9. Medidas de expresión en chips de alta densidad

El preprocesamiento de datos de chips de alta densidad de Affymetrix consiste en tres pasos:

- Corrección por background
- Normalización entre arreglos
- Obtención de una medida resumen del probe set, para cada gen

Describiremos brevemente los procedimientos más utilizados y con más detalle la corrección del background dado por el modelo de convolución exponencial - normal.

9.1 Notación

i = 1,…, I (cantidad de chips, desde 1 a cientos)
j= 1,…, J (cantidad de probes en cada probe set, generalmente 11 ó 20)
n = 1,…, N (cantidad de genes = probe sets, entre 8 000 y 35 000)

y

PMijn = intensidad de un "perfect match"
MMijn= intensidad de un "mismatch"
E_{ijn} = intensidad corregida

en el chip i, probe j, gen n

9.2 Métodos

En las tablas siguientes resumimos los procedimientos más utilizados para corrección del fondo, normalización y obtención de índices del nivel de expresión basados en los probe sets. El gen n está representado por el probe set con J probes.

Método	**Corrección del Background**	**Normalización entre arreglos**
MAS 5	$E_{ijn} = PM_{ijn} - MM_{ijn}^*$ donde MM_{ijn}^* se elije de manera que E_{ijn} sea no negativo	**Método de escala**: A nivel del probe set, sobre la medida resumen (o anivel de probe). Un arreglo fijo i*(baseline), $\bar{x}_{i^*}$ =media podada de las intensidades del arreglo i* $\bar{x}_i$ =media podada de las intensidades del arreglo i $\beta_i = \bar{x}_{i^*} / \bar{x}_i$ es el factor multiplicativo para todas las intensidades x_{in} del arreglo i: $x'_{in} = \beta_i x_{in}$ Es "casi" equivalente a ajustar una recta por el origen a los pares (x_i, x^*_i) y reemplazar x_i por su valor predicho por la recta. **Métodos no lineales**: Ajustar un suavizado f (x) a los pares (x_i, x^*_i) y reemplazar x_i por $f(x_i)$
Model Based Expression Index (MBEI) - dChip	$E_{ijn} = PM_{ijn} - MM_{ijn}$	

Robust Multichip Analysis RMA	Ajusta el modelo a PM PM = Background ($N(\mu, \sigma^2)$)+ Señal (exponencial(λ)) E_{ijn} = E(Señal | PM_{ijn}) Tiene una expresión cerrada que veremos al final	**Normalización por cuantiles** **Loess cíclico**

El análisis basado en los modelos MBEI y RMA requieren de múltiples arreglos para la estimación de los parámetros de afinidad del probe.

Método	**Modelo**	**Resumen del Probe Set**
MAS 5	$\log_2(E_{ijn}) = \log_2(\theta_{in}) + \varepsilon_{ijn}$ θ_{in} : índice de expresión del chip i para el gen n	log(señal del probe set) = TukeyBiweight(log E_{ijn})
Model Based Expression Index (MBEI) - dChip	$E_{ijn} = \theta_{in}\Phi_{jn} + \varepsilon_{ijn}$ Φ_{jn} = efecto de afinidad del probe j del gen n ε_{ijn} Normales, estimación Máx. Veros.	θ_{in} índice de expresión del gen n del arreglo i
Robust Multichip Analysis RMA	$\log_2 E_{ijn} = e_{in} + a_{jn} + e_{ijn}$ a_{jn} efecto de afinidad del probe j del gen n en escala logarítmica Estimación por median polish.	e_{in} índice de expresión del gen n del arreglo i

Loess cíclico

Let X be a $p \times n$ matrix with columns representing arrays and rows probes or probesets.
log transform the data: $X \leftarrow \log X$
repeat
  for $i = 1$ to $n - 1$ **do**
    for $j = i + 1$ to n **do**
      for $k = 1$ to p **do**
        Compute $M_k = x_{ki} - x_{kj}$ and $A_k = \frac{1}{2}(x_{ki} + x_{kj})$
      end for
      fit a loess curve for M on A. Call this $\hat{f}$.
      for $k = 1$ to p **do**
        $\hat{M}_k = \hat{f}(A_K)$
        set $a_k = (M_k - \hat{M}_k)/n$
        $x_{ki} = x_{ki} + a_k$ and $x_{kj} = x_{ki} - a_k$
      end for
    end for
  end for
until convergence or the maximum number of iterations is reached
Revert to the original scale $X \leftarrow \exp(X)$

Table 2.4. Cyclic Loess Algorithm.

(Página 23 Gentleman 2005)

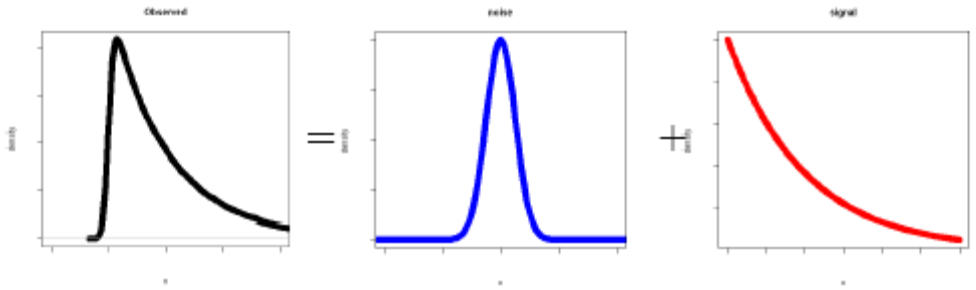
Terry Speed

RMA in slightly more detail
Background Correction

Ignore MM, fit model to PM

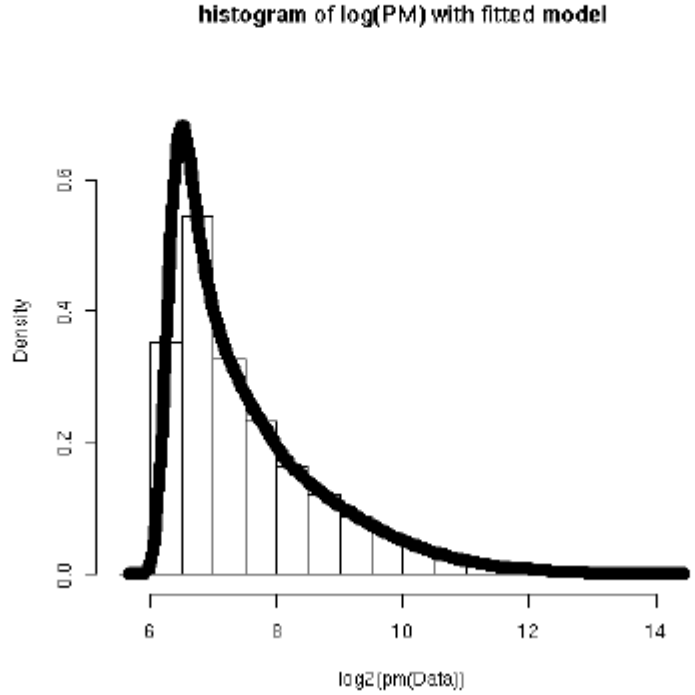
PM = Background + Signal

$N(0,\sigma^2)$ Exponential(α)

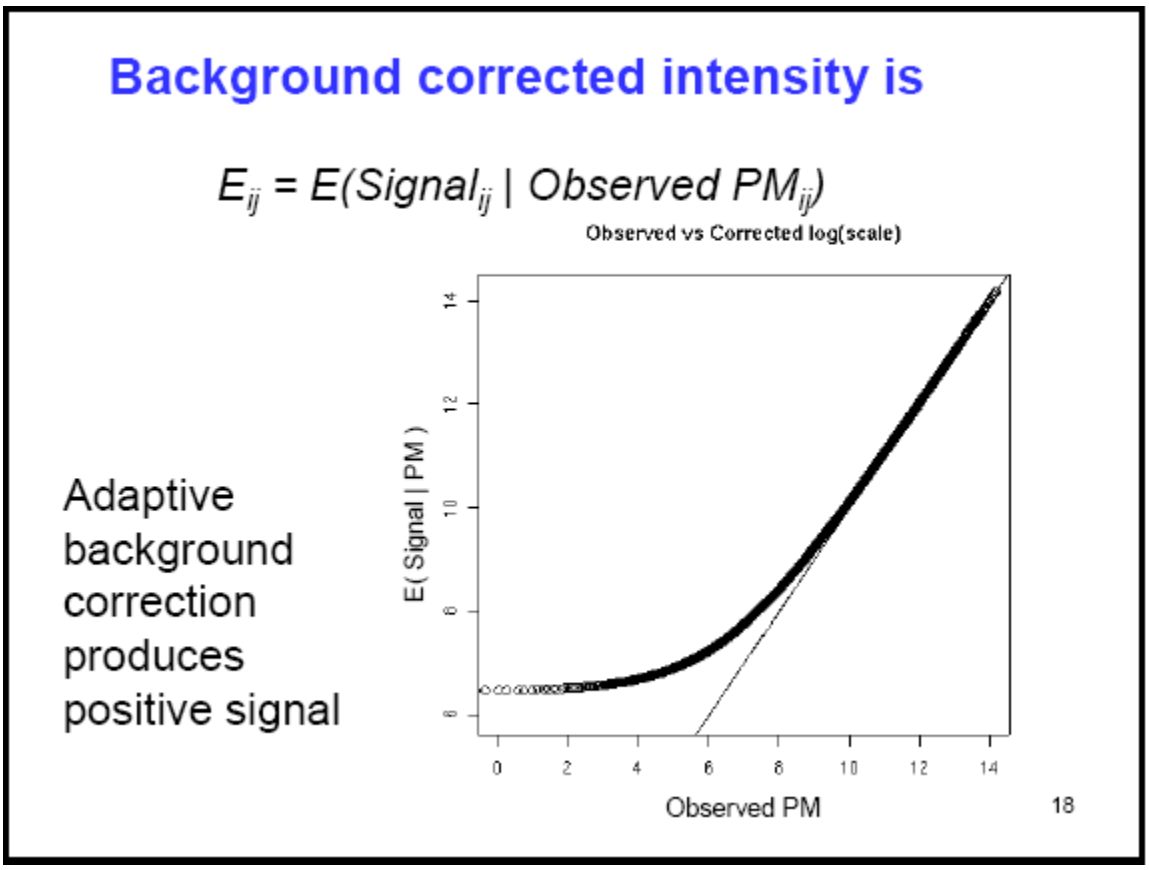


16

PM data on $\log_2$ scale: raw and fitted model



17



9.3 Corrección por background modelo de convolución normal exponencial

El metodo normexp se propuso originalmente como parte del algoritmo de RMA para los datos de microarreglos de Affymetrix (Bolstadt 2004). Para los datos microarry de dos colores, el método de corrección de fondo normexp fue introducido y comparado con otros métodos por Ritchie et al (2007). Una mejora en la estimación fue propuesta por Silver et al (2009).

Bolstadt propone el siguiente modelo para las intensidades observadas S:

$$S = X + Y,$$

donde X es la señal e Y es el fondo (background). Se supone que X tiene distribución exponencial de parámetro α ($X \sim e(\alpha)$) e $Y \sim N(\mu,\sigma^2)$, con X e Y independientes. Por otra parte, se supone que $Y > 0$ para evitar la producción de valores negativos. Por lo tanto, Y sigue una distribución normal truncada en cero.

Bajo este modelo las intensidades de las sondas corregidas por el fondo estarán dadas por $E(X| S = s)$

Bolstadt (2004), páginas 17-20, demuestra que

$$E(X|\ S=s) = a + b\frac{\varphi(a/b) - \varphi((x-a)/b)}{\Phi(a/b) + \Phi((x-a)/b) - 1}$$

donde $a = s - \mu - \alpha\,\sigma^2$ y $b=\sigma$, φ y Φ son respectvamente las funciones de densidad y de distribución de la Normal estándar.

y dice que en la mayoría de las aplicaciones $\varphi((x-a)/b)$ es despreciable y $\Phi((x-a)/b)$ es practicamente 1 (estos son términos que aparecen por truncar la Normal). Por lo tanto en la práctica será necesario calcular solo el primer término del numerador y el primer término del denominador.

Xie (2009) propone el mismo modelo normal exponencial para microarrays de Illumina y no restringe los valores del background a ser positivos, llegando a la expresión más simple:

$$E(X|\ S=s) = a + b\frac{\varphi(a/b)}{\Phi(a/b)}$$

Señalan que bajo el modelo los valores de fondo negativos pueden ocurrir con muy baja probabilidad de manera que los pueden ignorar.

El problema ahora se encuentra en la estimación de μ, α y σ^2.

Bioconductor estima de la siguiente manera:

- estima una moda global, $m0$, a partir de una estimación de densidades de las intensidades
- $\hat{\mu}$ = moda de las observaciones que se encuentran a la izquierda de $m0$
- utiliza los valores a la izquierda de $\hat{\mu}$ para estimar σ
- utiliza los valores a la derecha de $m0$ para estimar α de una exponencial

Referencias

- Algoritmos de AffymetrixMAS 5 o GCOS 1.0
- dChip http://www.dchip.org Li and Wong (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *PNAS* 98, 31-36.
- RMA (Robust Multichip Analysis) Irizarry *et al* (2003), Summaries of Affymetrix GeneChip probe level data. *NAR* 31(4):e15
- Bioinformatics and Computational Biology Solutions Using R and Bioconductor Editado por R. Gentleman, V. Carey, W. Huber, R. Irizarry, y S. Dudoit (2005). Springer.
- http://bmbolstad.com/Dissertation/Bolstad_2004_Dissertation.pdf

- Ritchie, M. E., Silver, J., Oshlack, A., Silver, J., Holmes, M., Diyagama, D., Holloway, A., and Smyth, G. K. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics* 23, 2700-2707. http://bioinformatics.oxfordjournals.org/cgi/reprint/23/20/2700
- Silver, J., Ritchie, M. E., and Smyth, G. K. (2009). Microarray background correction: maximum likelihood estimation for the normal-exponential convolution model. *Biostatistics* 10, 352-363 http://biostatistics.oxfordjournals.org/cgi/reprint/10/2/352
- Xie Y, Wang X, Story M: Statistical methods of background correction for Illumina BeadArray data. Bioinformatics 2009, 25:751-757