Expresión diferencial (marginal)

Guillermo Ayala Gallego

2024-04-09

Table of contents

El problema

- ¿Hay diferencias de la expresión de genes entre dos (o más) grupos distintos?
- Dicho de otro modo, la expresión de un gen es distinta bajo distintos tratamientos.
- Se adopta la aproximación gen-a-gen, esto es, de momento solamente buscamos genes que se expresan diferencialmente sin atender al comportamiento conjunto que puedan tener.

Log fold-change

El logaritmo del cociente de las medias

- Queremos comparar dos grupos.
- x_{ij} (y_{ij}) es la expresión de la i-ésima característica en la j-ésima muestra del primer grupo (segundo grupo).
- Tomamos los logaritmos (en base 2 o `log2`) de los valores originales:

$$u_{ij} = \log_2(x_{ij}),\ v_{ij} = \log_2(y_{ij}).$$

- Para cada gen, tendremos las expresiones medias para la i-ésima característica

$$\bar{x}_{i\cdot} = \sum_{j=1}^{n_1} \frac{x_{ij}}{n_1}$$

  $\bar{y}_{i\cdot}$, $\bar{u}_{i\cdot}$, $\bar{v}_{i\cdot}$.
- ¿Qué se entiende por **fold-change**?
- Dos son las interpretaciones **distintas** de este valor.

$$FC_i^{(1)} = \frac{\bar{x}_{i\cdot}}{\bar{y}_{i\cdot}}.$$

  y

$$FC_i^{(2)} = \bar{u}_{i\cdot} - \bar{v}_{i\cdot},$$

- El log-fold change es el logaritmo en base 2 de los fold-change que acabamos de definir.
- Si el cociente anterior es mayor que `c` entonces diríamos que el gen se expresa de un modo diferencial en ambos grupos.
- Tenemos una expresión diferencial del gen i si

$$\left|\log_2\left(FC_i\right)\right| \geq c$$

Comparando dos grupos y un solo gen

Dos grupos

- Utilizamos los datos gse21942.

```
data(gse21942,package="tamidata")
```

- Nos fijamos en un gen.

```
library(Biobase)
```

```
Loading required package: BiocGenerics
```

```
Attaching package: 'BiocGenerics'
```

```
The following objects are masked from 'package:stats':

    IQR, mad, sd, var, xtabs
```

```
The following objects are masked from 'package:base':

    anyDuplicated, aperm, append, as.data.frame, basename, cbind,
    colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
    get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
    match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
    Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
    table, tapply, union, unique, unsplit, which.max, which.min
```

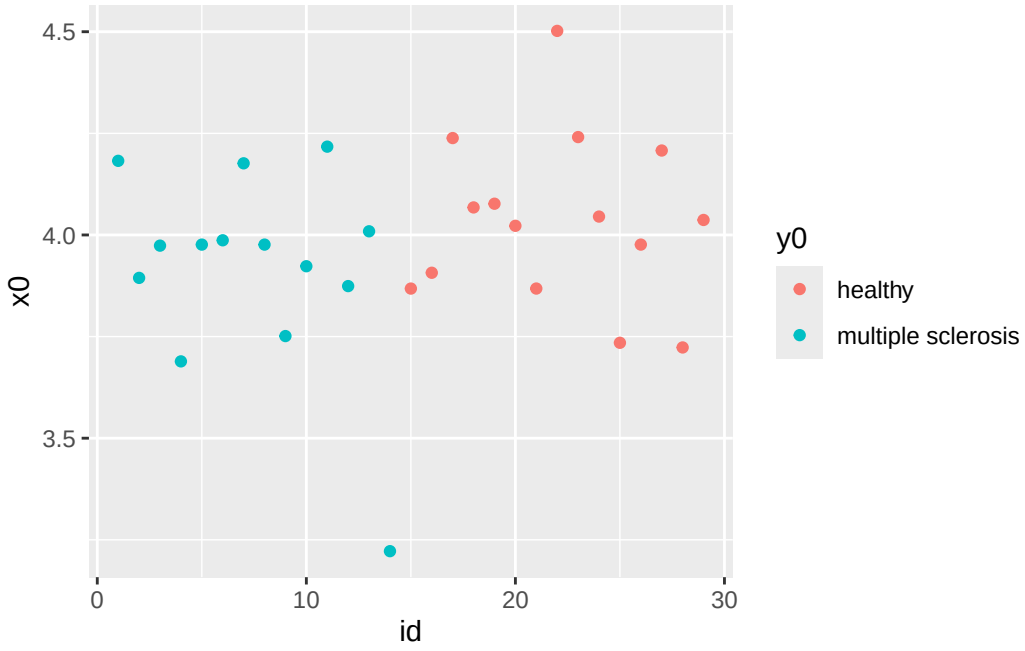
```
Welcome to Bioconductor

    Vignettes contain introductory material; view with
    'browseVignettes()'. To cite Bioconductor, see
    'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
x0 = exprs(gse21942)[2058,]
y0 = pData(gse21942)[,"FactorValue..DISEASE.STATE."]
```

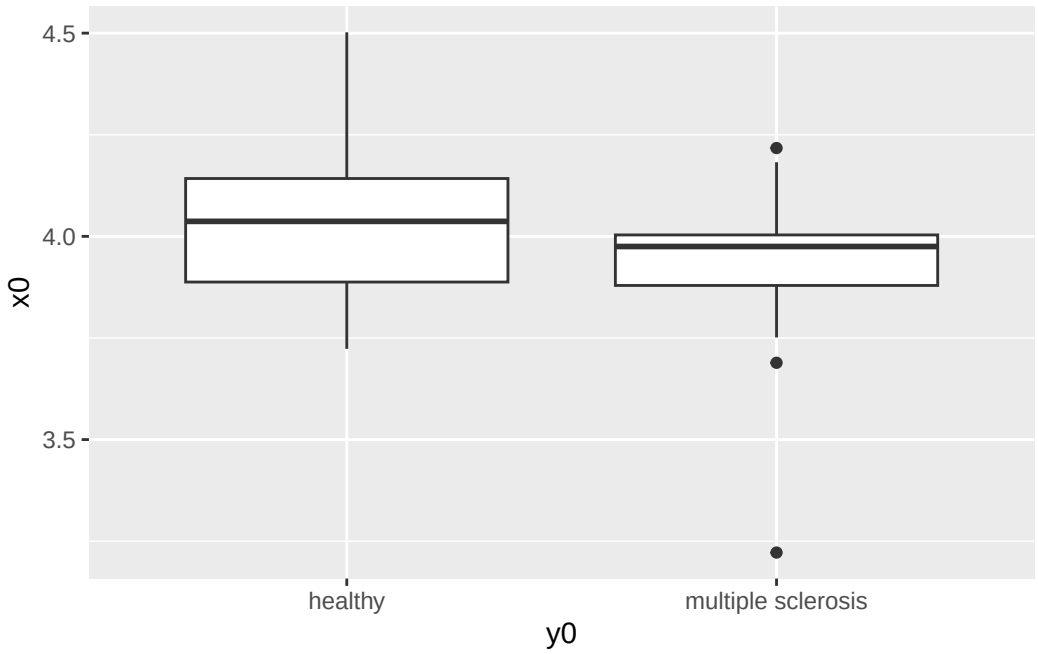
Perfil de expresión

```
pacman::p_load(ggplot2)
df = data.frame(id=1:length(x0),x0,y0)
ggplot(df,aes(x=id,y=x0)) + geom_point(aes(colour=y0))
```



Un diagrama de cajas

```
df = data.frame(id=1:length(x0),x0,y0)
ggplot(df,aes(x=y0,y=x0)) + geom_boxplot()
```



Comparamos dos condiciones

- Tenemos datos relativos a niveles de expresión bajo dos condiciones experimentales.
- Denotamos por X el nivel de expresión aleatorio que observamos bajo la primera condición y por Y lo mismo pero con la segunda condición.
- $X \sim N(\mu_X, \sigma_X^2)$
- $Y \sim N(\mu_Y, \sigma_Y^2)$
- Nos planteamos el contraste

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X \neq \mu_Y.$$

t-test con varianzas iguales

- Calculamos

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

con

$$S_p = \hat{\sigma}^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

siendo

$$S_X^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}.$$

y S_Y^2 lo mismo sustituyendo X_i por Y_i.

- Rechazamos la hipótesis nula cuando

$$|T| > t_{\nu,1-\alpha/2}$$

- El p-valor viene dado por

$$p = P(|T| \geq t_0)$$

siendo t_0 el valor observado de T en cada caso.

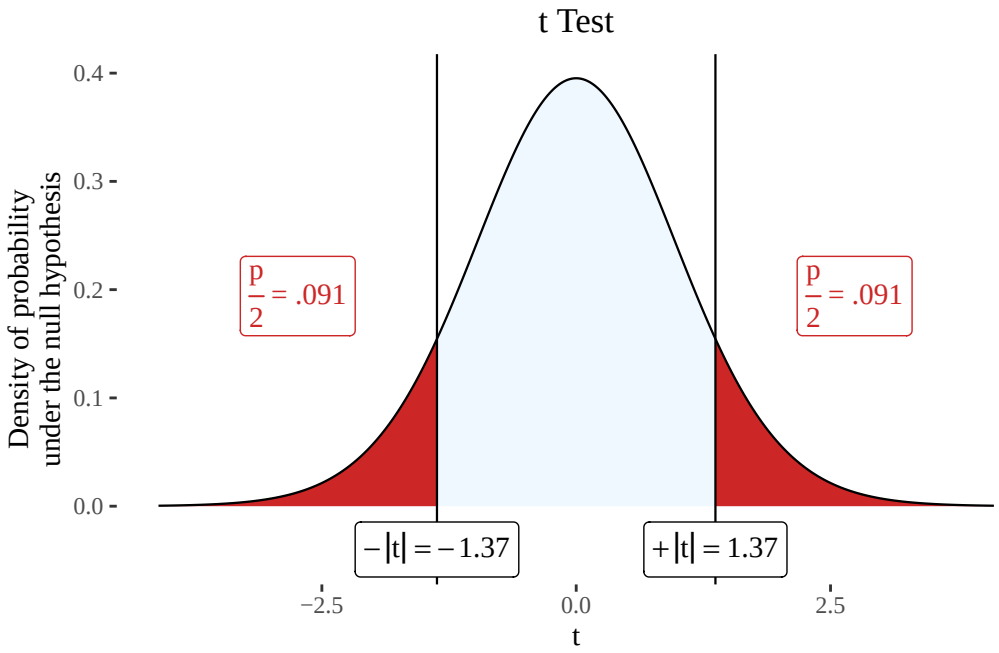
t.test

```
t.test(x0 ~ y0,var.equal=TRUE)
```

```
    Two Sample t-test

data:  x0 by y0
t = 1.3685, df = 27, p-value = 0.1824
alternative hypothesis: true difference in means between group healthy and group multiple sc
95 percent confidence interval:
 -0.0581131  0.2908712
sample estimates:
           mean in group healthy mean in group multiple sclerosis
                        4.034384                          3.918005
```

...



Comparación simultánea de todos los genes

Hacemos un t-test para todos los genes

```
library(genefilter)
tt = genefilter::rowttests(gse21942,pData(gse21942)[,"FactorValue..DISEASE.STATE."])
```

```
head(tt)
```

```
            statistic          dm     p.value
1007_s_at -2.29869931 -0.15981906 0.029492008
1053_at    3.44084102  0.20940361 0.001901206
117_at    -0.08505071 -0.01110444 0.932848609
121_at    -0.53362792 -0.02274832 0.597965094
1255_g_at -1.01536731 -0.04339509 0.318943716
1294_at   -1.05030996 -0.07873761 0.302886126
```

¿Cuándo tests mostrarían una expresión diferencial entre ambos grupos?

```r
table(tt$p.value < 0.01)
```

```
FALSE  TRUE
44014 10661
```

- Como vemos tenemos demasiados genes que muestran expresión diferencial.
- Y no parece muy razonable esto.
- El nivel de significación α es una cota superior del error tipo I cuando realizamos **un solo test**.
- Pero no estamos aplicando un solo test.
- Estamos rechazando muchas hipótesis nulas (no diferencia entre grupos para cada gen) cuando no deberíamos de hacerlo.