptional modulation in non-smoking female
        lung cancer in Taiwan",
             abstract = "Although smoking is the major risk factor for lung cancer, only 7% of female
             lung cancer patients in Taiwan have a history of cigarette smoking, extremely
             lower than those in Caucasian females.
        This report is a comprehensive analysis of the molecular signature of
        non-smoking female lung cancer in Taiwan.")

finput = system.file("externaldata","atributos.csv", package = "carlosmarrero")
pd0 = read.csv(finput, header=TRUE, row.names = 1)
metadatos = data.frame(labelDescription = colnames(pd0),row.names=colnames(pd0))
datosfenotipo = new("AnnotatedDataFrame", data = pd0, varMetadata = metadatos)
geod19804  = new("ExpressionSet",
                exprs=matrizexpr,
                phenoData = datosfenotipo,
                experimentData = infoData,
                annotation='hgu133plus2')
```

Anotación

```
a = AnnotationDbi::select(hgu133plus2.db::hgu133plus2.db,keys=featureNames(geod19804),
                          column=c("ENTREZID","ENSEMBL","GENENAME"),keytype="PROBEID")
a = a[!is.na(a[,"ENTREZID"]),]
b = match(featureNames(geod19804),a[,"PROBEID"])
fData(geod19804) = a[b,]
dim(geod19804)
fData(geod19804)
save(geod19804,file="geod19804.rda")
```