Appendix: Further details on the background correction models and their implementation

Matthew E. Ritchie, Jeremy Silver, Alicia Oshlack, Melissa Holmes,
Dileepa Diyagama, Andrew Holloway and Gordon K. Smyth

20 July 2007

This appendix provides further specifics on the models described briefly in the Methods section (2) of the paper *A comparison of background correction methods for two-colour microarrays.*

Kooperberg: Kooperberg *et al.* (2002) suggest an empirical Bayes model made up of a convolution of normal distributions to background adjust the signals. Observed foreground and background mean intensities (X_f and X_b) and their standard deviations (SD_f and SD_b), along with the number of foreground and background pixels (n_f and n_b) for each spot in a given channel are used in the model,

$$p(\mu|\sigma_b, \sigma_f, X_b, X_f) = \frac{\phi\left(\frac{X_f-\mu-X_b}{\sigma_d}\right)\Phi\left(\frac{(X_f-\mu)\sigma_b^2+X_b\sigma_f^2}{\sigma_f\sigma_b\sigma_d}\right)}{\sigma_d\int_0^\infty \Phi\left(\frac{X_f-v}{\sigma_f}\right)\phi\left(\frac{X_b-v}{\sigma_b}\right)dv} \quad (1)$$

where $\phi(.)$ is the density of the standard normal distribution, $\Phi(x) = \int_{-\infty}^{x}\phi(y)dy$ is the cumulative standard normal distribution, $\sigma_f = aSD_f/\sqrt{n_f}$, $\sigma_b = aSD_b/\sqrt{n_b}$, $\sigma_d = \sqrt{\sigma_b^2+\sigma_f^2}$ and a is a scaling factor. Numerical integration is applied to obtain the expected value of the true signal $E(\mu|X_b, X_f, \sigma_b, \sigma_f)$ in each channel for each spot. The a-values were estimated separately for each channel by regressing the observed background variability ($SD_b/\sqrt{n_b}$) on the empirical standard deviation of the background from the 3 (for spots on an outer row/column) or 4 nearest neighbour spots within a print-tip group. This model-based method avoids missing values and was demonstrated to reduce the high variability of low intensity log-ratios when used on data from a self-self hybridisation where there is no differential expression.

The function *kooperberg* was implemented in the *limma* package (Smyth, 2004) to adjust the foreground signals according to Equation 1. The R code was modified from Charles Kooperberg's S-Plus code (supplied in personal communication). This method was applied to GenePix data in this study, with the local mean estimate used for the background.

Edwards: A simpler background correction method is suggested in Edwards (2003), who adjusts the foreground intensities as follows:

$$R = \begin{cases} R_f - R_b & \text{if } R_f - R_b > \delta \\ \delta \exp[1 - (R_b + \delta)/R_f] & \text{otherwise} \end{cases}$$
$$G = \begin{cases} G_f - G_b & \text{if } G_f - G_b > \delta \\ \delta \exp[1 - (G_b + \delta)/G_f] & \text{otherwise} \end{cases} \tag{2}$$

In this model, subtraction of the background is done as usual when the difference between the foreground and background is larger than a threshold value δ, however when the difference is small or negative ($\leq \delta$), subtraction is replaced by a smooth monotonic function.

A value for the parameter δ was adaptively chosen from the data, as per the original code from David Edwards (supplied in personal communication). The quantile of the difference between foreground and background in each channel which was 10% above the number of negative background corrected values was chosen when negatives were present. If there were no negative values, the minimum value was used. Adjustments as per Equation 2 were made using GenePix data with local median estimates of the background for R_b and G_b by the *backgroundCorrect* function in *limma* with *method="edwards"*.

Normexp: The *normexp* convolution model which has been used to background correct Affymetrix data (Irizarry *et al.*, 2003) will be derived here. Further detail is given in Bolstad (2004) (Chapter 2) and McGee and Chen (2006). The motivation comes from looking at the distribution of the observed foreground signals (X) in each channel for a given array (Figure 1). Assume that the foreground X_f, true signal X and background signal X_b are additive ($X_f = X + X_b$), independent, and that X is exponentially distributed with mean α, and X_b is normally distributed with mean μ and standard deviation σ.

The joint density of X and X_b is

$$f(x, x_b; \mu, \sigma, \alpha) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_b - \mu)^2\right) \frac{1}{\alpha} \exp\left(-x/\alpha\right)$$

for $x > 0$. The joint density of X_f and X is therefore

$$f(x_f, x; \mu, \sigma, \alpha) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_f - x - \mu)^2\right) \frac{1}{\alpha} \exp\left(-x/\alpha\right).$$

Now

$$\begin{aligned} -\frac{1}{2\sigma^2}(x - x_f + \mu)^2 - x/\alpha &= -\frac{1}{2\sigma^2}\left(x^2 - 2(x_f - \mu - \sigma^2/\alpha)x + (x_f - \mu)^2\right) \\ = \quad -\frac{1}{2\sigma^2}\left(x - (x_f - \mu - \sigma^2/\alpha)\right)^2 &- \frac{1}{2\sigma^2}\left((x_f - \mu)^2 - (x_f - \mu - \sigma^2/\alpha)^2\right) \end{aligned}$$

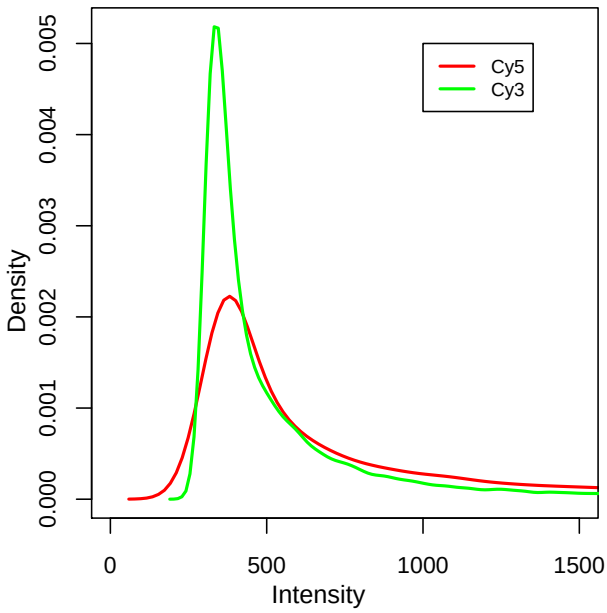

Figure 1: Smoothed histograms of the red (Cy5) and green (Cy3) foreground intensities for an array from the Mixture experiment.

$$
\begin{aligned}
&= -\frac{1}{2\sigma^2}\left(x - (x_f - \mu - \sigma^2/\alpha)\right)^2 - \frac{1}{2\sigma^2}\left(2(x_f-\mu)\sigma^2/\alpha) - (\sigma^2/\alpha)^2\right) \\
&= -\frac{1}{2\sigma^2}\left(x - (x_f - \mu - \sigma^2/\alpha)\right)^2 - (x_f-\mu)/\alpha + \frac{1}{2}\sigma^2/\alpha^2
\end{aligned}
$$

so we can re-write the joint density as

$$
f(x_f, x; \mu, \sigma, \alpha) = \frac{1}{\alpha}\exp\left(-(x_f-\mu)/\alpha + \frac{1}{2}\sigma^2/\alpha^2\right)\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_{x.x_f})^2\right)
$$

with $\mu_{x.x_f} = x_f - \mu - \sigma^2/\alpha$.
The marginal distribution of X_f arises from integrating with respect to x as follows

$$
\begin{aligned}
f(x_f; \mu, \sigma, \alpha) &= \frac{1}{\alpha}\exp\left(-(x_f-\mu)/\alpha + \frac{1}{2}\sigma^2/\alpha^2\right)\int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_{x.x_f})^2\right)dx \\
&= \frac{1}{\alpha}\exp\left(-(x_f-\mu)/\alpha + \frac{1}{2}\sigma^2/\alpha^2\right)\left(1 - \Phi(0; \mu_{x.x_f}, \sigma^2)\right) \qquad (3)
\end{aligned}
$$

where Φ is the cumulative normal distribution. The parameters α, μ and σ can be estimated via maximum likelihood (ML) using the log likelihood

$$
\log f(x_f; \mu, \sigma, \alpha) = -\log\alpha - (x_f-\mu)/\alpha + \frac{1}{2}\sigma^2/\alpha^2 + \log\left(1 - \Phi(0; \mu_{x.x_f}, \sigma^2)\right).
$$

The conditional distribution of X given X_f is the truncated normal distribution

$$f(x|x_f;\mu,\sigma,\alpha) = \frac{f(x,x_f;\mu,\sigma,\alpha)}{f(x_f;\mu,\sigma,\alpha)} = \frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_{x.x_f})^2\right)}{1-\Phi(0;\mu_{x.x_f},\sigma^2)}$$

for $x > 0$. Now

$$\frac{\partial \log f(x|x_f)}{\partial \mu_{x.x_f}} = \frac{1}{\sigma^2}(x-\mu_{x.x_f}) - \frac{\phi(0;\mu_{x.x_f},\sigma^2)}{1-\Phi(0;\mu_{x.x_f},\sigma^2)}$$

which gives

$$E(X|X_f = x_f) = \mu_{x.x_f} + \sigma^2 \frac{\phi(0;\mu_{x.x_f},\sigma^2)}{1-\Phi(0;\mu_{x.x_f},\sigma^2)}$$

where ϕ is the normal density function.

The first distinction between the convolution model in Irizarry *et al.* (2003) and our model, is the incorporation of the observed background in the calculations. For the ith spot we observe the foreground ($x_{f,i}$) and background ($x_{b,i}$) intensities. Assuming that $X_{b,i}|x_{b,i} \sim N(\mu_i, \sigma^2)$ and $X_i \sim \exp(\alpha)$, where $\mu_i = \beta + x_{b,i}$ and $\mu_{x.x_{f,i}} = x_i - \mu_i - \sigma^2/\alpha$, then the parameters α, β and σ need to be estimated in each channel, on each array.

A second distinction is that we use a saddle-point approximation to simplify the mathematical form of the likelihood function to make ML numerically feasible. Let $\tilde{f}$ be the saddle-point approximation to the density function f, also called the tilted Edgeworth expansion. Following Barndorff-Nielsen and Cox (1989), the saddle-point approximation to $f(x_f;\mu,\sigma,\alpha)$ can be written as

$$\log \tilde{f}(x_f;\mu,\sigma,\alpha) = -\frac{1}{2}\log\{2\pi K''(\tilde{\theta})\} - x_f\tilde{\theta} + K(\tilde{\theta})$$

where $K()$ is the cumulant generating function of the convolution and $\tilde{\theta}$ satisfies

$$K'(\tilde{\theta}) = x_f.$$

The cumulant generating function is the sum of the normal and exponential cumulant generating functions,

$$K(\theta) = \mu\theta + \frac{1}{2}\sigma^2\theta^2 - \log(1-\alpha\theta).$$

Write $\kappa_k = K^{(k)}(\tilde{\theta})$, $k = 0, 1, \ldots$, for the derivatives of the cumulant generating function at $\tilde{\theta}$. We use the second-order saddle-point approximation, which can be written as

$$\log \tilde{f}(x_f;\mu,\sigma,\alpha) = -\frac{1}{2}\log\{2\pi\kappa_2\} - x_f\tilde{\theta} + \kappa_0 - \frac{1}{8}\frac{\kappa_4}{\kappa_2^2} - \frac{5}{24}\frac{\kappa_3^2}{\kappa_2^3}. \quad (4)$$

The Nelder-Mead simplex algorithm was used to obtain ML estimates of $\hat{\alpha}, \hat{\beta}$ and $\hat{\sigma}$ from the saddle-point approximation to the log likelihood (Equation 4). These estimates are used to obtain adjusted signals according to

$$E(X_i|X_{f,i} = x_{f,i}) = \mu_{x.x_{f,i}} + \sigma^2 \frac{\phi(0; \mu_{x.x_{f,i}}, \sigma^2)}{1 - \Phi(0; \mu_{x.x_{f,i}}, \sigma^2)}. \tag{5}$$

Data can be corrected in this way using the *limma* function *backgroundCorrect* with *method="normexp"*.

Vsn: The variance stabilisation method of Huber *et al.* (2002) measures differential expression using a 'difference statistic'

$$\begin{aligned} \Delta h &= \operatorname{arcsinh}(z_R) - \operatorname{arcsinh}(z_G) \\ &= \log \frac{z_R + \sqrt{z_R^2+1}}{z_G + \sqrt{z_G^2+1}} \end{aligned} \tag{6}$$

where $z_R = a_R + b_R R$ and $z_G = a_G + b_G G$, with calibration parameters a_R, a_G, b_R and b_G estimated for each channel from an array in an experiment. Since the arcsinh function is defined for negative values, corrected signals (R, G) which are negative do not pose a problem. At high intensities, the arcsinh transform is equivalent to the regular log-ratio, whereas at low intensities it is close to the difference $z_R - z_G$. This method is implemented in the *vsn* software and can be accessed in *limma* using the function *normalizeBetweenArrays* by choosing the *method="vsn"* option. Note that the returned intensity and expression measures from this function are log base 2, to allow comparability with the other methods. In contrast to the other 7 methods in this study, vsn background corrects and normalises/calibrates the data together. For the other alternatives, a separate normalisation step is neccesary.

References

Barndorff-Nielsen, O.E. and Cox, D.R. (1989). *Asymptotic techniques for use in statistics.* Chapman and Hall, London.

Bolstad, B. M. (2004). *Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization.* Ph.D. thesis, Department of Biostatistics, University of California, Berkeley.

Edwards, D. (2003). Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics*, **19**(7), 825–833.

Huber, W. et al. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**(Suppl 1), S96–S104.

Irizarry, R. A. et al. (2003). Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2), 249–264.

Kooperberg, C. et al. (2002). Improved background correction for spotted DNA microarrays. *J Comput Biol*, **9**(1), 55–66.

McGee, M. and Chen, Z. (2006). Parameter estimation for the exponential-normal convolution model for background correction of Affymetrix GeneChip data. *Stat Appl Genet Mol Biol*, **5**(1), Article 24.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, **3**(1), Article 3.