BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 20 2007, pages 2700–2707
doi:10.1093/bioinformatics/btm412

Gene expression

A comparison of background correction methods for two-colour microarrays

Matthew E. Ritchie[1], Jeremy Silver[2], Alicia Oshlack[2], Melissa Holmes[3], Dileepa Diyagama[4], Andrew Holloway[4] and Gordon K. Smyth[2,*]

[1]Department of Oncology, University of Cambridge, CRUK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK, [2]Bioinformatics Division, [3]Immunology Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3050 and [4]The Peter MacCallum Cancer Centre, St Andrews Place, East Melbourne, Victoria 3002, Australia

Received on April 16, 2007; revised on July 20, 2007; accepted on August 9, 2007

Advance Access publication August 25, 2007

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Microarray data must be background corrected to remove the effects of non-specific binding or spatial heterogeneity across the array, but this practice typically causes other problems such as negative corrected intensities and high variability of low intensity log-ratios. Different estimators of background, and various model-based processing methods, are compared in this study in search of the best option for differential expression analyses of small microarray experiments.

Results: Using data where some independent truth in gene expression is known, eight different background correction alternatives are compared, in terms of precision and bias of the resulting gene expression measures, and in terms of their ability to detect differentially expressed genes as judged by two popular algorithms, SAM and limma eBayes. A new background processing method (*normexp*) is introduced which is based on a convolution model. The model-based correction methods are shown to be markedly superior to the usual practice of subtracting local background estimates. Methods which stabilize the variances of the log-ratios along the intensity range perform the best. The *normexp+offset* method is found to give the lowest false discovery rate overall, followed by morph and vsn. Like vsn, normexp is applicable to most types of two-colour microarray data.

Availability: The background correction methods compared in this article are available in the R package *limma* (Smyth, 2005) from http://www.bioconductor.org.

Contact: smyth@wehi.edu.au

Supplementary information: Supplementary data are available from http://bioinf.wehi.edu.au/resources/webReferences.html.

1 INTRODUCTION

Two-colour microarray experiments quantify the relative gene expression between experimental samples for thousands of probes simultaneously. The pixel intensities from the Cy3 (green, *G*) and Cy5 (red, *R*) images are surrogate measures for the amount of hybridization between the RNA samples and the immobilized probe sequences.

Image analysis software returns foreground and background intensities for each spot. The foreground is an overall measure of the intensity of the spot while the background is a measure of the ambient signal. Background fluorescence can arise from many sources, such as non-specific binding of labelled sample to the array surface, processing effects such as deposits left after the wash stage or optical noise from the scanner. Removal of ambient, non-specific signal from the total intensity is known as 'background correction'.

Most image analysis programs return 'local' background intensities, obtained from the mean or median of the pixel intensity values surrounding each spot. Local background is arguably an unbiased estimate of the local non-specific signal, so subtracting it from the foreground intensity gives in principle an unbiased estimator of the true signal due to hybridization. Although well motivated, this traditional approach produces corrected intensities with undesirable statistical properties. It produces negative intensities whenever the background intensity is larger than the foreground intensity, leading to missing log-ratios, sometimes for a substantial proportion of probes on an array. Even when not missing, the log-ratios are highly variable for low intensity spots.

The problems caused by this 'fanning' phenomenon have been widely noted (Beiβbarth *et al.*, 2000; Bilban *et al.*, 2002; Finkelstein *et al.*, 2002; Kooperberg *et al.*, 2002). The most common reaction in the applied literature has been to filter out low intensity spots, even though this is another cause of missing values. Another response has been to develop methods which incorporate variance-intensity dependence into the differential expression analysis (Baggerly *et al.*, 2001; Newton *et al.*, 2001; Yang *et al.*, 2002a). Yet another has been to transform the corrected intensities to try to stabilize the variability over the intensity range (Cui *et al.*, 2003; Durbin and Rocke, 2004; Durbin *et al.*, 2002; Huber *et al.*, 2001, Kafadar and Phang, 2003; Rocke and Durbin, 2003).

The above strategies take background correction as a given. In this article, we take a step back and consider whether

*To whom correspondence should be addressed.

different background correction methods might avoid or mitigate the problems in the first place. Avoiding background correction altogether is recommended by Yang *et al*. (2001a) and Tran *et al*. (2002). Qin *et al*. (2004) examined the effect of background correction on bias and the ability to detect spike-in ratio controls. Background correcting the intensities did not improve the detection of DE genes, but the log-ratios from the spike-in genes were systematically underestimated when no background correction was performed.

Another possibility is to replace local background with a different image analysis measure. The *morph* background measure in the Spot software (CSIRO, North Ryde, Australia) and the TV+L[1] model background of Yin *et al*. (2005) are non-linear filters which give lower, less variable values than local background. Yang *et al*. (2002b) found the morph background estimate performed best in terms of bias-variance trade-off, compared with no background, constant background or local background, producing more extreme *t*-statistics for known differentially expressed (DE) genes.

The third possibility is to process the background estimate other than by subtraction. Edwards (2003) tapers the background to avoid negative corrected values. Kooperberg *et al*. (2002) gives an empirical Bayes model to estimate the true signal. The limma software (Smyth, 2005) provides a model-based adjustment (*normexp*) which requires less input information than that of Kooperberg *et al*. (2002). This method has proven successful in applied studies (Gilad *et al*., 2006; Peart *et al*., 2005) but has not yet been the subject of a comparative study. The variance stabilizing models of Huber *et al*. (2002) and Durbin and Rocke (2004) incorporate additive components which avoid negative intensities.

The aim of this article is to compare available background correction alternatives. We consider a specific context which is very common in experimental medicine, where the aim is to detect DE genes from a microarray experiment with a relatively small number of biological replicates. The popular algorithms SAM (Tusher *et al*., 2001) and limma eBayes (Smyth, 2004) are selected as representative of state-of-the-art statistical differential expression methods. Both of these algorithms increase statistical power in small microarray experiments by 'borrowing' information between probes. This article examines which background correction methods perform best in concert with these differential expression methods.

Section 2 outlines the eight background correction methods considered. Section 3 describes the data sets used to assess the methods. Section 4 discusses the major results in terms of bias, variance and differential expression and Section 5 presents our recommendations.

2 CORRECTION METHODS

The usual assumption of background correction for two-colour microarray data is that the background signals, R_b and G_b, are additive to the true signals, R and G on the raw intensity scale. Given the observed foreground intensities, R_f and G_f, this allows the true signal to be estimated by subtracting the foreground and background values, such that $R = R_f - R_b$ and $G = G_f - G_b$. The corrected intensities are then used to form the log-ratio, $M = \log_2(R/G)$, and average log-intensity, $A = \frac{1}{2}\log_2(RG)$, for each spot.

Table 1. Summary of the background correction methods considered

Method	Data extraction software	Bg estimate	Adjustment
Standard	GenePix Pro 3.0/4.0	Local median	Subtraction
Kooperberg	GenePix Pro 3.0/4.0	Local mean	Model
Edwards	GenePix Pro 3.0/4.0	Local median	Model
Normexp	GenePix Pro 3.0/4.0	Local median	Model
Normexp+offset	GenePix Pro 3.0/4.0	Local median	Model
Vsn	GenePix Pro 3.0/4.0	Local median	Model
Morph	Spot 2.0	Morph	Subtraction
No background	GenePix Pro 3.0/4.0	None	None

We compare eight background correction methods (Table 1) which use different estimates for R_b and G_b and different processing methods (variants on subtraction) for removing background signal. The methods are outlined below with details in Supplementary Material. All are implemented in the *backgroundCorrect* function of the *limma* software package. The standard method can produce negative corrected intensities while the other methods produce strictly positive corrected intensities.

Standard: The traditional correction method uses local background estimates for R_b and G_b and subtracts them from the foreground values. In this article, mean foreground and local median background estimates from GenePix Pro 3.0 (Axon Instruments, Union City, CA, USA) are used.

Kooperberg: Kooperberg *et al*. (2002) suggest an empirical Bayes model involving a convolution of normal distributions to background adjust the signal from each spot. Observed foreground and background mean intensities and their SDs, along with the number of foreground and background pixels for each spot in a given channel are used in this model. Numerical integration is applied to obtain the expected value of the true signal in each channel for each spot. We implemented this method by modifying Charles Kooperberg's S-Plus code (supplied in a personal communication). In practice, the method is restricted to GenePix data.

Edwards: A simpler strategy to avoid negative intensities is suggested by Edwards (2003), who adjusts the foreground intensities by subtracting the background when the difference between the foreground and background is larger than a small threshold value. When the difference is less than the threshold, subtraction is replaced by a smooth monotonic function. This method is applied with local median background estimates from GenePix.

Normexp: The *normexp* method is based on the same normal plus exponential convolution model which has previously been used to background correct Affymetrix data as part of the popular RMA algorithm (Irizarry *et al*., 2003; McGee and Chen, 2006). Two changes have been made to the method for use with two-colour arrays. First, the convolution model is fitted to the background subtracted signals for each channel separately. Second, the kernel density parameter estimation method used in RMA has been replaced by maximum-likelihood estimation, which is more sensitive to the true parameter values. In order to make maximum likelihood numerically feasible, a saddle-point approximation is used to simplify the mathematical form of the likelihood function. In this article, GenePix data (mean foreground, median background) was corrected using this model.

Normexp+offset: A slight variation on the normexp method is to add a small positive offset, k, to move the corrected intensities away from zero. This is a simple variance-stabilizing technique, analogous to the started-log approach described by Rocke and Durbin (2003). It should reduce the variation of the low intensity M-values, since $\log_2[(R+k)/(G+k)]$ will be close to 0 for R and G both small relative

to k. The use of an offset is effective here because normexp produces corrected intensities which are positive but may be close to zero. The value $k = 50$ was chosen on the basis of our previous experience with cDNA microarray experiments.

Vsn: The variance stabilization method of Huber *et al.* (2002) calibrates the data from each channel between arrays and uses a generalized arcsinh transformation of the data instead of the logarithm. The arcsinh function is defined for negative values, which ensures negative corrected signals can be handled. At high intensities, the arcsinh transform is equivalent to the regular log-ratio, whereas at low intensities it is close to the difference between the transformed intensities. 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, rather than the vsn default of natural logarithms, to allow comparability with the other methods. In contrast to the other seven methods in this study, vsn operates on all the arrays together rather than for each array separately, and background corrects and normalizes the intensity data in one operation. For all other methods, a separate normalization step is required.

Morph: The morph background, described in Yang *et al.* (2002b), gives lower, less variable estimates of the background than the local estimates. The morph background is obtained by performing a morphological opening, which involves applying an erosion operator (local minimum) followed by a dilation (local maximum) using a window of fixed size for each image. This background measure is available in the image analysis software Spot (CSIRO, North Ryde, Australia) and GenePix Pro 6.0 (Axon Instruments, Union City, CA, USA). In this article, the mean foreground and morph background obtained from the Spot software are used.

No background: equivalent to setting $R_b = G_b = 0$. In this article, GenePix mean foreground is used with this option.

3 TEST DATA

3.1 Spike experiment

This article uses three test data sets. The first uses Lucidea Universal ScoreCard (LUS) controls (Amersham Biosciences) to assess bias. Twelve copies of the LUS control probe set were printed on nine cDNA microarrays. The spots were printed in side-by-side duplicate pairs, to make 24 spots in total for each control probe. The arrays were also printed with a 13K clone library, but here we analyse the control probes only. Test and reference control RNA was spiked into the RNA samples prior to labelling to produce known fold changes (Table 2). The arrays were scanned on a GenePix 4000B scanner and image analysed using Spot and GenePix. All eight background correction methods were applied and, for methods other than vsn, the resulting log-ratios were normalized using global loess (Yang *et al.*, 2001b) with a span of 0.6. The larger than usual span is appropriate because of the smaller than usual number of spots. The duplicate spots for each probe were combined using the method of Smyth *et al.* (2005), as would be done for a differential expression analysis, to give an estimated $\log_2$-fold change, $\hat{\beta}$, for each copy of each probe.

3.2 Mixture experiment

The second data set is from Holloway *et al.* (2006). Six RNA mixtures consisting of mRNA from MCF7 and Jurkat cell-lines in known relative concentrations (100%:0%, 94%:6%, 88%:12%, 76%:24%, 50%:50% and 0%:100%) were compared to pure Jurkat reference mRNA on 12 cDNA microarrays printed with a Human 10.5K clone set. A dye-swap pair of arrays is available for each of the six mixtures, making 12 arrays in total. All eight background correction methods were applied and, except for vsn, the data was normalized using print-tip loess (Yang *et al.*, 2001b). The data was analysed by fitting probe-wise non-linear regression equations to the 12 arrays, to evaluate the precision and dynamic range of the microarrays, as described by Holloway *et al.* (2006). This analysis produces an estimate $\hat{T}_i$ of the MCF7 to Jurkat fold change for each probe i, a reliable estimate because it combines information from the entire mixture series. It also provides a SD $\hat{\sigma}_i$ which estimates the between-array measurement error for that probe.

Table 2. Summary of LUS controls in the Spike experiment

Control	True log fold change	A	DE?
U03Med	$\log^2(3)=1.58$	Medium	Yes
U10Med	$\log^2(10)=3.32$	Medium	Yes
D03Med	$-\log^2(3)=-1.58$	Medium	Yes
D10Med	$-\log^2(10)=-3.32$	Medium	Yes
U03High	$\log^2(3)=1.58$	High	Yes
U10High	$\log^2(10)=3.32$	High	Yes
D03High	$-\log^2(3)=-1.58$	High	Yes
D10High	$-\log^2(10)=-3.32$	High	Yes
Calibration	$\log^2(1/1)=0.00$	Low/medium/high	No

3.3 Quality control study

The final data set is a quality control study of 111 of the same 10.5K Human cDNA arrays used for the Mixture experiment. It is used as a source of independent truth for our false discovery rate comparison. All arrays were hybridized with MCF7 (Cy3) and Jurkat mRNA (Cy5). Spot image data was morph background corrected and print-tip loess normalized. A large proportion of the genes are expected to be DE between the two samples. We chose the top 30% of genes ranked by moderated t-statistic (Smyth, 2004) as DE, and the bottom 40% as non-DE. This gave 3098 DE and 4130 non-DE genes.

4 RESULTS

4.1 Heavy versus light background correction

The normalized M- and A-values for one array from the Mixture experiment are shown in Figure 1. This array has 100% Jurkat on both channels, so there is no true differential expression. The differences between the background correction methods for the same raw data are striking. Most obvious is that some background correction methods produce M-values which are much more variable than others, and this fanning is most apparent at low A-values. The differences would be visually even more striking if the vertical scale of the plot was not truncated for compactness. The M-values for this array are actually as large as 7.5 for standard

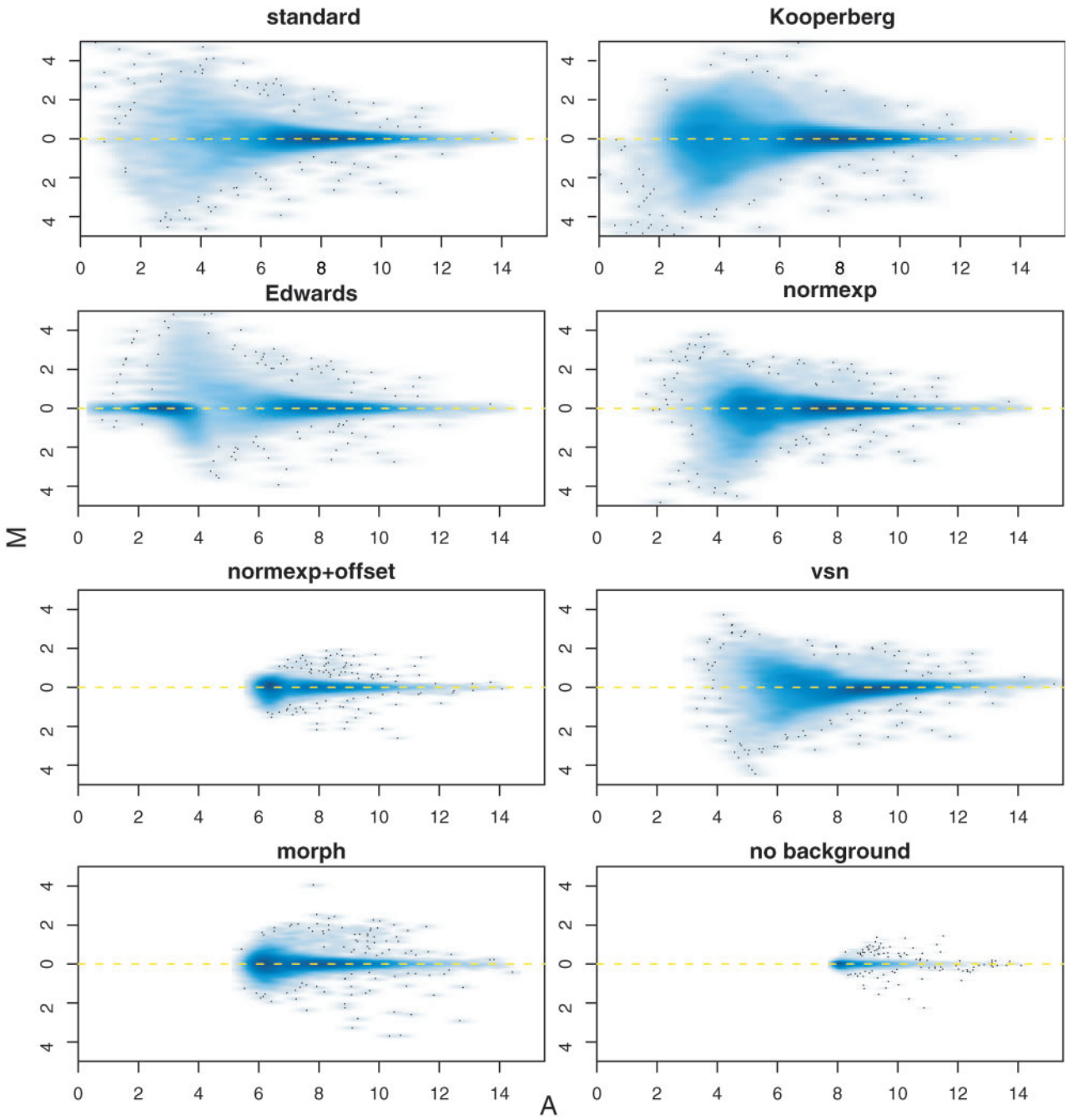

Fig. 1. MA-plots obtained using different background correction methods for a self-self hybridization of pure Jurkat from the Mixture experiment.

background, 9.6 for Kooperberg, 7.4 for Edwards, 5.5 for normexp and 4.6 for vsn. Only normexp+offset, morph and no background show all the data on the plots. MA plots for the other mixture comparisons are supplied as Supplementary Material.

The second striking feature of Figure 1 is that the background methods with less variable M-values also give compressed ranges of A-values. The most extreme is no background, for which the A-values start at nearly 8 rather than at 0.

The hidden cost of standard subtraction, which is not depicted in Figure 1, is missing values. Across all 12 arrays of the Mixture experiment, 16.8% of the M-values are missing for the standard method. The Kooperberg method also gave an occasional missing value, 14 or 0.01% in total, and the other methods gave none.

Taking these features into account, we can place the background correction methods broadly on a continuum for which standard background and no background form the extremes. At one end are methods which change the foreground

intensities relatively little giving intensities which are offset away from zero and low M-value variability. At the other end are methods which change the foreground intensities the most, giving a full range of intensities down to zero and highly variable M-values. The background methods can be ordered in this way from low to high offset as: standard, Kooperberg, Edwards, normexp, vsn, morph, normexp+offset and no background.

For the array in Figure 1, high offset and low M-value variability is desirable because the true M-values are zero. For other arrays with genuine differential expression, compression of the M-values may appear as bias. A major aim of this study is to determine where this trade-off should be drawn for confident assessment of differential expression.

4.2 Precision

The overall results from the Mixture experiment are now presented. The residual SD for each probe ($\hat{\sigma}_i$) is a measure of the precision with which the expression values returned by the microarrays follow the pattern of the mixing proportions. Figure 2 shows the trend in variability for each background method. For ease of comparability, the A-values have been standardized to be the same for each method. The vertical scale is $\log_2$-variance, so each unit on the vertical axis corresponds to a 2-fold change in variance or a halving of statistical information.

Most of the background methods show a trend to increasing precision at higher intensities. The trend is strongest for low offset methods and weakest for high offset methods, with no background and normexp+offset actually reversing the trend at the lowest intensities. It is clear that higher offsets give higher precision, although the different methods converge at higher intensities. The Kooperberg and Edwards methods give unexpectedly low precision at medium intensities and vsn gives unexpectedly low precision at high intensities. Interestingly, we found that by varying the offset k used for normexp, we could nearly match the precision curves obtained from the standard method, from morph and from the no background method (see Supplementary Material). We conclude that precision is largely a function of offset, but that the Kooperberg, Edwards and vsn methods are somewhat less precise than their effective offsets would predict.

4.3 Bias

It is to be expected that higher precision, purchased by compressing the intensity range, will result in attenuated signal as well. This is confirmed by examining the MCF7-Jurkat log-fold changes, ($\log_2 \hat{T}_i$), estimated from the Mixture experiment. Figure 3 shows boxplots of the log-fold changes arising from each method. The spread of fold changes is clearly less for the higher offset methods, although the largest fold changes are nearly as great in all cases.

To examine whether a smaller spread of fold changes can be interpreted as bias, we turn to the Spike experiment data. The predicted log-ratios for the LUS controls (see Table 2) were recovered more closely by some background correction methods than others. Figure 4 shows the M-values for the non-DE calibration controls and the DE D03Med ratio

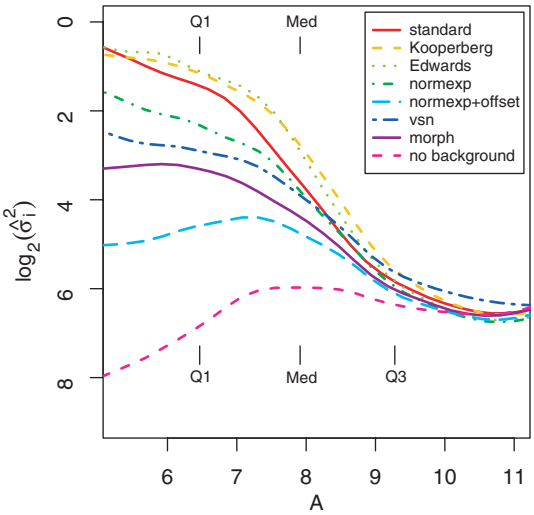

Fig. 2. Smoothed $\hat{\sigma}^2$ from the non-linear fits versus intensity for the Mixture experiment. The A-values have been standardized between methods, and plotted from the 5th to the 95th percentiles. The quantiles of the A-values are marked.

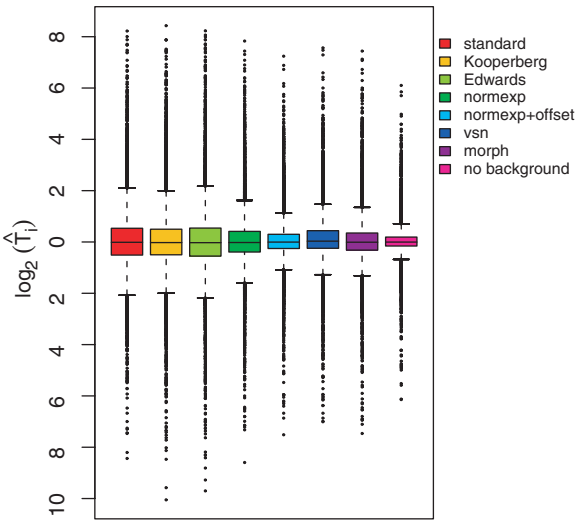

Fig. 3. Estimated $\log_2$-fold changes from the non-linear models fitted to the Mixture experiment data for each background processing method.

controls for a typical slide. The vsn method produced M-values which were systematically off target, being too low for both calibration and ratio controls. Using vsn to normalize the LUS controls together with the experimental probes did not alleviate the problem. The other methods produced M-values which were unbiased for the calibration controls and slightly attenuated towards zero for the ratio controls. The amount of attenuation increases steadily with the size of the effective offset, with the exception of vsn. The pattern was similar but less pronounced for the high intensity ratio controls (see Supplementary Material). Similar results were found for all the arrays in the Spike experiment (Supplementary Material).

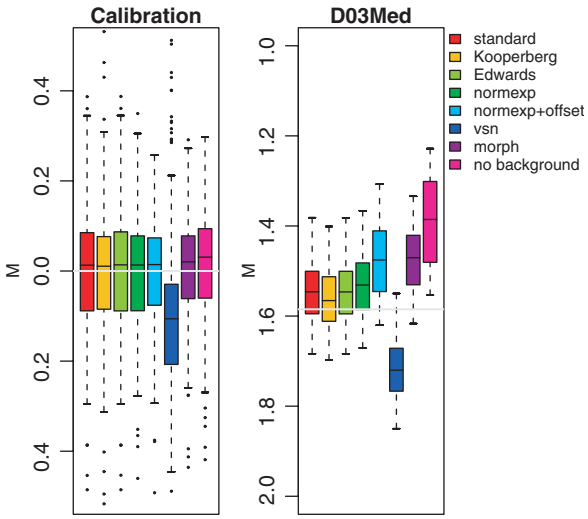

Fig. 4. *M*-values for the Calibration and D03Med controls from array 6 of the Spike experiment for different background correction methods. The solid gray lines show the theoretical spike-in log-ratios (0 and -1.58 respectively).

To summarize the bias for each background method, the mean absolute bias of the estimated $\log_2$-fold changes, $|\hat{\beta}_i - \mu|$, was computed over all the control probes, where μ is the true $\log_2$-fold change given in Table 2. In order of increasing bias, the methods were Kooperberg (0.213), standard (0.219), Edwards (0.219), normexp (0.238), vsn (0.263), morph (0.284), normexp+offset (0.305) and no background (0.342). Again, the ordering is from low to high offset methods. This shows that high offset for background correction translates to high bias as well as high precision.

4.4 Assessing DE

We now assess differential expression using arrays from the Mixture experiment. Ignoring the self-self hybridizations, the Mixture experiment consists of five dye-swap pairs of arrays. We assessed differential expression between MCF7 and Jurkat using each pair of arrays separately. The RNA mixtures vary from 100 to 50% MCF7, so the magnitude of the fold changes will vary from one pair of the arrays to another, but the set of DE genes should be the same in each case.

Using only two arrays to find DE genes presents a challenging problem, because there is only one degree of freedom available to estimate genewise SDs. The level of difficulty further increases with the concentration of Jurkat in the MCF7:Jurkat RNA mixture. The use of ordinary *t*-tests or other traditional univariate statistics to assess differential expression would be disastrous (Smyth, 2004). Instead we use two of the most popular algorithms for microarray differential expression which have the characteristic of 'borrowing' information between genes. These algorithms have the ability to make statistical inferences with some confidence even for small numbers of replicate arrays. Genes were ranked in terms of evidence for differential expression using SAM regularised *t*-statistics (Tusher *et al.*, 2001 and using empirical Bayes moderated *t*-statistics (Smyth, 2004). The statistics were calculated using the *samr* (http://www-stat.stanford.edu/~tibs/SAM) and *limma* software packages, respectively.

To assess the success of the differential expression analyses, an independent determination of which genes are truly DE is required. We selected the top 30% of genes from the quality control study as unambiguously DE and the bottom 40% as unambiguously non-DE. The remaining 30% of genes were treated as indeterminate and are not used in the analysis.

Figure 5 shows the number of false discoveries for each method versus the number of genes selected by ranking the genes using $|t|$-statistics, from largest to smallest for limma eBayes (a) and SAM (b). The curves have been averaged across the five dye-swap pairs. The curves show that the high offset background correction methods generally do better than low offset methods, with the best performance achieved by normexp + offset, then morph, then vsn. The standard method of background subtraction is by far the worst method.

Comparing the SAM and limma results, SAM gives somewhat more false discoveries but this effect is more marked for some background methods than others. SAM does nearly as well as limma for the best three background correction methods, but its performance trails off more rapidly than that of limma when presented with less than optimal background values. In particular, SAM is much worse than limma for the no background choice. With limma, the no background option does nearly as well as the best methods, whereas it is the second worst method with SAM. SAM is also noticeably poorer for the low offset methods Kooperberg and Edwards.

We investigated why SAM should perform so poorly with the no background option. The issue appears to be the amount of moderation which is done of the genewise sample variances or standard errors. Limma smoothes the genewise variances towards a common value, controlling the degree of smoothing by the 'prior degrees of freedom' (pdf) (Smyth, 2004). With the no background option, limma estimates the pdf to be 5.4 on average, indicating that a large amount of smoothing is occurring. By comparison, limma uses pdf of only 1.7 on average for the standard background method, indicating much less smoothing of the variances. The general trend for the limma analyses is that more smoothing is done for the high offset background methods. In SAM, the amount of smoothing is controlled by the exchangeability factor s_0 added to the SD in the denominator of the *t*-statistic. The exchangeability factor is a percentile of the SD values, and the percentile indicates the amount of smoothing. For the no background method, s_0 was a low percentile (35th, 20th, 65th, 35th and 65th for each mixture) which indicates only modest smoothing is done. By comparison, SAM uses the 100th percentile, the maximum allowed value, in every case with normexp. It appears that SAM is less able than limma to adapt the amount of smoothing to all situations.

5 DISCUSSION

Our study has shown that the different background correction options differ markedly in terms of bias and precision, and that bias and precision need to be traded off as they are competing

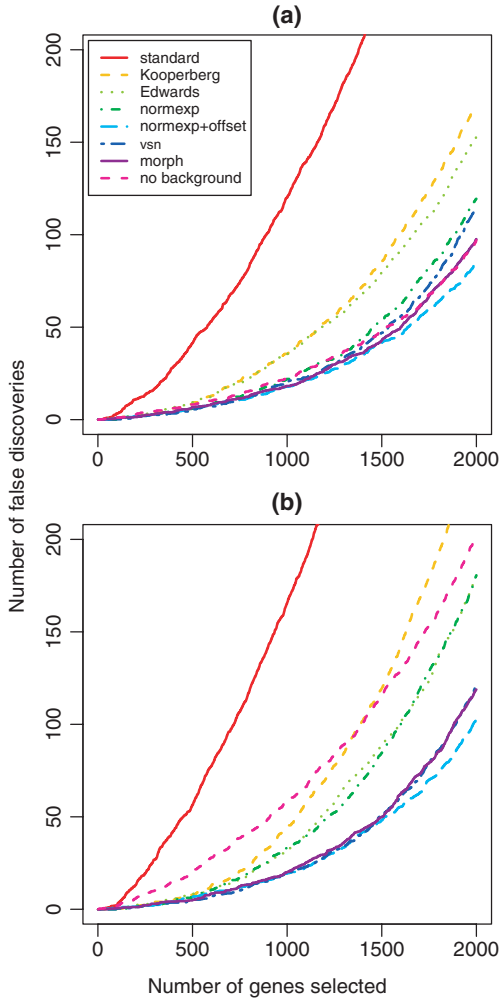

Fig. 5. Number of false discoveries from the Mixture data set using moderated *t*-statistics from (a) limma eBayes and (b) SAM. Each curve is an average over the 5 mixtures.

requirements. We assessed this trade-off in terms of false discovery rates in the context of a small-scale microarray experiment.

Our first result is that the standard and most common method of background correcting microarray data is far worse than other methods, resulting in a sizeable number of false discoveries. We hope that the traditional practice of subtracting local background will gradually disappear given that better options are readily available. Abandoning the standard background correction method would also avoid the chronic problem of missing values, which greatly complicate downstream data analysis. All the other methods considered return strictly positive intensities and avoid missing values. The performance of the other extreme, no background correction, was better but still poor when used with the SAM algorithm.

The best performing background methods are those which best stabilize the variance as a function of intensity, namely normexp+offset, morph and vsn, in that order. Despite the fact that these methods are more biased than the standard method, tending to attenuate the signals somewhat, the improvement in precision has been shown to outweigh the increase in bias for the purpose of detecting differential expression. It is interesting to note that normexp+offset and morph appear to give the best stabilization of the variance as a function of intensity in Figure 2, even beating vsn, which is explicitly designed to stabilize the variance.

The morphological opening background estimator, morph, is to date available only in Spot software or as a user-determined option in GenePix Version 6.0. The model-based method normexp+offset appears to give the same benefits as morph but can be used with any image analysis software.

For this study, we chose the offset k for normexp+offset based on our previous experience with cDNA microarray experiments. In practice, the value of k is chosen by examining MA-plots of the microarray data, and choosing the offset so as to visually stabilize the variability of the M-values as a function of intensity. A more numeric algorithm could be devised, for example k could be chosen to maximize the prior degrees of freedom estimated by limma in a linear model analysis. However, we have not found results to be sensitive to the specific value of k used, i.e. good results are typically obtained for a range of sensible values (Supplementary Material).

Durbin and Rocke (2004) have previously shown how an additive offset can behave as a simple but effective variance stabilizing transformation. Our offset k is similar to the offset c in their started-log transformation and to the free parameter in their generalized-log transformation. This relates directly to the continuum that we observe in Section 4 from light to heavy background correction methods. Our normexp+offset has many properties similar to the started-log transformation. A key advantage of normexp however is that positivity of the corrected intensities is enforced before the offset is applied. This ensures that the offset is comparable for all spots and arrays regardless of the background level, and allows the offset to be chosen purely to stabilize the variance rather than having to achieve positivity at the same time.

The two differential expression methods, SAM and limma, had comparable performance with the best background correction methods, but limma was better able to adapt to background options with different characteristics.

Our study did not focus on normalization methods, but the unusual biases noted on some arrays after vsn normalization were less evident in the loess normalized data. This suggests that the removal of intensity-dependent trends in the data can improve the accuracy of the expression measures. Removal of such trends are not possible using the linear normalization procedure used in vsn.

ACKNOWLEDGEMENTS

Thanks to Terry Speed for valuable discussions, James Wettenhall for image analysis of the Spike experiment and Rachel Uren for proofreading this article.

Conflict of Interest: none declared.

REFERENCES

Baggerly,K.A. *et al*. (2001) Identifying differentially expressed genes in cDNA microarray experiments. *J. Comput. Biol.*, **8**, 639–659.

Beißbarth,T. *et al*. (2000) Processing and quality control of DNA array hybridization data. *Bioinformatics*, **16**, 1014–1022.

Bilban,M. *et al*. (2002) Defining signal thresholds in DNA microarrays: exemplary application for invasive cancer. *BMC Genomics*, **3**, 19.

Cui,X. *et al*. (2003) Transformations for cDNA microarray data. *Stat. Appl. Genet. Mol. Biol.*, **2**, Article 4.

Durbin,B.P. and Rocke,D.M. (2004) Variance-stabilizing transformations for twocolor microarrays. *Bioinformatics*, **20**, 660–667.

Durbin,B.P. *et al*. (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18** (Suppl. 1), S105–S110.

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

Finkelstein,D. *et al*. (2002) Microarray data quality analysis: lessons from the AFGC project. *Plant. Mol. Biol.*, **48**, 119–131.

Gilad,Y. *et al*. (2006) Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature*, **440**, 242–245.

Holloway,A.J. *et al*. (2006) Statistical analysis of an RNA titration series evaluates microarray precision and sensitivity on a whole-array basis. *BMC Bioinformatics*, **7**, Article 511.

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**, 249–264.

Kafadar,K. and Phang,T. (2003) Transformations, background estimation, and process effects in the statistical analysis of microarrays. *Comput. Stat. Data. Anal.*, **44**, 313–338.

Kooperberg,C. *et al*. (2002) Improved background correction for spotted DNA microarrays. *J. Comput. Biol.*, **9**, 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**, Article 24.

Newton,M.A. *et al*. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from micro-array data. *J. Comput. Biol.*, **8**, 37–52.

Peart,M.J. *et al*. (2005) Identification and functional significance of genes regulated by structurally diverse histone deacetylase inhibitors. *Proc. Natl Acad. Sci. USA*, **102**, 3697–3702.

Qin,L. *et al*. (2004) Empirical evaluation of data transformations and ranking statistics for microarray analysis. *Nucleic Acids Res.*, **32**, 5471–5479.

Rocke,D.M. and Durbin,B. (2003) Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics*, **19**, 966–972.

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

Smyth,G.K. (2005) Limma: linear models for microarray data. In Gentleman,R. *et al*. (eds.), *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York, 397–420.

Smyth,G.K. *et al*. (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, **21**, 2067–2075.

Tran,P.H. *et al*. (2002) Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic Acids Res.*, **30**, e54.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Yang,I.V. *et al*. (2002a) Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol.*, **3**, R62.

Yang,Y.H. *et al*. (2001a) Analysis of cDNA microarray images. *Brief. Bioinform.*, **2**, 341–349.

Yang,Y.H. *et al*. (2001b) Normalization for cDNA microarray data. In Bittner,M.L. *et al*. (eds.) *Microarrays: Optical Technologies and Informatics,* Vol 4266 of *Proceedings of SPIE*.

Yang,Y.H. *et al*. (2002b) Comparison of methods for image analysis on cDNA microarray data. *J. Comput. Graph. Stat.*, **11**, 108–136.

Yin,W. *et al*. (2005) Background correction for cDNA microarray images using the TV+L1 model. *Bioinformatics*, **21**, 2410–2416.