Explorando el Transcriptoma con Datos de Expresión Genética

Datos de Expresión Genética

Yered Pita-Juárez

6/1/2015

Datos de Microarreglos

- Muestra de ARN preparada, etiquetada e hibridizada al microarreglo
- El escáner genera una imagen (`.DAT`)
- Se separan los pixeles por sonda
- La intensidad luminosa se promedia por sonda (`.CEL`)

Datos de Microarreglos

- Descarga y extrae los estos archivos en tu working directory
  `https://www.dropbox.com/s/qeg81zxoxfzjknr/celfiles.zip`
- working directory

```
wd <- getwd()
basedir <- paste0(wd, "/celfiles")
setwd(basedir)
```

- Primero vamos a extraer la información acerca de estas muestras

```
library(affy)
tab <- read.delim("sampleinfo.txt",
   check.names=FALSE,as.is=TRUE)
rownames(tab) <- tab$filenames
tab
```

Datos de Microarreglos

- Vamos a ver que archivos `CEL` estan disponibles

```
fns <- list.celfiles()
fns
```

- ¿Tenemos todos los archivos?

```
fns %in% tab[,1]
```

- Vamos a crear un objecto AffyBatch incluyendo la informacion provista acerca de estas muestras

```
ab <- ReadAffy(phenoData=tab)
dim(pData(ab))
```

- ¿Que plataforma?

```
annotation(ab)
```

- Affymetrix Human Genome U95 (`hgu95a`)

Procesando los Datos

Robust Multichip Average (RMA)

1. Corrección de fondo
2. Normalización
3. Resumen

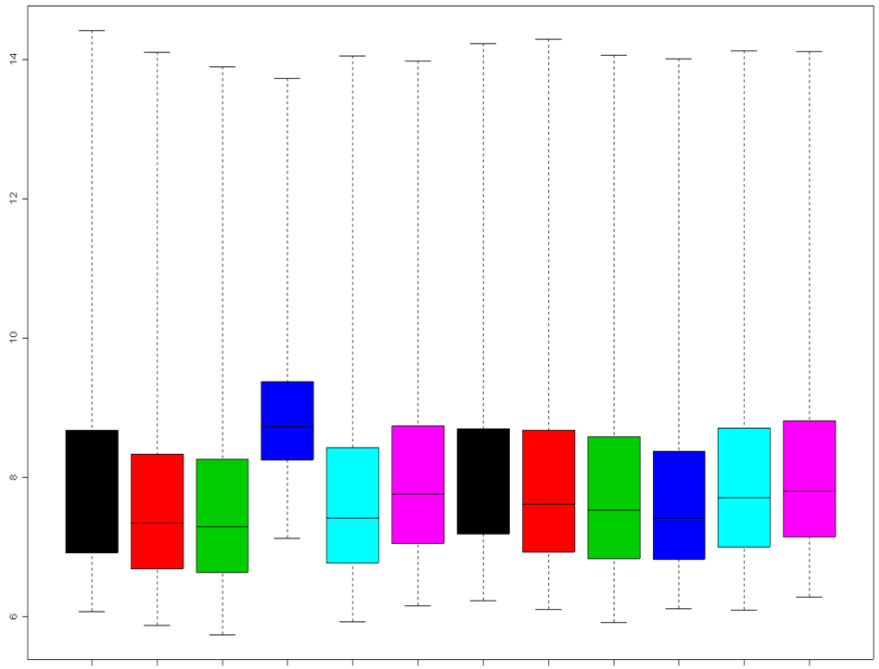
Corrección de fondo

- Observado = Señal + Fondo
- Nos interesa la señal
- Estimar la señal usando métodos de inferencia estadística

Procesando los Datos

Normalización

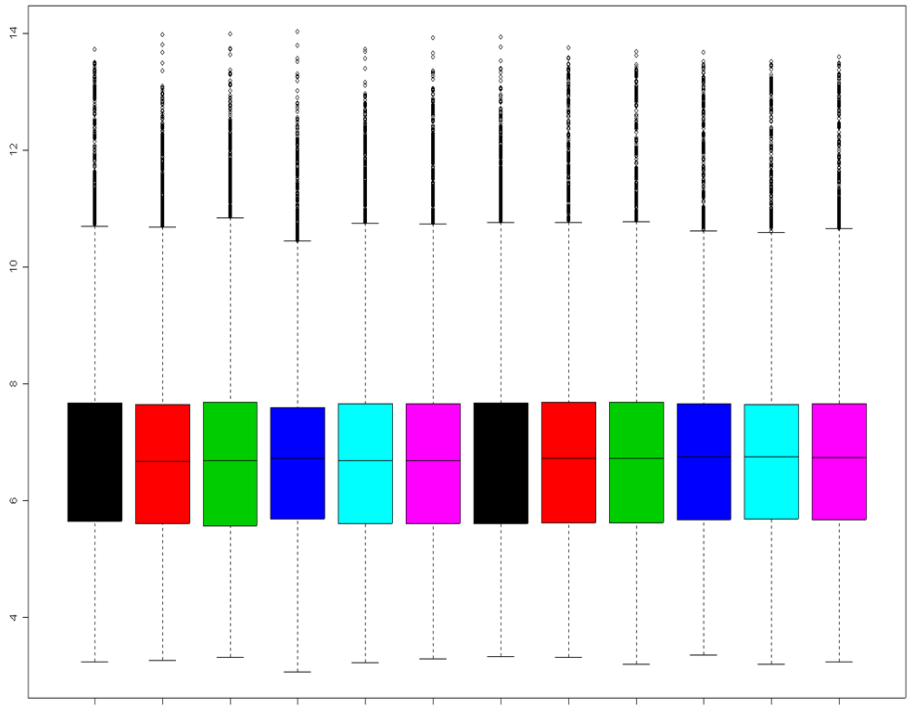
- Ajustar las mediciones de los arreglos para que esten en la misma escala
- Poder comparar las muestras



Procesando los Datos

Normalización

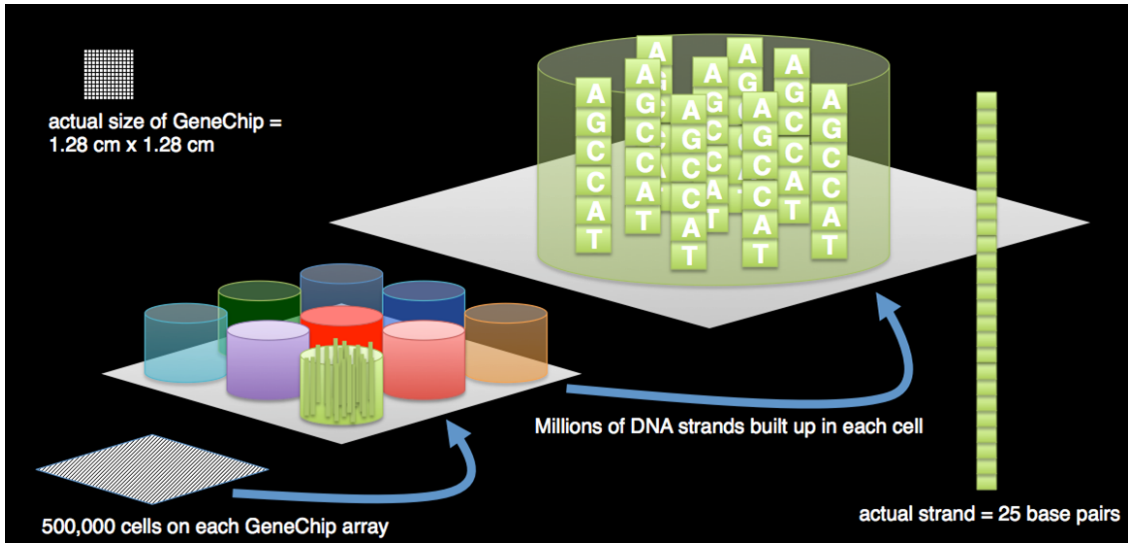
- Ajustar las mediciones de los arreglos para que esten en la misma escala
- Poder comparar las muestras



Procesando los Datos

Resumen

- Estimar el promedio de la intensidad por grupos de sondas



En R

```
e <- rma(ab)
```

Datos de Microarreglos

- Hay varias formas de leer los datos de microarreglos
- Ahora vamos a usar el paquete `oligo`

```
detach("package:affy")
library(oligo)
```

- Repetir los pasos anteriores

```
basedir <- paste0(wd,"/celfiles")
setwd(basedir)
tab <- read.delim("sampleinfo.txt",check.names=FALSE,
   as.is=TRUE)
```

- Revisar que tenemos todos los archivos

```
fns <- list.celfiles(listGzipped=TRUE)
fns %in% tab[,1]
```

Datos de Microarreglos

- Ingresar la información acerca de las muestras

```
pd <- as(tab, "AnnotatedDataFrame")
```

- Leer los archivos `CEL`

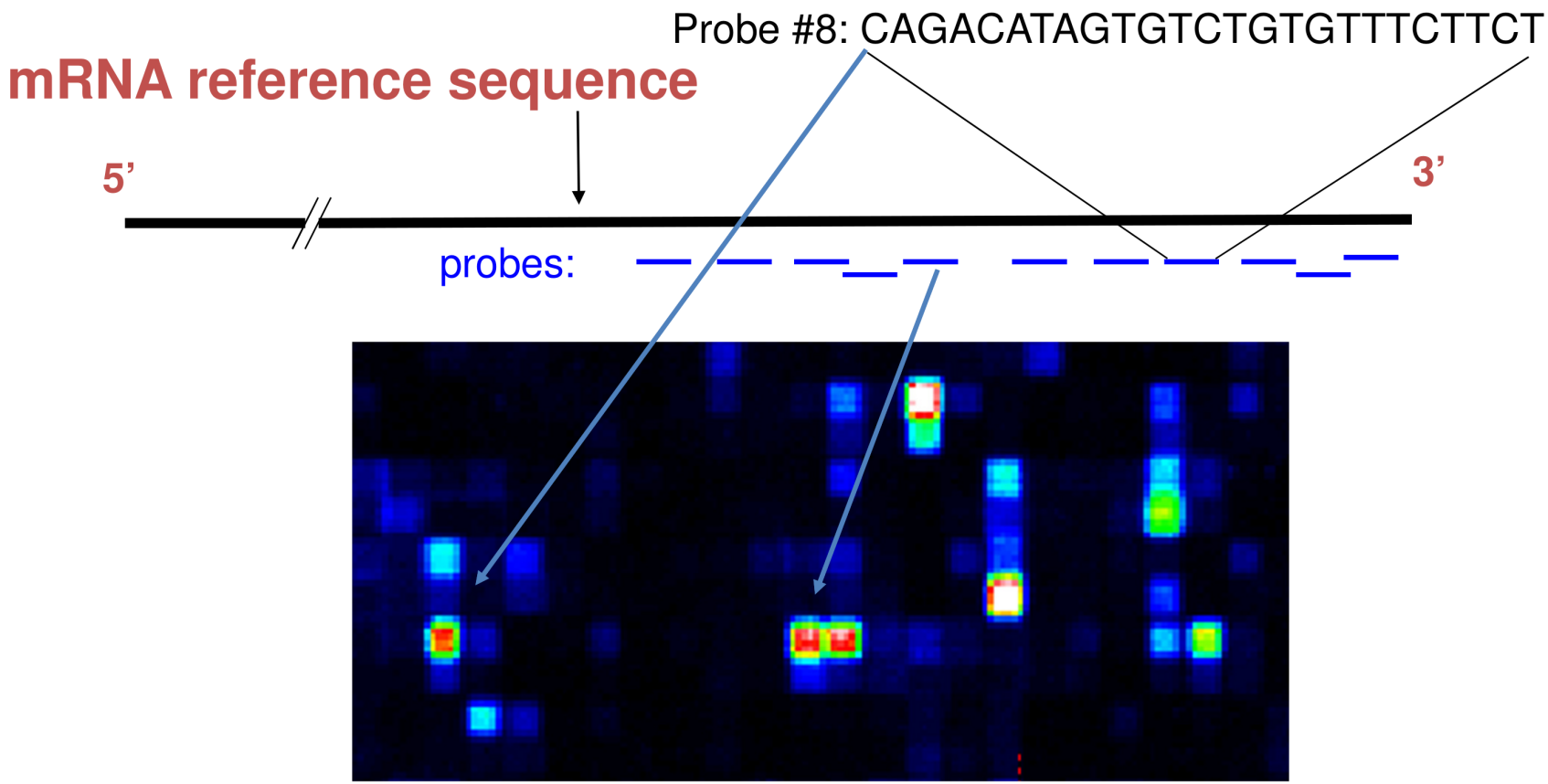
```
efs <- read.celfiles(filenames=tab[,1],
                     phenoData=pd,sampleNames=sampleNames(pd))
```

- Procesar los datos

```
setwd(wd)
e <- rma(efs)
```

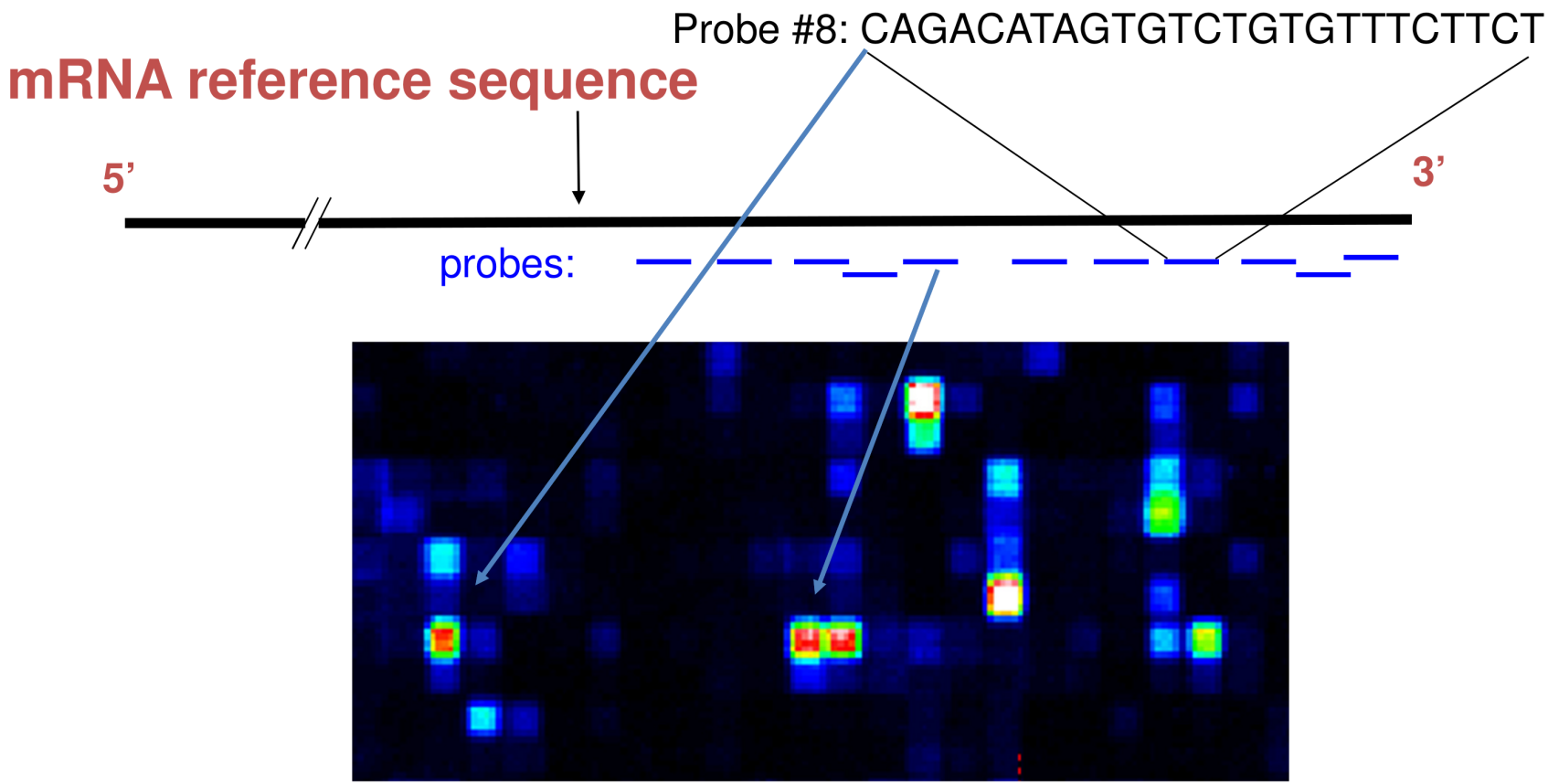
Anotación

- ¿Dónde quedaron los genes?



Anotación

- Anotación: que sondas corresponden a que genes



Anotación

- Para este ejemplo vamos a usar los datos de `maPooling`
- Descarga el archivo `maPooling.RData` en to working directory `https://dl.dropboxusercontent.com/u/21912429/CdeC/maPooling.RData`
- Los archivos `RData` son archivos de R para guardar datos

```
library(Biobase)

load("maPooling.RData")
e <- maPooling
head(rownames(e))
```

Anotación

- Plataforma

  `annotation(e)`
- Affymetrix Rat Expression Set 230 (`rae230a`)

Anotación

- Paquetes para anotación

```
library(rae230a.db)
library(AnnotationDbi)
```

- Campos disponibles

```
columns(rae230a.db)
```

- `Keys`: campos que se pueden como palabras claves

```
keytypes(rae230a.db)
```

- Por ejemplos, podemos usar los nombres de las sondas para acceder a los otros campos de la anotación

```
head(keys(rae230a.db, keytype="PROBEID"))
```

- En nuestas muestras

```
head(rownames(e))
```

Anotación

Nombres de los Genes

- Ensembl (European Bioinformatics Institute & Wellcome Trust Sanger Institute)
- Entrez (National Center for Biotechnology Information)
- Symbol (Human Genome Organisation)

```
res <- select(rae230a.db, keys=rownames(e),
              columns=c("ENTREZID","ENSEMBL","SYMBOL"),
              keytype="PROBEID")

head(res)
```

Anotación

- Vamos a agregar esta información a nuestras muestras

```
idx <- match(rownames(e), res$PROBEID)
head(rownames(e))
head(res$PROBEID,7)
fData(e) <- res[idx,]
head(fData(e),10)
```

- Vamos a asegurarnos que los nombres corresponden

```
all.equal(fData(e)$PROBEID, rownames(e))
```

GEO

- Gene Expression Omnibus (GEO)
- Repositorio de datos genómicos del National Center for Biotechnology Information (NCBI)
- Alrededor del 90% de los datos son estudios de expresión genética
- Los datos tienen 2 identificiones principalmente:
  - GEO Sample (GSM)
  - GEO Series (GSE)
- Vamos a usar el paquete `GEOquery` para descargar datos de GEO directamente a R

GEO

- Datos procesados

```
library(GEOquery)
gse <- getGEO("GSE21653", GSEMatrix=TRUE)
show(gse)
```

- Archivos sin procesar: si los archivos `.CEL` estan disponibles los podemos descargart usando la funcion `getGEOSuppFiles`
- Como argumento para esta funcion usamos una identificacion de GEO
- Esta función crea un folder en el working directory para guardar los archivos sin procesar

```
filePaths = getGEOSuppFiles('GSE21653')
filePaths
```