9.3 Normalización - Continuación

Hemos visto que la normalización es un proceso para identificar y eliminar errores sistemáticos que no se deban a diferencias reales entre el RNA de los distintos grupos de tejidos a comparar.

- Las fuentes de estos errores sistemáticos se pueden deber a una o más de las siguientes razones
    - Diferente eficiencia en la incorporación de tintes
    - Diferencias en la cantidad de mRNA
    - Diferencias en los parámetros de escaneo
    - Diferencias entre los grupos de cada aguja
    - Efectos espaciales
    - Efectos debido al plato
- Estas variaciones sistemáticas se reflejan en la dependencia de los cocientes en:
    - La intensidad total de fluorescencia (A)
    - La heterogeneidad espacial
    - La aguja
    - El plato

Existen diferentes propuestas respecto de qué genes utilizar en las normalizaciones.

- Todos los genes del arreglo.
- Genes cuya expresión es constante.
- Controles agregados -spiked in- . Por ejemplo genes de plantas.
- Serie de titulaciones de DNA genómico (Genomic DNA titration series).
- Un conjunto con rango invariante (Rank invariant set).

¿Cuáles utilizar? Depende del tipo de experimento. ¿Qué tipo de propuesta utilizaría para los siguientes experimentos?

• Se comparan muestras en las cuales la mayoría de los genes muestran distinto grado de diferenciación (por ejemplo, distintas etapas de crecimiento de la célula)

• Se comparan muestras en las cuales sólo una fracción pequeña de los genes está alterada (por ejemplo, pérdida de la actividad de un gen en un camino metabólico alterado)

9.3.1 Tipos de normalizaciones.

9.3.1.1 Notación

Cuando la normalización está basada en los

- log-ratios ($\log_2$ R/G) nos referimos a
  - *normalización de 2 canales*

Cuando la normalización está basada en cada una de

- $\log_2$ R, $\log_2$ G nos referimos a
  - *normalización de un canal* (canales por separado)

9.3.1.2 Objetivos

Los tipos de métodos de normalización a utilizar dependen del tipo de preguntas que se desea responder.

1) Determinar los **niveles de expresión relativos** de cada gen entre dos tipos de muestras (tejido sano vs. tejido patológico)

2) Determinar **cuáles son los genes que se han expresado** en cada una de las muestras o en diferentes tiempos .

El objetivo principal de la *normalización de dos canales* es el ajuste de sesgos *dentro* de cada arreglo. Es necesario para responder al primer problema.

Para responder al segundo problema es necesario realizar una normalización de los dos canales por separado. Además las normalizaciones de un canal son necesarias para que el análisis basado en microarreglos de un canal sea una opción válida para los investigadores.

El siguiente ejemplo muestra la necesidad de realizar un análisis de los canales por separado y por lo tanto también su normalización, en microarrays de dos canales.

Consideremos un ejemplo en el que se comparan animales jóvenes y viejos para los mutantes y wild-type

Nombre del Archivo	Cy3	Cy5
Archivo1	wt.joven	wt.viejo
Archivo2	wt.viejo	wt.joven
Archivo3	mu.joven	mu.viejo
Archivo4	mu.viejo	mu.joven

Cada arreglo de este experimento realiza una comparación directa entre el RNA de joven-viejo. Pero no hay arreglos que comparen mutantes con wild-type. Este es un ejemplo de diseño no conectado en el cual no hay arreglos ligando mutantes con wild-type.

No es posible realizar una comparación entre mutantes y wild-type utilizando log-ratios únicamente. Es necesario analizar las intensidades de los canales rojo y verde por separado. Esto es analizar log (R) y log(G) en vez de log -ratios.

9.3.2 Métodos de normalización

Dos canales

Ya hemos visto cómo realizar las siguientes normalizaciones de dos canales utilizando el paquete marray

- Global: Mediana
- Local, dependientes A: Loess -Print-tip

La siguiente es una nueva propuesta que propone un diseño de replicaciones de algunos genes en distintos bloques para estimar el efecto de cada bloque como aditivo de fila y columna, lo denominan:

- Semilinear in-slide model (FAN et al 2005)

$$Y_{gi} = \alpha_g + \beta_{r_{gi}} + \gamma_{c_{gi}} + m(X_{gi}) + \varepsilon_{gi}, \quad (1)$$

donde

$$Y_{gi} = \log_2 \frac{R_{gi}}{G_{gi}}, \qquad X_{gi} = \frac{1}{2}\log_2(R_{gi}G_{gi}).$$

α_g es el efecto del tratamiento asociado con el gen g

r_{gi} fila del bloque al que pertenece el gen g de la iésima replicación

c_{gi} columna del bloque al que pertenece el gen g de la iésima replicación

β_r es el efecto fila y γ_c es el efecto columna, con las restricciones:

$$\sum_{i=1}^{r} \beta_i = 0 \quad \text{y} \quad \sum_{j=1}^{c} \gamma_j = 0$$

$m(.)$ es una función suave que representa el efecto de la intensidad

Si el microarreglo está formado por 8x4 bloques r = 8 y c = 4

Luego de hallar los estimadores $\hat{\beta}_r$, $\hat{\gamma}_c$ y $\hat{m}(.)$ para el modelo (1), la normalización consiste en calcular

$$Y_g^* = Y_g - \widehat{\beta}_{r_g} - \widehat{\gamma}_{c_g} - \widehat{m}(X_g)$$

De acuerdo con el modelo (1) los valores normalizados son:

$$Y_g^* \approx \alpha_g + \varepsilon_g$$

en los cuales se ha eliminado los efectos confundentes.

Los autores proponen que el método puede utilizarse también en microarrays de un canal si se toman los tratamientos y controles como los canales rojo y verde.

Un canal

- ANOVA (Kerr et al 2000, Wolfinger et al 2001)
- Quantile normalization (Bolstand et al 2003)
- VSN (Huber et al 2002, Durbin et al 2002, 2003)

9.4 Normalización de los canales por separado

El paquete marray no realiza este tipo de normalizaciones, utilizaremos el paquete **`limma`**.

Recordemos que **`beta7norm`** contiene las intensidades normalizadas por Print-tip-loess

```
> beta7norm <- maNorm(beta7, norm= "p")
```

Para seguir con los procedimientos de **`limma`** tendremos que transformar los datos de **`marray`** crudos o normalizados mediante la función **`as(,)`** .

```
> library(convert)#para usar la función as(,)

#datos sin normalizar
> beta7.l <- as(beta7,"RGList")

#datos normalizados
> beta7.p <- as(beta7norm,"MAList")
```

Observación: Un aspecto a tener en cuenta antes de realizar la normalización entre arreglos es cómo se ha manejado la corrección por back ground para evitar los valores faltantes en los log-ratios, que puedan surgir de valores corregidos cero o negativos. La función **`backgroundCorrect()`** da una cantidad de opciones útiles

9.4.1 Normalización por cuantiles-Quantile Normalization

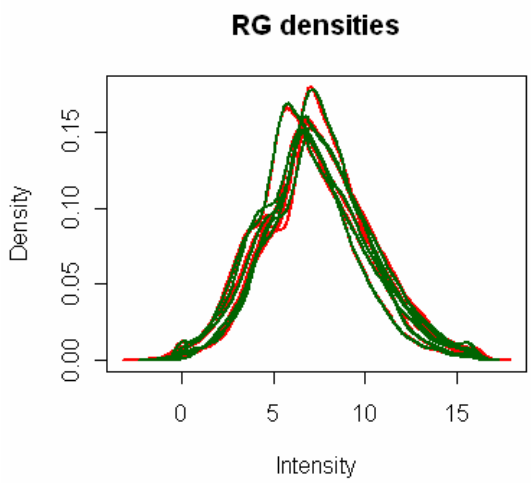
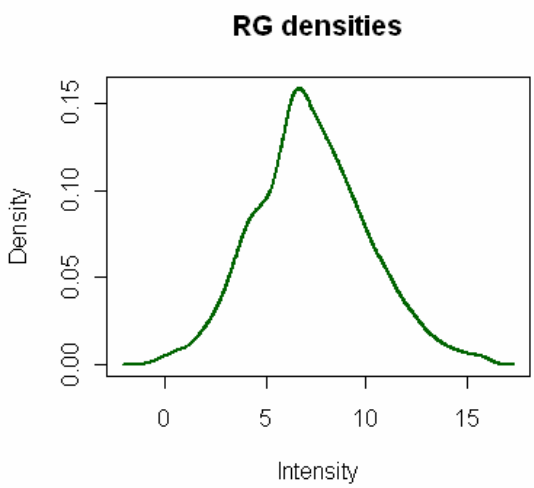
Un enfoque para realizar normalizaciones para cada canal está basado en las funciones **`normalizeWithinArrays`** y **`normalizeBetweenArrays`** del paquete **`limma`**. La primera de las funciones provee una normalización *dentro de cada arreglo* similar a la del marray y la otra *entre arreglos* permite que cada canal tenga una distribución similar entre los arreglos.

```
> beta7.pq <- normalizeBetweenArrays(beta7.p,method="quantile")
```

Mediante el método de normalización por cuantiles, las distribuciones de las intensidades de los canales rojo y verde de todos los arreglos quedan idénticas. Mediante

```
> plotDensities(beta7.p)
> plotDensities(beta7.pq)
```

obtenemos las densidades de las intensidades de los canales rojo y verde antes y después de la normalización por cuantiles.

	
Canales verdes y rojos por separado. Normalización por Print-tip	Canales verdes y rojos por separado. Normalización por Print-tip + quantile

¿En que consiste la normalización por cuantiles?

En transformar los valores de las intensidades en cada canal y en cada muestra, sin cambiarles el orden, de manera que resulten con idéntica distribución.

Se ordenan cada uno de los microarreglos-color en orden creciente. Se toma el promedio o mediana por fila. Se reemplaza la media ó la mediana en todos las columnas y se vuelve al orden original. De esta manera el valor máximo es el mismo en todos los arreglos-color, ..., el valor mínimo es el mismo.

9.4.2 Transformaciones estabilizadoras de la varianza

Otro enfoque de normalización de canales por separado es el "*variance stabilizing transformation*" **`vsn`**. Es una opción de la función **`normalizeBetweenArrays.`** Consiste en ajustar a los datos dados en la matriz x_{ki} (k filas, i columnas -arrays-) la siguiente transformación normalizadora

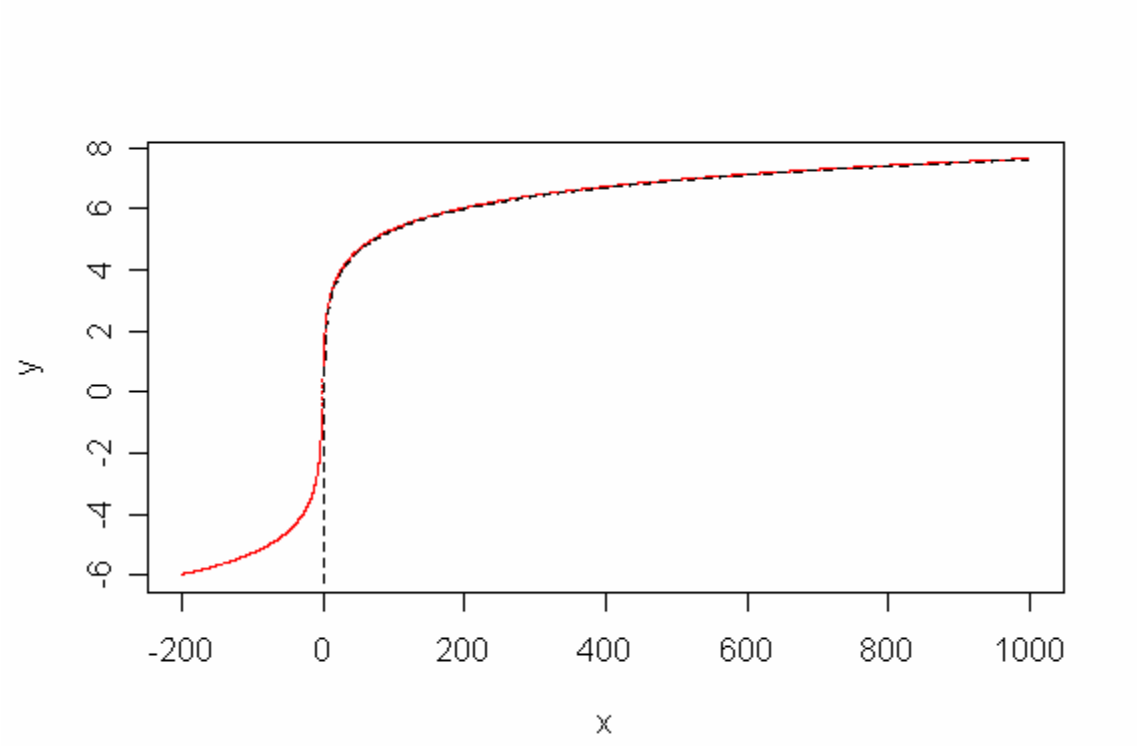
$$x_{ki} \rightarrow h(x_{ki}) = g \log(\frac{x_{ki} - a_i}{b_i})$$

donde glog(x) = $\log(x + \sqrt{x^2 + 1})$ es llamado logaritmo generalizado: es una función

similar al logaritmo natural para valores grandes (comparados con el ruido del background) pero es menos empinada para valores pequeños (ver figura) .

glog (x) ≈ log (x) + log(2)

```
> x <- seq(-200,1000,0.1)
> y <- log(x+ sqrt(x*x+1))
> plot(x,y, pch='.',col="red")
> z <- seq(0.001,1000)
> w <- log(z)+log(2)
> lines(z,w,lty=2)
```



Las diferencias entre los valores transformados son los log-ratios generalizados. Son estimadores comprimidos (shrinkage estimators) del logaritmo natural del cociente de las intensiddes (fold change).

Las siguientes instrucciones

```
> e2 <- vsn(e1)
> M <- exprs(e2)[, i] - exprs(e2)[, j]
```

producen un vector de log-ratios generalizados entre las muestras i y j. Observemos que

estos están en base e. Para cambiarlos a base 2 deben dividirse por log(2). Para obtener el fold change (en unidades originales) es necesario exponenciar

```
> M.base2 <- M/log(2)
> fold.change <- exp(M)
```

Los **log-ratios generalizados** pueden interpretarse como estimadores shrinkage por que **siempre son menores que los log-ratios puros** (naive), la igualdad se alcanza en forma asintótica a medida que aumentan las intensidades. Tienen la ventaja de no sufrir una divergencia de la varianza como le ocurre a los log-ratios naive en intensidades bajas y se mantienen bien definidos y con significado estadístico cuendo los datos se acercan a cero e incluso son negativos.

La transformación también es una opción de la función **normalizeBetweenArrays:**

```
 > beta7.vsn <-
normalizeBetweenArrays(as(beta7,"RGList"),method="vsn")
Loading required package: vsn
vsn: 23184 x 12 matrix (1 stratum). 100% done.

> summary(beta7.vsn)

              Length Class      Mode
Rb            139104 -none-     numeric
Gb            139104 -none-     numeric
weights       139104 -none-     numeric
printer            4 -none-     list
genes              6 data.frame list
targets            7 data.frame list
notes              1 -none-     character
M             139104 -none-     numeric
A             139104 -none-     numeric
preprocessing      3 -none-     list
```

Referencias:

Bolstad, B (2001) *Probe Level Quantile Normalization of High Density Oligonucleotide Array Data*. Unpublished manuscript

Bolstad, B. M., Irizarry R. A., Astrand, M, and Speed, T. P. (2003) *A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance.* Bioinformatics 19(2) , pp 185-193. http://www.stat.berkeley.edu/~bolstad/normalize/normalize.html

Jianqing FAN, Heng PENG, and Tao HUANG (2006) "Semilinear High-Dimensional Model for Normalization of Microarray Data: A Theoretical Analysis and Partial Consistency". JASA Theory and Methods, Vol. 100, No. 471, 781-813

W. Huber, A. von Heydebreck, H. S¨ultmann, A. Poustka, and M. Vingron. Variance stablization applied to microarray data calibration and to quantification of differential expression (2002). Bioinformatics, 18:S96–S104.

W. Huber, A. von Heydebreck, andM. Vingron. Analysis of microarray gene expression data. To appear in the Handbook of Statistical Genetics, (2003). Eds.: D. J. Balding, M. Bishop, C. Cannings. John Wiley & Sons, Inc.

W. Huber, A. von Heydebreck, H. S¨ultmann, A. Poustka, and M. Vingron. Parameter estimation for the calibration and variance stabilization of microarray data (2003). Statistical Applications in Genetics and Molecular Biology, Vol. 2: No. 1, Article 3. http://www.bepress.com/sagmb/vol2/iss1/art3 2

Kerr, MK, Martin, M, and Churchill, GA (2000). Analysis of variance for gene expression microarrays. J. Comput. Biol. 7, 819–837

David M. Rocke and Blythe Durbin. A model for measurement error for gene expression analysis. Journal of Computational Biology, 8:557–569, 2001.

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. J. Comp. Biol 8: 625-637