Multiplicative background correction for spotted microarrays to improve reproducibility

DABAO ZHANG[1]*, MIN ZHANG[1] AND MARTIN T. WELLS[2,3]

[1] *Department of Statistics, Purdue University, West Lafayette, IN 47907, USA*
[2] *Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA*
[3] *Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA*

(*Received 16 December 2005 and in revised form 15 March 2006*)

Summary

We propose a simple approach, the multiplicative background correction, to solve a perplexing problem in spotted microarray data analysis: correcting the foreground intensities for the background noise, especially for spots with genes that are weakly expressed or not at all. The conventional approach, the additive background correction, directly subtracts the background intensities from foreground intensities. When the foreground intensities marginally dominate the background intensities, the additive background correction provides unreliable estimates of the differential gene expression levels and usually presents *M–A* plots with 'fishtails' or fans. Unreliable additive background correction makes it preferable to ignore the background noise, which may increase the number of false positives. Based on the more realistic multiplicative assumption instead of the conventional additive assumption, we propose to logarithmically transform the intensity readings before the background correction, with the logarithmic transformation symmetrizing the skewed intensity readings. This approach not only precludes the 'fishtails' and fans in the *M–A* plots, but provides highly reproducible background-corrected intensities for both strongly and weakly expressed genes. The superiority of the multiplicative background correction to the additive one as well as the no background correction is justified by publicly available self-hybridization datasets.

1. Introduction

With a two-colour competitive hybridization process, spotted microarrays provide a genome-wide measure of differential gene expression levels in two samples as well as controlling for undesirable effects (Schena *et al.*, 1995; Brown & Botstein, 1999). Analysis of spotted microarrays starts with quantifying each cDNA array into image files, and segmenting each pixel of the images into either the spotted or unspotted regions. As shown in Fig. 1*a*, pixel intensities of the spotted (or unspotted) region within each spot are summarized into the median foreground (or background) intensities for the two channels (traditionally red and green), say R_f (or R_b) and G_f (or G_b). While the foreground intensities R_f and G_f measure the fluorescence intensities caused by specific hybridization of the mRNA samples to the spotted cDNA, the background intensities R_b and G_b measure the fluorescence intensities of the background noise. The goal of background correction is to correct the foreground intensities for the background noise within the spotted region. With R_b and G_b as estimates of the background noise within R_f and G_f, it is appropriate to correct R_f and G_f with R_b and G_b, respectively.

Assuming the foreground intensities are affected additively by the background noise, the conventional background correction, the additive background correction (ABC), proceeds by a direct subtraction of R_b and G_b from R_f and G_f, respectively. Then the *log-ratio M* (the logarithm of the ratio between the background-corrected spot intensities) and *log-intensity A* (the average of the logarithmic background-corrected spot intensities) are calculated. The procedure is displayed in the left-hand panel

[*] Corresponding author. Tel: +1 (765) 4946046. Fax: +1 (765) 4940558. e-mail: zhangdb@stat.purdue.edu

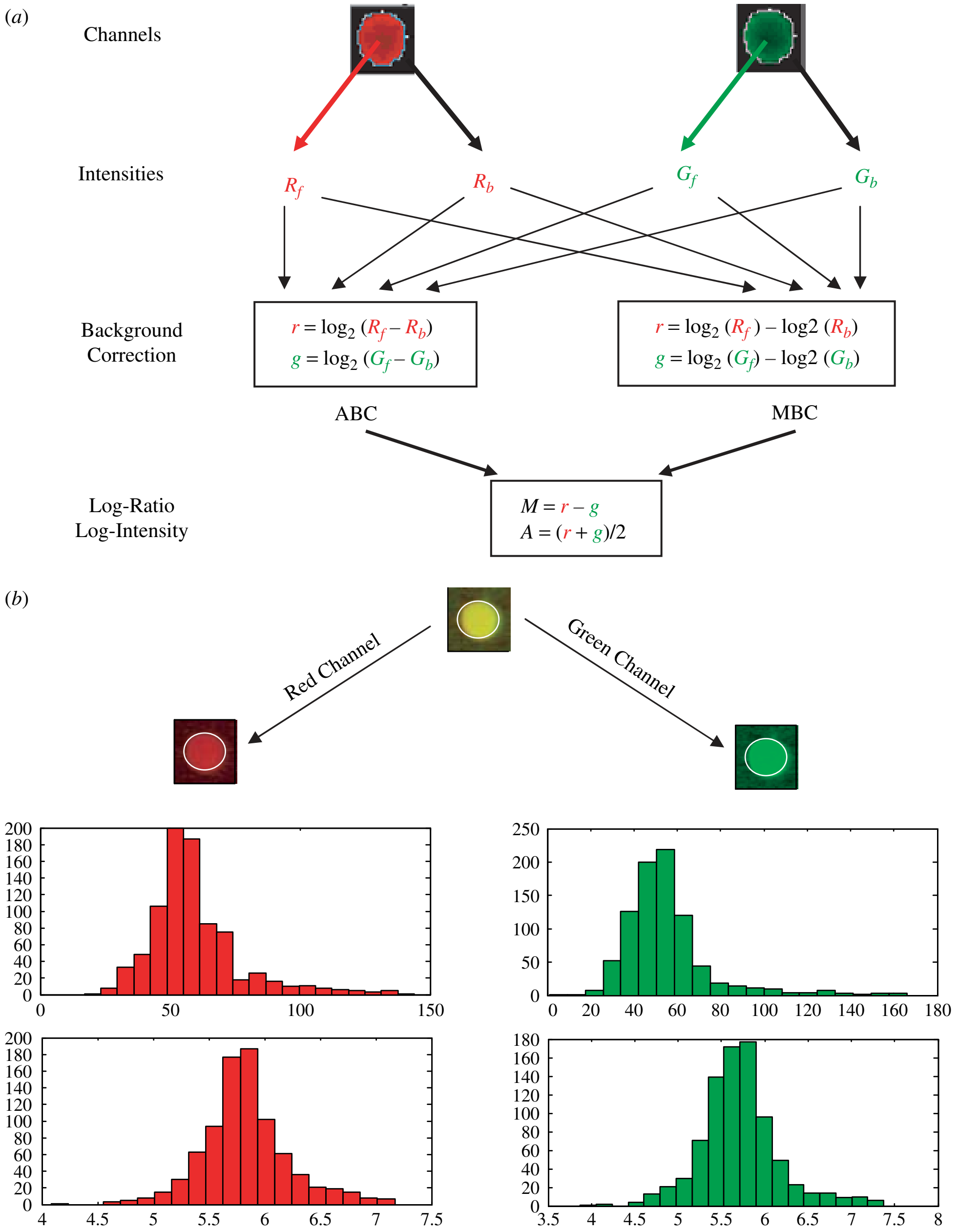

Fig. 1. Illustration of the additive background correction (ABC) and the multiplicative background correction (MBC). (*a*) ABC directly subtracts the background intensities from the foreground intensities. MBC logarithmically transforms both background intensities and foreground intensities before the background correction, with the logarithmic transformation symmetrizing the skewed statistics. (*b*) The top two histograms are for the original background intensities, i.e. the intensities of the pixels outside the ellipsoids; the bottom two are for the logarithmic background intensities.

of Fig. 1*a*. However, this strategy excludes the non-positive background-corrected intensities and gives rise to other problems in the downstream analyses. First, for genes expressed very weakly (or not at all) in either channel, the foreground intensities of some spots are overwhelmed by the background intensities. Missing log-ratios and missing log-intensities will be reported for such spots in all downstream analyses. The differential expression information of these genes is essentially lost by discarding non-positive background-corrected spot intensities, even though these genes may be highly differentially expressed or their expression patterns may change significantly across time. Secondly, spots with small positive background-corrected intensities are usually unreliable estimates of *M* and *A*, which manifest as 'fishtails' or 'fans' in *M*–*A* plots. These spots have wildly varying log-ratios that challenge the downstream analyses, and may therefore lose differential expression information of these genes.

Background noise, as the major component of microarray noise, is not due to specific sticking

https://doi.org/10.1017/S0016672306008196 Published online by Cambridge University Press

of target molecules to the array surface. Principles of fluorescence spectroscopy indicate that the feature readings are proportional to the density of the fluorescent molecules (Schena, 2003). However, the multiplicative coefficient may vary locally, affected by spot-specific factors such as the assay surface and fluorescence emission from the surrounding spots. Assuming that the background noise affects the feature readings multiplicatively, we propose the multiplicative background correction (MBC), shown in the right-hand panel for Fig. 1*a*, to obtain more precise background-corrected intensities for both strongly and weakly expressed genes. This approach avoids losing differential expression information for the genes weakly expressed in either channel. Skewness of the background intensity and symmetry of the logarithmic background intensity (see Fig. 1*b*) further confirm that the multiplicative error structure is a more realistic modelling assumption. Application to publicly available self-hybridization datasets shows the excellent performance of MBC.

2. Methods

While empirical observations reveal a non-addictive relation between the foreground and background intensities (Brown *et al.*, 2001), the principles of fluorescence spectroscopy indicate that the background noise affects the foreground intensities multiplicatively. Therefore, instead of subtracting the background intensities directly from the foreground intensities as ABC, we propose MBC to subtract the estimates of the logarithmic background intensities from the logarithmic foreground intensities (see Fig. 1). The immediate advantage of MBC is that the background-corrected log-ratio and log-intensity for each spot are well defined. Second, the logarithmic transformation roughly symmetrizes the background intensities, since background intensities are skewed to the right (Kim *et al.*, 2002) and are roughly distributed lognormally (see Fig. 1*b*). Obviously, for genes expressed weakly or not at all, we have background-corrected logarithmic spot intensities fluctuating around zero.

Because the median is invariant to any monotonic transformation, we suggest summarizing a spot's foreground (or background) intensities with the median of the corresponding spotted (or unspotted) region for each channel, i.e. R_{f} (or R_{b}) for the traditional red channel and G_{f} (or G_{b}) for the traditional green channel. Define $r_{\mathrm{f}}=\log_2(R_{\mathrm{f}})$, $r_{\mathrm{b}}=\log_2(R_{\mathrm{b}})$, $g_{\mathrm{f}}=\log_2(G_{\mathrm{f}})$ and $g_{\mathrm{b}}=\log_2(G_{\mathrm{b}})$. Assume that the logarithmic intensity medians r_{f} and g_{f} have additive noise effects $\tilde{r}_{\mathrm{b}}=\log_2(\tilde{R}_{\mathrm{b}})$ and $\tilde{g}_{\mathrm{b}}=\log_2(\tilde{G}_{\mathrm{b}})$, respectively, where $\tilde{R}_{\mathrm{b}}$ and $\tilde{G}_{\mathrm{b}}$ are the unobservable median background intensities of the spotted region for the two distinct channels. It is reasonable to estimate $\tilde{r}_{\mathrm{b}}$ by r_{b} and $\tilde{g}_{\mathrm{b}}$ by g_{b}. The background-corrected logarithmic spot intensities are calculated as $r=r_{\mathrm{f}}-r_{\mathrm{b}}$ and $g=g_{\mathrm{f}}-g_{\mathrm{b}}$; the background-corrected spot intensities are then calculated as $R=2^{r_{\mathrm{f}}-r_{\mathrm{b}}}$ and $G=2^{g_{\mathrm{f}}-g_{\mathrm{b}}}$. Because fold changes of gene expressions between two mRNA samples are of interest to researchers, we can focus on the well-defined r and g to extract information from the microarray experiment.

With the background-corrected spot intensities r and g, the log-ratio and log-intensity can be calculated following the conventional definition, i.e. $M=r-g$ as the log-ratio to measure the differential gene expression and $A=(r+g)/2$ as the log-intensity to measure the overall gene expression. Because, for empty spots, the paired r_{f} and r_{b}, g_{f} and g_{b} are assumed to have identical mean values, we expect both M and A will stay close to zero even before any normalization of the data. Negative A implies that the corresponding gene is expressed weakly or not at all in both mRNA samples.

3. Results

(i) *The CAGE self-hybridization data*

We use two replicate cDNA microarray datasets, all hybridized using one sample of species *Arabidopsis thaliana*, from the Compendium of *Arabidopsis* Gene Expression project (CAGE; http://www.ebi.ac.uk/microarray/Projects/cage/). These self-hybridization data are publicly available at ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) with experiment accession number E-CAGE-2. Each array has 19 992 spots in total, including 243 empty spots. With an ideal background correction, log-intensities of the empty spots are expected to be identically zero, and log-ratios of all spots are expected to be zero because both arrays are from a self-hybridization experiment; the between-array *M–M* plot is also expected to have all points tightly clustering around zero.

When using ABC, about one-quarter of the spots (i.e. one array with 4960 spots and the other with 4807 spots) need to be discarded from each of the two CAGE self-hybridization arrays owing to the dominant background noise; more than 60 % of the empty spots are discarded. Furthermore, the *M–A* plots (see the top plots in Fig. 2) have the notorious 'fishtail' patterns with the remaining empty spots scattering within the fishtails. In contrast, MBC provides *M–A* plots with tight bands and all empty spots clustering around the origin (see the bottom plots in Fig. 2). When employing no background correction (NBC), the shapes of the corresponding *M–A* plots are similar to those with MBC (see the central plots in Fig. 2). The empty spots from NBC

https://doi.org/10.1017/S0016672306008196 Published online by Cambridge University Press

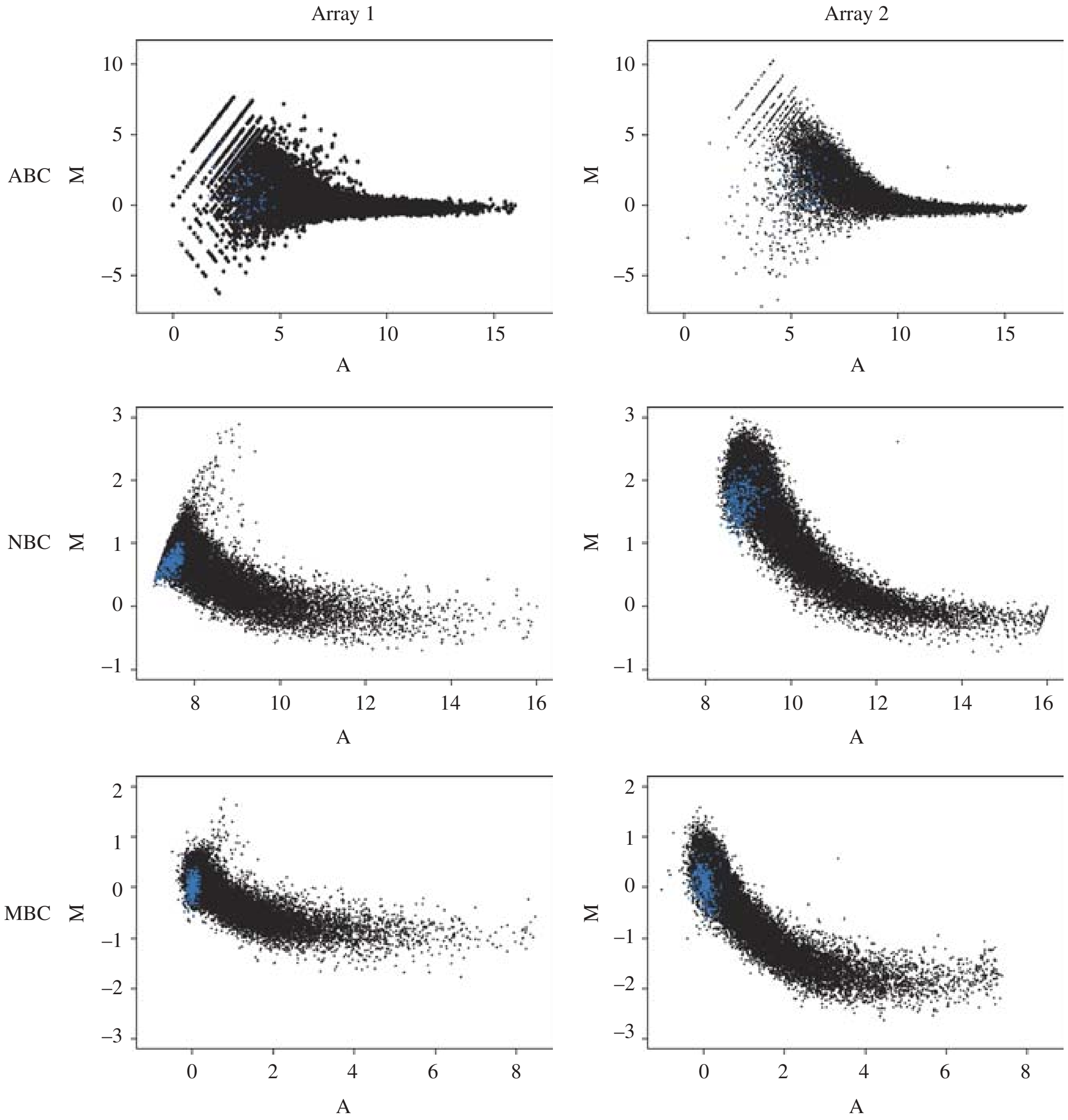

Fig. 2. The *M–A* plots for the CAGE self-hybridization data. The two plots are from ABC, the two middle plots are from NBC (i.e. no background correction) and the two bottom plots are from MBC. The empty spots are in red.

are away from the origin and have slightly greater variability. As shown in a later example, the background correction is necessary to correct the foreground intensities from the background noise, especially when the spots are contaminated heavily by their highly expressed neighbouring spots.

The reproducibilities of different background correction methods are shown in Fig. 3 by plotting the estimated differential gene expression levels (i.e. the normalized *M* after the background correction) across replicated arrays. Either before or after an intensity-dependent normalization (Yang *et al.*, 2001), the *M–M* plots of the replicated arrays have points scattered wildly when employing ABC (see the top two plots in Fig. 3). In particular, the estimated differential expression levels for empty spots are in no way reproducible as they vary largely from array to array (see the red points of the top two plots in Fig. 3). In contrast, when using MBC, the *M–M* plot of the normalized *M* has a very tight cluster around the origin (see the bottom right plot in Fig. 3), showing the estimated differential expression levels are highly reproducible. A linear pattern is shown in the *M–M* plot of the non-normalized *M* (see the bottom left plot in Fig. 3), which implies that the same systematic errors lie across different arrays. The disappearance of the linear pattern from the *M–M* plot of the normalized *M* validates the intensity-dependent normalization in removing these systematic errors. More extreme spots and the irregular shape of the *M–M* plot for NBC suggest that MBC should be favoured among the three approaches.

https://doi.org/10.1017/S0016672306008196 Published online by Cambridge University Press

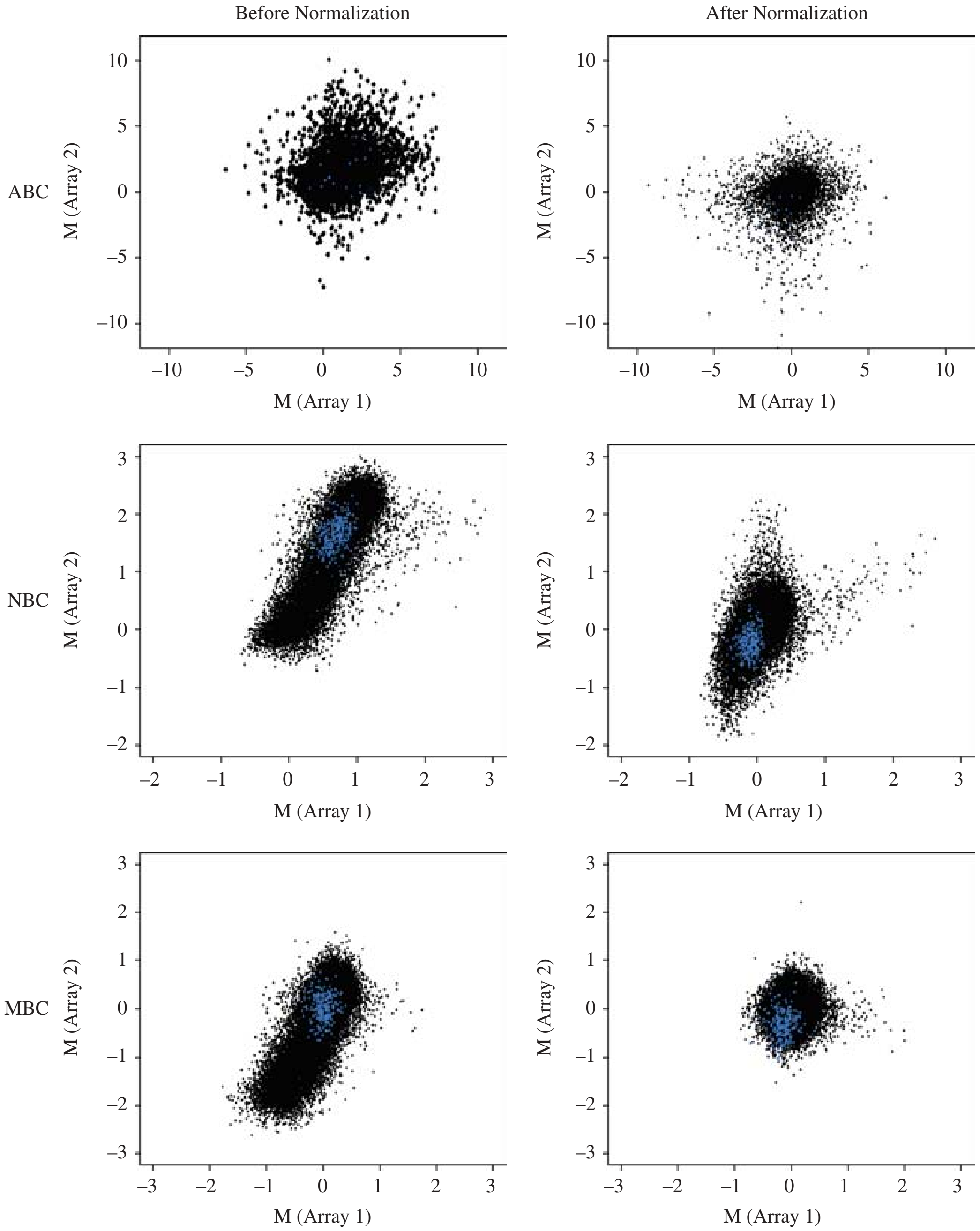

Fig. 3. The *M–M* plots for the CAGE self-hybridization data. Shown are the plots of *M* across the replicated CAGE self-hybridization arrays. The left-hand plots are before normalizing the data; the right-hand plots are after normalizing the data. The empty spots are in red.

(ii) *The yeast data with spiked controls*

The yeast data with spiked controls were collected to evaluate a design of microarray experiment (see van de Peppel *et al*., 2003). This dataset is available at ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) with experiment accession number E-UMCU-1. It is a self-hybridization experiment for *Saccharomyces cerevisiae* with controls spiked at three different ratios 1:1, 1:2, and 1:10. For each ratio, a pair of dye-swap microarrays was hybridized. Each of the six microarrays incorporated 6371 gene probes for *Saccharomyces cerevisiae* with each gene spotted twice. There were also 864 spots left empty in each microarray. Each of the 14 controls (nine normalization controls and five ratio controls) was spotted at least twice onto each subgrid of the microarrays to generate sufficient data points. The normalization control RNAs were spiked in the total RNA at concentrations varying over three orders of magnitude to cover a range of mRNA expression levels. The ratio control RNAs were spiked into paired mixes

https://doi.org/10.1017/S0016672306008196 Published online by Cambridge University Press

of all other external control RNAs at ratios 1:1, 1:3 and 1:7 before the mixes of the external contol RNAs were added at ratios of 1:1, 1:2 and 1:10 to paired aliquots of a single yeast total RNA preparation. Therefore, there are three different experiments, i.e a 1:1 experiment with ratio controls mixed at ratio 1:1, 1:3, 1:7, 3:1, 7:1, but other controls (including normalization controls) mixed at ratio 1:1; a 1:2 experiment with ratio controls mixed at ratio 1:2, 1:6, 1:14, 3:2, 7:2, but other controls mixed at ratio 1:2; and a 1:10 experiment with ratio controls mixed at ratio 1:10, 1:30, 1:70, 3:10, 7:10, but other controls mixed at ratio 1:10. We apply different background correction approaches to all arrays. After the background correction, each array is normalized locally with an intensity-dependent normalization procedure by using only the yeast spots to estimate the non-linear trends of M versus A, since none of the yeast spots is differentially expressed and they have a wide range of concentrations.

The superiority of MBC to ABC is clearly demonstrated in Fig. 4, where the estimated log-ratio M from MBC is plotted against that from ABC for each array (the within-array M–M plots). An obvious linear relationship exists between the log-ratios M from two different approaches for the ratio controls, confirming that, for genes with high log-intensities, the estimated log-ratio M is comparable between these two different approaches. For the empty spots, however, the estimated M stays around zero when using MBC but varies widely when using ABC, which is revealed as horizontal lines in the within-array M–M plots. Therefore, MBC provides reliable estimated log-ratios M for weakly expressed genes, i.e. genes with low log-intensities, whereas ABC cannot.

The estimated log-ratio M based on MBC is also more reproducible than that based on ABC, especially for genes expressed weakly or not at all. As shown in the M–M plot for each pair of dye-swap arrays (see Fig. 5), the ratio controls line up away from the origin when using either approach, indicating that both approaches have good reproducibility for genes with high levels of log-intensities. For those genes with low levels of log-intensities, however, MBC has superior reproducibility, because the empty spots and yeast spots cluster tightly around the origin when using MBC but spread out when using ABC.

Because of the overwhelming background noise, M and A for most empty spots are not defined when using ABC. Furthermore, the remaining empty spots with well-defined background-corrected M and A scatter widely to the left of the M–A plots after normalization (see Fig. 6). The spots with low log-intensities have inflated variations, which make downstream analyses rather difficult if not impossible. In contrast, MBC provides, M–A plots with all the empty spots closely clustering around the origin. The normalized log-ratio M, when using MBC, has rather stable variation across different levels of the long-intensity A, which makes downstream data analyses more transparent.

Ideally, the log-ratios of all the normalization spots should be at the same level for each array because the normalization controls were spiked at a specific ratio into the yeast RNAs. As displayed in Fig. 6, the estimated M of the normalization controls are not constant within each array after either ABC or MBC. When using ABC, the estimated M of the normalization controls merge into other estimated M, which scatter widely to the left of the M–A plot and lack reproducibility (shown in Fig. 5). When using MBC, however, the estimated M of the normalization controls converges to tight clusters around the origin formed by the estimated M of empty spots, which is highly reproducible across arrays (see Fig. 5).

As shown by the plots in Figs 5 and 6, MBC provides a consistent systematic connection between the true and our estimated differential expression levels. Clearly, this systematic connection can be described by a non-linear function, which also depends on the total gene expression levels. Let M_{theo} and A_{theo} be the theoretical log-ratio and log-intensity, respectively, for a gene with estimated log-ratio M and its mean μ_{M}. Mathematically, the systematic connection may be expressed as $\mu_{\text{M}} = g(M_{\text{theo}}, A_{\text{theo}})$, where $g(.,.)$ is a smoothing function with $g(M_{\text{theo}}, 0) = 0$ and $g(M_{\text{theo}}, \infty) = M_{\text{theo}}$. Consistency of this systematic connection is supported by Fig. 5, where the normalization controls, after using MBC, lie either in a tight cluster (with controls spiked 1:1) or on a straight line (with controls spiked 1:2 or 1:10). In contrast, ABC cannot establish such a consistent systematic connection.

4. Discussion

(i) *MBC versus NBC*

With the empirical observation that ABC is inferior to NBC in some microarray experiments, many researchers have proposed discarding the information in the background intensities and ignoring background correction issues (Qin & Kerr, 2004; Parmigiani *et al*., 2003). However, as shown by the M–A plots of yeast spots only (Fig. 7), some yeast spots have noisy background due to the neighbouring ratio controls and therefore their estimated M from NBC are far from zero, which results in false positives. In contrast, the estimated M from MBC have been corrected for the effects of neighbouring spots, whereas ABC provides a plot with an unstable estimate of M for weakly expressed genes. This example shows that the background correction is

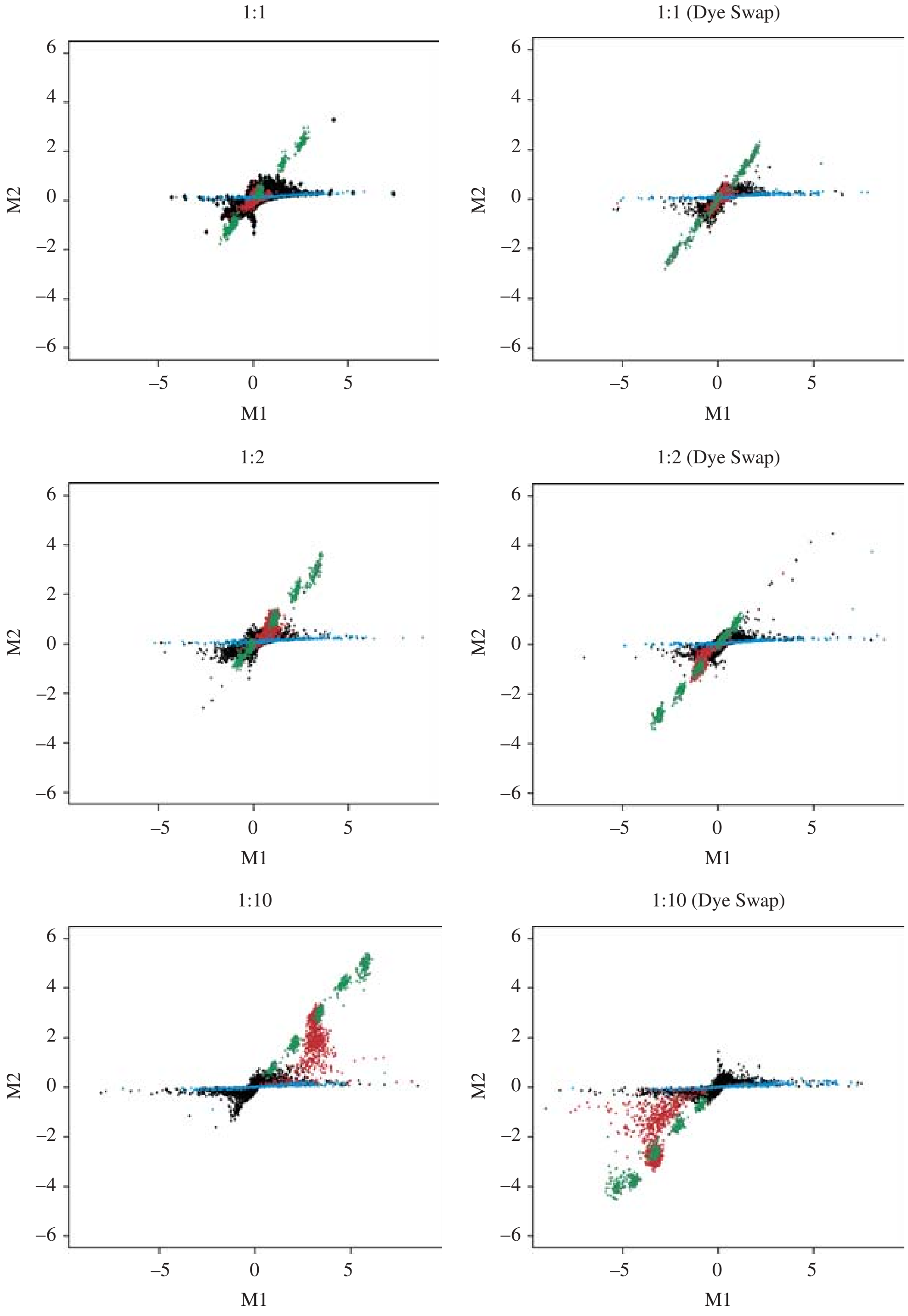

Fig. 4. Within-array *M–M* plots for yeast data with spiked controls. Shown in each plot is the estimated *M* based on the MBC (along *y*-axis) against the estimated *M* based on ABC (along *x*-axis) for each array. The yeast RNA spots, empty spots, ratio controls and normalization spots are plotted in black, blue, green and red, respectively.

necessary to reduce the number of false positives in identifying differentially expressed genes.

To further address the necessity of background correction for some microarray data, we applied both NBC and MBC to the carp microarray data of Gracey *et al.* (2004), which are available at ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) with experiment accession number E-MAXD-1. Fig. 8 includes the plot of MBC-based *M* versus NBC-based *M* for the carp microarray B30D4B-Forward data, where the spots are clustered along two lines, i.e. the line $y=x$ and the line $y=x-1$. These two clusters imply two different backgrounds, one with similar background noises from the two channels and the

https://doi.org/10.1017/S0016672306008196 Published online by Cambridge University Press

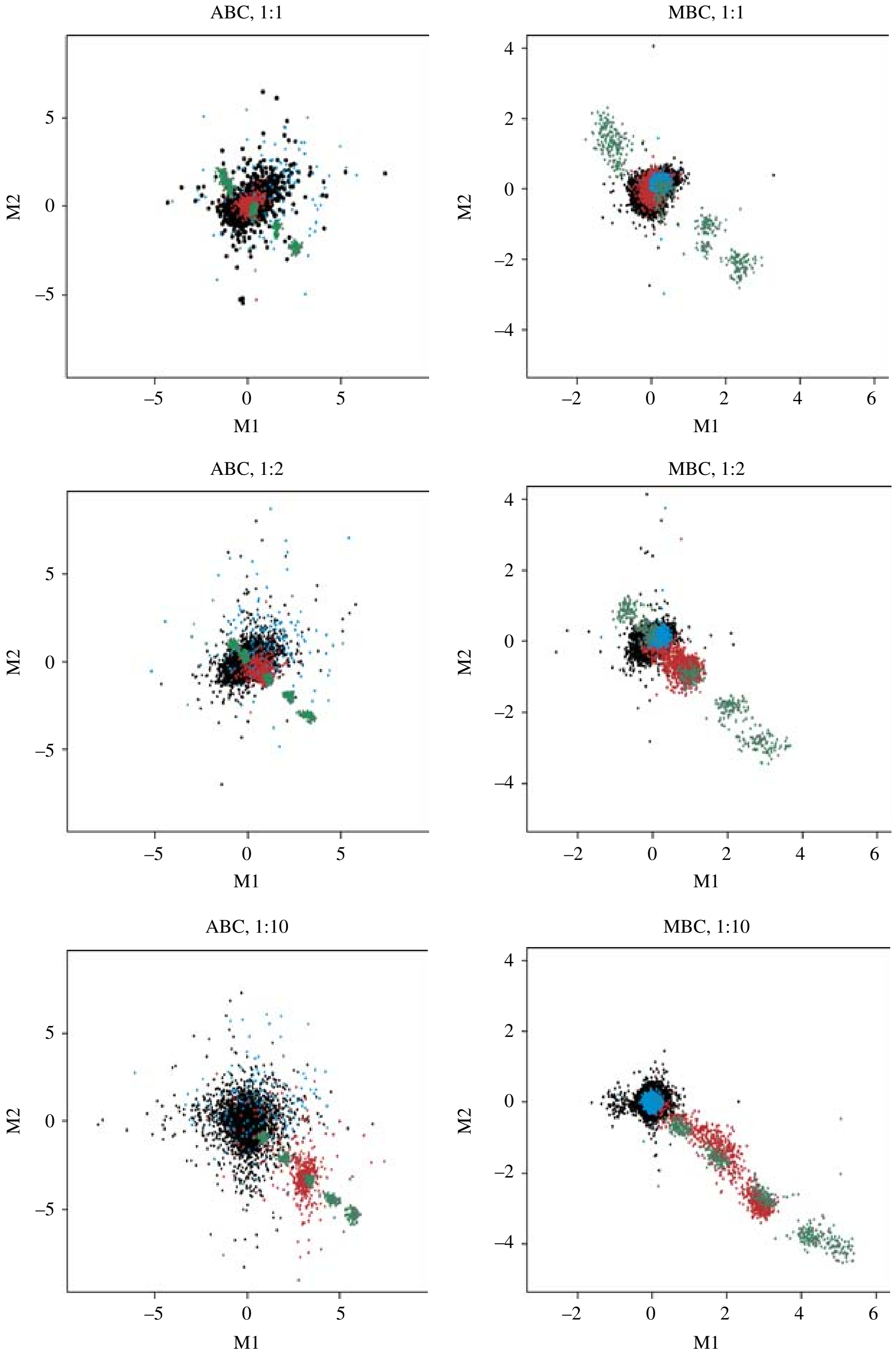

Fig. 5. The *M–M* plots for yeast data with spiked controls. Each *M–M* plot is for a pair of the dye-swap arrays with controls spiked at the same ratio. The left-hand plots are based on ABC, and the right-hand ones are based on MBC. The yeast RNA spots, empty spots, ratio controls and normalization spots are plotted in black, blue, green and red, respectively.

other with different background noises from the two channels. Therefore, background correction is necessary to complement the normalization because of the heterogeneous background noises. For the carp microarray B30D4B-Reverse data, the background noises are rather homogeneous and therefore MBC-based *M* agrees with NBC-based *M* except at several spots (see Fig. 8). A recent simulation study identified

https://doi.org/10.1017/S0016672306008196 Published online by Cambridge University Press

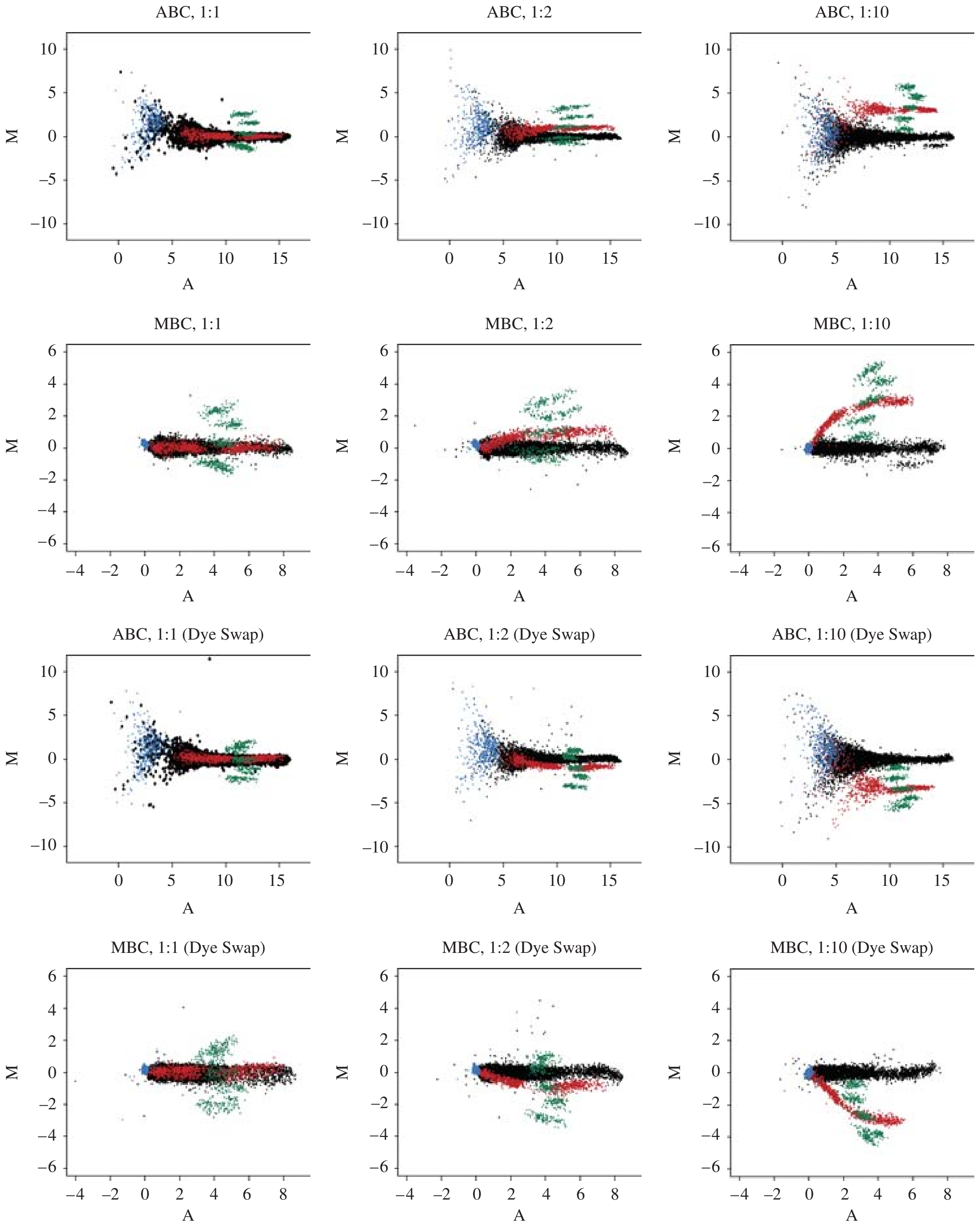

Fig. 6. The *M*–*A* plots for yeast data with spiked controls after normalization. For each array, the upper plot is based on ABC and the lower one is based on MBC. The yeast RNA spots, empty spots, ratio controls and normalization spots are plotted in black, blue, green and red, respectively.

factors that are important for determining whether to dispense with ABC (Scharpf *et al.*, 2004) or not. The agreement between NBC-based *M* and MBC-based *M* in the case of homogeneous backgrounds implies that one can improve microarray data analysis with MBC and reduce false positives caused by heterogeneous backgrounds.

MBC's outperformance of other background-correction approaches does not rely on the combination of scanner and feature-extraction software,

https://doi.org/10.1017/S0016672306008196 Published online by Cambridge University Press

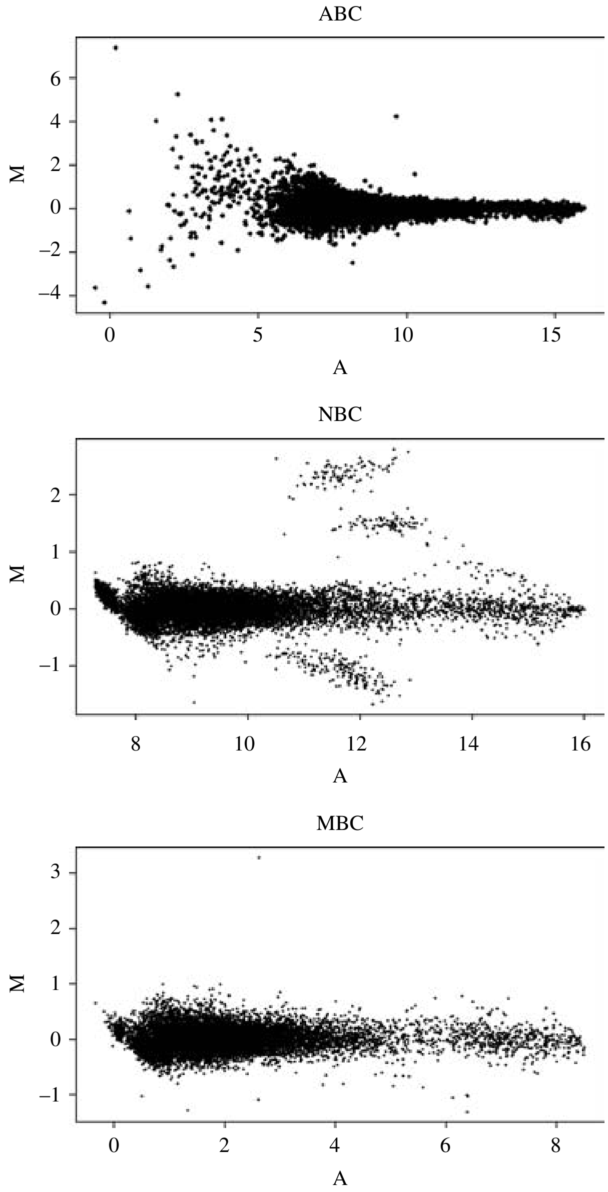

Fig. 7. The *M–A* plots for yeast spots only after normalization. Shown in the plots is the first array of the yeast data with controls spiked at 1:1. The top plot is based on ABC, the middle one is based on NBC and the bottom one is based on MBC.

although some scanners and softwares may have certain advantages over others. Both experiments E-CAGE-2 and E-UMCU-1 used the scanner ScanArray (PerkinElmer) and the software ImaGene (Biodiscovery), but the experiment E-MAXD-1 used a GenePix 4000A Scanner and GenePix 3.0 software (Axon Instruments). It may be of further interest to compare the performance of different scanners and feature-extraction softwares when using MBC.

(ii) *A statistical perspective of MBC*

Decomposing $r_{\mathrm{f}}=\mu_{\mathrm{r}}+\tilde{r}_{\mathrm{b}}$ and $g_{\mathrm{f}}=\mu_{\mathrm{g}}+\tilde{g}_{\mathrm{b}}$, where μ_{r} and μ_{g} are logarithmic spot intensities corresponding to specific hybridizations. Here $\tilde{r}_{\mathrm{b}}$ and $\tilde{g}_{\mathrm{b}}$ are logarithmic spot intensities of the fluorescences not due to specific hybridizations and may have spot-specific non-zero means which are the primary interest in the background correction. Usually it is assumed that r_{b} and g_{b} have identical non-zero means to $\tilde{r}_{\mathrm{b}}$ and $\tilde{g}_{\mathrm{b}}$, respectively. Replacing $\tilde{r}_{\mathrm{b}}$ with r_{b} and $\tilde{g}_{\mathrm{b}}$ with g_{b} we expect to remove spot-specific undesirable mean shifts from r_{f} and g_{f}, that is,

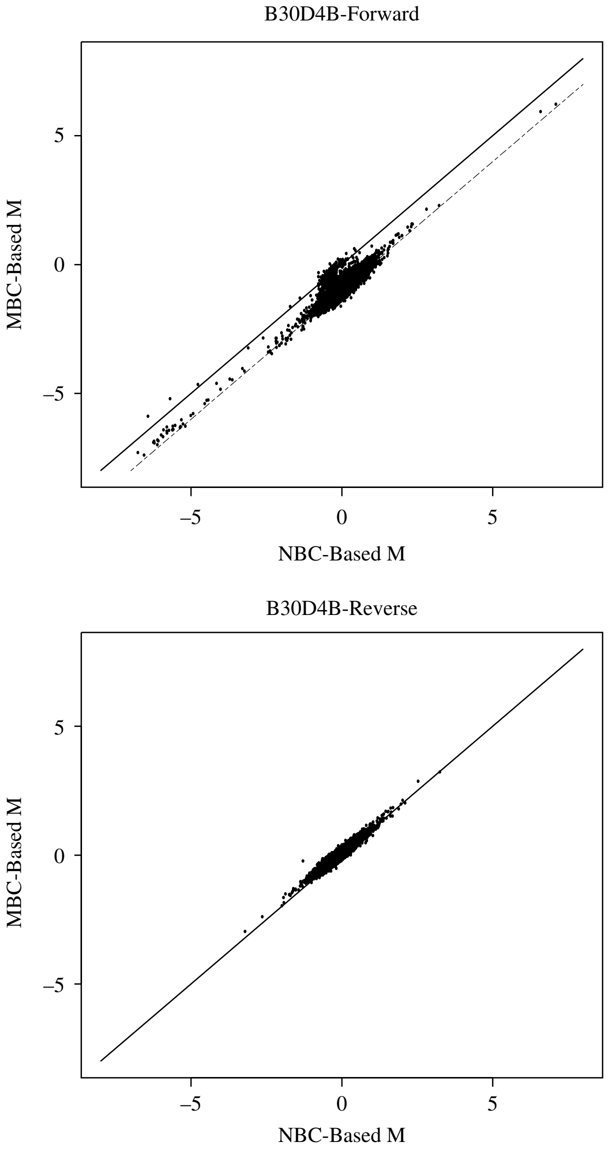

Fig. 8. The *M–M* plots for the carp microarray B30D4B-Forward (top) and B30D4B-Reverse (bottom). Shown in the plots are the estimated *M* based on the MBC (along *y*-axis) against the estimated *M* based NBC (along *x*-axis) with lines $y=x$ (unbroken) and $y=x-1$ (dashed).

$$r=r_{\mathrm{f}}-r_{\mathrm{b}}=\mu_{\mathrm{r}}+\varepsilon_r,$$
$$g=g_{\mathrm{f}}-g_{\mathrm{b}}=\mu_{\mathrm{g}}+\varepsilon_g,$$

where $\varepsilon_r=\tilde{r}_{\mathrm{b}}-r_{\mathrm{b}}$ and $\varepsilon_g=\tilde{g}_{\mathrm{b}}-g_{\mathrm{b}}$ distribute symmetrically with zero means. Then, *r* and *g* are used to

https://doi.org/10.1017/S0016672306008196 Published online by Cambridge University Press

estimate μ_{r} and μ_{g} but with symmetric mean-zero errors ε_r and ε_g, respectively. The log-ratio $\mu_{\mathrm{M}} = \mu_r - \mu_g$ and log-intensity $\mu_{\mathrm{A}} = \frac{1}{2}(\mu_r + \mu_g)$ are estimated by the background-corrected log-ratio M and the background-corrected log-intensity A, that is,

$$M = r - g = \mu_M + (\varepsilon_r - \varepsilon_g),$$
$$A = (r+g)/2 = \mu_A + (\varepsilon_r + \varepsilon_g)/2.$$

Ideally μ_M indicates the differential expression of one gene under two distinct conditions and should not depend on μ_A. However, systematic errors usually result in a non-linear relation between observed M and A, which has to be removed from M before using it to estimate the differential gene expression level. As shown by Zhang *et al.* (2005), a rigorous statistical normalization of M has to consider measurement errors in both M and A, although the measurement error in A is usually smaller than that in M. The measurement error in A, however, is usually ignored in conventional data normalization, which may inflate the number of false positives.

Because the variance of ε_{r} equals the sum of the variance $\tilde{r}_{\mathrm{b}}$ and r_{b} minus twice the covariance between $\tilde{r}_{\mathrm{b}}$ and r_{b}, it follows that as the correlation of $\tilde{r}_{\mathrm{b}}$ and r_{b} increases, the variance of r decreases (similarly for g). When r_{b} (or g_{b}) is so closely related to $\tilde{r}_{\mathrm{b}}$ (or $\tilde{g}_{\mathrm{b}}$) that the correlation coefficient is over 0·5, the variance of ε_r (or ε_g) is then smaller than that of r_{b} (or g_{b}). In this case, using r_{b} (or g_{b}) to estimate $\tilde{r}_{\mathrm{b}}$ or ($\tilde{g}_{\mathrm{b}}$) not only removes spot-specific undesirable drifts from μ_r (or μ_g) but reduces the variation of the measurement error. Therefore, the measurement-error issue should be considered when developing the image analysis approach. Furthermore, MBC provides another advantage, even if the logarithmic background intensities r_{b}, $\tilde{r}_{\mathrm{b}}$, g_{b} and $\tilde{g}_{\mathrm{b}}$ are non-symmetric, the disturbance errors in r and g, i.e. ε_r and ε_g, are still symmetric; hence we have symmetric measurement errors in the background-corrected log-ratio M and log-intensity A.

(iii) *Fishtails in using ABC*

In the case that there are many genes expressed either weakly or not at all, 'fishtail' patterns will inevitably appear in the M–A plot when using ABC. Let $\hat{R}$ and $\hat{G}$ be the spot intensities from ABC. For all spots with $\hat{R}$ at a specific level, say $\hat{R} = 2^{\mathrm{c}}$, we have

$$\frac{\hat{R}}{\hat{G}} = 2^{2c}(\hat{R}\hat{G})^{-1}, \text{ or equivalently,}$$

$$\log_2 \frac{\hat{R}}{\hat{G}} = 2c - 2 \times \frac{1}{2} \log_2 (\hat{R}\hat{G}).$$

This implies that the spots with $\hat{R}$ at a specific level line up in the M–A plot under the x-axis with the slope equal to -2. Similarly, the spots with $\hat{G}$ at a specific level line up in the M–A plot above the x-axis with the slope equal to $+2$. When ABC generates many spots with either $\hat{R}$ or $\hat{G}$ at each low level, the corresponding M–A plot is shown with many lines on its left and therefore has a 'fishtail' pattern. In contrast, MBC provides a M–A plot with spots clustered around the origin on its left by avoiding the unstable calculation of $\hat{R}$ and $\hat{G}$.

(iv) *Negative background-corrected intensities*

To avoid the negative background-corrected intensities for the spots with dominant background noise in either channel, different approaches have been proposed to modify ABC. For example, Smyth (2004) suggested replacing with small positive values, and Edwards (2003) proposed using log-linear interpolation. Based on the conventional assumption that background noise affects the foreground intensities in either channel additively, a Bayesian approach was proposed to avoid negative intensities (Kooperberg *et al.*, 2002). This Bayesian approach is computation-intensive, however, and still needs to exclude some spots with dominant background noise.

In practice, unbiased estimates are always pursued and used to identify differentially expressed genes. Therefore, the negative background-corrected logarithmic intensities form MBC are reasonable and expected for some of the genes expressed either weakly or not at all. Accordingly, the corresponding log-ratios and log-intensities from MBC may also be negative. With the necessity of background correction and the excellent performance of this simple strategy, it is unnecessary to pursue sophisticated strategies to overcome the problem. Instead, a strict statistical normalization of the data is necessary to consider measurement errors in both M and A since, as shown in a recent study (Zhang *et al.*, 2005), ignoring the measurement errors in A may increase the number of false positives.

We thank Frank C. Holstege and Jeroen van de Peppel for providing detailed description of the microarray experiments for the yeast data with spiked controls. We also thank Andrew G. Clark for suggestions that improved this manuscript.

References

Brown, P. O. & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics* **21**, 33–37.

Brown, C. S., Goodwin, P. C. & Sorger, P. K. (2001). Image metrics in the statistical analysis of DNA microarray data. *Proceedings of the National Academy of Sciences of the USA* **98**, 8944–8949.

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

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D. & Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell* **11**, 4241–4257.

Gracey, A. Y., Fraser, E. J., Li, W., Fang, Y., Taylor, R. R., Rogers, J., Brass, A. & Cossins, A. R. (2004). Coping with cold: an integrative, multitissue analysis of the transcriptome of a poikilothermic vertebrate. *Proceedings of the National Academy of Sciences of the USA* **101**, 16970–16975.

Kim, J. H., Shin, D. M. & Lee, Y. S. (2002). Effect of local background intensities in the normalization of cDNA microarray data with a skewed expression profiles. *Experimental and Molecular Medicine* **34**, 224–232.

Kooperberg, C., Fazzio, T. G., Delrow, J. J. & Tsukiyama, T. (2002). Improved background correction for spotted DNA microarray. *Journal of Computational Biology* **9**, 55–66.

Parmigiani, G., Garrett, E. S., Irizarry, R. A. & Zeger, S. L. (2003). The analysis of gene expression data: an overview of methods and software. In *The Analysis of Gene Expression Data: Methods and Software* (ed. G. Parmigiani, E. S. Garrett, R. A. Irizarry & S. L. Zeger, Springer, New York), pp. 1–45.

Qin, L.-X. & Kerr, K. F. (2004). Empirical evaluation of data transformations and ranking statistics for microarray analysis. *Nucleic Acids Research* **32**, 5471–5479.

Scharpf, R. B., Iacobuzio-Donahue, C. A. & Parmigiani, G. (2004). When should one subtract background fluorescence in cDNA microarrays? (www.bepress.com/jhubiostat/paper50).

Schena, M. (2003). *Microarray Analysis*. Hoboken, NJ: Wiley.

Schena, M., Shalon, D., Davies, R. W. & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with complementary DNA microarray. *Science* **270**, 467–470.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, No. 1, Article 3.

van de Peppel, J., Kemmeren, P., van Bakel, H., Radonjic, M., van Leenen, D. & Holstege, F. C. (2003). Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Reports* **4**, 387–393.

Yang, Y. H., Dudoit, S., Luu, P. & Speed, T. P. (2001). Normalization for cDNA microarray data. *SPIE BiOS 2001*, San Jose, California.

Zhang, D., Wells, M. T., Smart, C. D. & Fry, W. E. (2005). Bayesian normalization and inference for differential gene expression data. *Journal of Computational Biology* **12**, 391–406.

https://doi.org/10.1017/S0016672306008196 Published online by Cambridge University Press