Bernoulli **10**(6), 2004, 943–949

Intensity-dependent normalization in microarray analysis: a note of concern

MEI-LING TING LEE[1] and GEORGE A. WHITMORE[2]

[1]*Channing Laboratory, HMS/BHW, Harvard Medical School, 181 Longwood Avenue, Boston MA 02115, USA. E-mail: meiling@channing.harvard.edu*
[2]*Faculty of Management, McGill University, 1001 Sherbrooke Street West, Montreal, Canada, H3A 1G5. E-mail: george.whitemore@mcgill.ca*

We discuss the concern that intensity-dependent normalization can give biased estimates of differential expression and, therefore, can misclassify some moderately important genes as unexpressed.

Keywords: gene expression; intensity dependence; *MA* plot; microarray; normalization

1. Introduction

Normalization of gene expression intensity data in microarray studies aims to remove the influence of extraneous factors in order to provide precise and unbiased estimates of differential expression for genes across experimental samples. Normalization adjusts for sources of variability such as dye colour, pin tip effects and spatial anomalies on slides. A variety of methodologies have been proposed for this purpose. This note looks at one such methodology, intensity-dependent normalization described in Yang *et al*. (2002). We raise the concern that intensity-dependent normalization can give biased estimates of differential expression and, therefore, can misclassify some moderately important genes as unexpressed.

2. *MA* plots and intensity-dependent normalization

Dudoit *et al*. (2002), among others, have noted a dependence of differential expression on average intensity for the red and green intensities of a spot in cDNA arrays. Denoting the red and green log-intensity readings for any spot by $y^{(\mathrm{R})}$ and $y^{(\mathrm{G})}$, Dudoit *et al*. demonstrate this dependence in a plot of the log-intensity difference M against the mean log-intensity A, where

$$M = y^{(\mathrm{R})} - y^{(\mathrm{G})}, \qquad A = \tfrac{1}{2}(y^{(\mathrm{R})} + y^{(\mathrm{G})}). \tag{1}$$

This plot, referred to as an *MA* plot, provides a clear picture of the relationship.

Yang *et al*. (2002) suggest a normalization for gene expression data that uses lowess smoothing of the *MA* plot. This approach is referred to as *intensity-dependent normalization*. In it, a lowess-fitted function $\ell(A)$ of the average intensity A is used to normalize the differential expression M by computing the difference $M^{(\mathrm{d})} = M - \ell(A)$,

where the superscript (d) denotes an intensity-dependent adjustment. The authors show that this plot may vary by the pin tip of the arrayer and, therefore, suggest that the intensity-dependent normalization be carried out separately for each pin tip, in essence, giving normalized log-intensity differences of the form $M_i^{(d)} = M_i - \ell_i(A)$, where i is an index for the pin tip and $\ell_i(A)$ is the lowess function fitted to the data from the ith pin tip.

3. Re-examining the intensity-dependent normalization method in a case study

We now introduce a case study that we use to illustrate an *MA* plot and the nature of intensity-dependent normalization. This microarray data set was first presented in Lee *et al*. (2002) and was collected to investigate differential gene expression in kidney tissue from mutant (type 1) and wild-type (type 2) mice with juvenile cystic kidneys. The experimental design involves eight readings for each gene in four microarray pairs, according to the pattern set out in Table 1. 'Array' in Table 1 refers to the four microarray pairs (arrays a_1 to a_4). 'Channel' refers to whether the expression reading comes from the Cy3 green fluorescent channel (channel 1) or the Cy5 red fluorescent channel (channel 2). The data set in this experiment contains ScanAlyze cDNA gene expression data for 1728 genes (Eisen and Brown, 1999). Additional details of the experiment and data may be found in Lee *et al*. (2002).

We wish to consider the inherent logic of intensity-dependent normalization. We begin by looking at separate *MA* plots for arrays a_1 and a_2 of the case study. We note that these two arrays involve a comparison of expression intensity for genes from mutant and wild-type tissue, with the colour assignment reversed for array a_2 in comparison to array a_1. The raw data correspond to the ScanAlyze variables CH1I and CH2I and, hence, are not background-corrected. We have done a preliminary normalization for array, colour and tissue type that centres the log-readings for all genes at each level of these factors.

Notice that, although the M and A in (1) are defined in terms of the red and green intensities, the colours actually stand in for two experimental samples that are labelled by the two dyes. Hence, in this paper, we discontinue the use of notation $y^{(R)}$ and $y^{(G)}$ in (1). Instead, we use the notation y_{adt} to clearly identify the design parameters for the intensity

Table 1. Experimental design for the case study

	Channel 1 (Green)	Channel 2 (Red)
Array a_1	mutant	wild type
Array a_2	wild type	mutant
Array a_3	mutant	mutant
Array a_4	wild type	wild type

reading. Specifically, the first index a denotes the array (for arrays a_1 and a_2), the second index d denotes the dye ($1 =$ green and $2 =$ red) and the third index t denotes the sample tissue ($1 =$ mutant and $2 =$ wild type). Table 2 shows the *reversed-colour design*.

Individual *MA* plots for the two arrays now involve M and A values defined as follows.

$$M_1 = y_{122} - y_{111}, \quad A_1 = \tfrac{1}{2}(y_{122} + y_{111}) \qquad \text{for array } a_1, \tag{2}$$

$$M_2 = y_{212} - y_{221}, \quad A_2 = \tfrac{1}{2}(y_{212} + y_{221}) \qquad \text{for array } a_2. \tag{3}$$

Figure 1(a) shows the difference M_1 in normalized log-intensity for wild-type and mutant tissue plotted against the average normalized log-intensity A_1 for all genes for array a_1. The graph shows the zero lines for M_1 and A_1 (the log-averages). The graph also shows the smoothed values of M_1 as a lowess-fitted function of A_1. The fitted function is curved and shows a strong dependence of M_1 on A_1. Figure 1(b) shows the corresponding plot of M_2 against A_2 for array a_2. Observe that the differences M_1 and M_2 are both defined as differences between wild-type and mutant tissue.

Also note that in each array the tissue type is *confounded* with the dye colour. The strong curvilinear dependence of M on A is clear in each of the *MA* plots, with the curvatures reversed (convex and concave, respectively) because of the colour reversal. As described previously, intensity-dependent normalization would be applied to these two arrays by calculating corrected differential expressions of the form $M_1^{(\mathrm{d})} = M_1 - \ell(A_1)$ and $M_2^{(\mathrm{d})} = M_2 - \ell(A_2)$ for each gene. Judgements about differentially expressed genes would then be based on the $M_1^{(\mathrm{d})}$ and $M_2^{(\mathrm{d})}$ statistics.

Arrays a_1 and a_2 in Table 2 have a reversed-colour design. The colour reversal compensates for potential interaction of dye and gene expression. Based on equations (2) and (3), a single *MA* plot for this design can easily be constructed using the following definitions of *MM* and *AA*:

$$MM = \bar{y}_2 - \bar{y}_1, \qquad AA = \tfrac{1}{2}(\bar{y}_2 + \bar{y}_1), \tag{4}$$

where, taking the colour (dye) reversal into account, the average normalized log-intensities for experimnetal samples t_1 and t_2 are denoted by

$$\bar{y}_2 = \tfrac{1}{2}(y_{122} + y_{212}), \qquad \bar{y}_1 = \tfrac{1}{2}(y_{111} + y_{221}),$$

respectively. Observe that *MM* in (4) captures the differential expression for the two

Table 2. Reversed-colour design embedded in the case-study design

	Green dye d_1	Red dye d_2
Array a_1	sample t_1	sample t_2
	intensity y_{111}	intensity y_{122}
Array a_2	sample t_2	sample t_1
	intensity y_{212}	intensity y_{221}

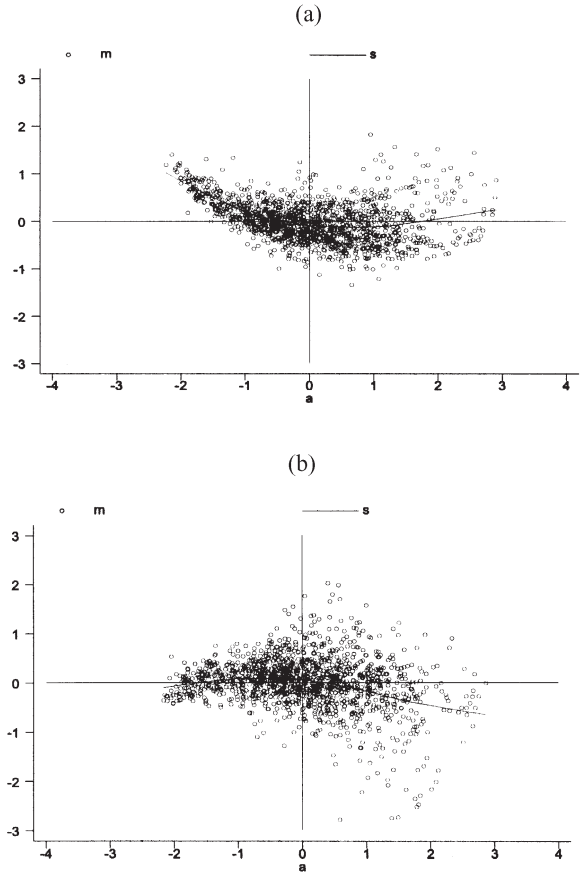

Figure 1. The difference M in normalized log-intensities for wild-type and mutant tissue plotted against the average normalized log-intensity A, shown for all genes. The intensity data for (a) and (b) are respectively from arrays a_1 and a_2 of the case study and are not background-corrected

experimental samples and AA measures the mean intensity (both on a logarithmic scale). As $\bar{y}_1$ and $\bar{y}_2$ are averages over both arrays and colours, the effects of these two factors are neutralized.

Thus, the MA plot in this case will show a relationship between differential expression and average intensity that is free of any additive colour or array influences. Figure 2 shows the plot of MM against AA for the case study-data (arrays a_1 and a_2). Again, the data are normalized by dye, array and type across all genes so the MA plot has axes centred on zero. Differential expression here represents the difference between wild-type and mutant tissue and, hence, is consistent with the plots in Figure 1 in this respect. A lowess function has been fitted to the plot. Observe its approximate linearity and the relatively uniform scatter of points about the smooth function. Observe also the steady decline in MM with average

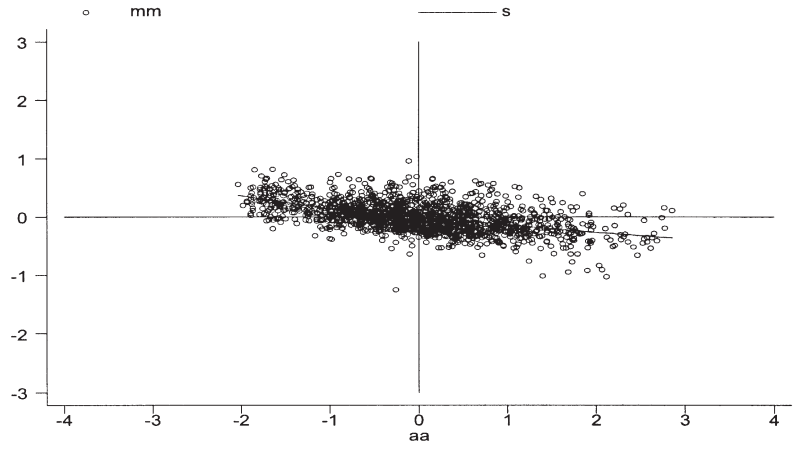

Figure 2. The difference *MM* in normalized log-intensity for wild-type and mutant tissues plotted against the average normalized log-intensity *AA*, shown for all genes. The design is a reversed-colour design. The intensity data are from arrays a_1 and a_2 of the case study and are not background-corrected

intensity *AA*. The correlation coefficient for *MM* and *AA* across all genes is -0.51 in this plot.

4. The concern about intensity-dependent normalization

Figure 2 shows a strong, nearly linear, relationship between differential expression and average intensity, even after cancelling out possible gene–array and gene–colour interactions. Intensity-dependent normalization, applied to this figure, would use the quantities

$$MM^{(d)} = MM - \ell(AA)$$

as a basis for judging differential expression, where *MM* and *AA* are the quantities calculated in (4). Our concern is that the adjustment $\ell(AA)$ being applied here yields biased estimates of differential expression for genes. We contend that, using a reversed-colour design, the study of a gene under two experimental samples produces two correlated estimates, a differential expression estimate *MM* and an average intensity estimate *AA*. These reflect two separate characteristics of expression for a gene, when comparing two experimental samples. There is no need to adjust the difference *MM* for its relationship with *AA* to create a centred value $MM^{(d)}$. The estimate *MM* is an unbiased estimate of differential expression under an ANOVA model for gene expression. The correlation of *MM* and *AA* is an empirical biological feature. To now adjust *MM*, using the intensity-dependent quantity $\ell(AA)$, introduces a bias in the estimate of differential expression.

To give a scenario that explains why intensity-dependent normalization should not be carried out here, imagine (in the context of our case study) that wild-type tissue has an assortment of genes that are absent in mutant tissue and, hence, the former show intensities ranging from low to high, whereas the latter show only background intensities (i.e., noise levels). Thus, the MM for these genes tend to be positive, while the averages AA are negative (on our centred log-scale). Furthermore, imagine that another set of genes are expressed in both tissues but are, nonetheless, weakly to strongly upregulated in mutant tissue relative to wild-type tissue. The differences MM for these genes tend to be negative, while the averages AA are positive. This is the pattern of the plot in Figure 2.

Intensity-dependent normalization based solely on $MM^{(d)}$ may misidentify differentially expressed genes in this setting because of the adjustment $\ell(AA)$. To illustrate this point, refer to the results in Table 3. In Lee *et al*. (2002), where the data from all four arrays are analysed, 12 genes are declared as differentially expressed. Columns 1 and 3 show the top 12 genes shown as differentially expressed based on the $MM^{(d)}$ and MM, respectively. Columns 2 and 4 show the MM and $MM^{(d)}$ values. The list for MM is identical to that in Lee *et al*. (2002) based on the full data set (i.e., all four arrays). In contrast, only 7 of these 12 genes are found in the $MM^{(d)}$ list. These are segregated in the table by a horizontal line. Not surprisingly, these seven genes are those for which both MM and $MM^{(d)}$ are largest. The adjustment $\ell(AA)$ is not large enough to affect their top ranked position, although their rank order is affected.

In conclusion, we raise the concern that intensity-dependent normalization can give

Table 3. Lists of differentially expressed genes with and without intensity-dependent adjustment. Statisitcs MM and $MM^{(d)}$ denote the respective differential expression estimates between wild-type and mutant tissue. Negative signs denote genes upregulated in mutant tissue

No intensity-dependent adjustment		Intensity-dependent adjustment	
(1) Gene	(2) MM	(3) Gene	(4) $MM^{(d)}$
1238	−0.758	1216	0.693
1691	−0.773	1006	0.697
1584	0.809	619	0.699
1224	0.816	1181	0.700
408	−0.830	1666	0.712
1198	−0.902	1198	−0.756
293	−0.918	293	−0.804
401	−0.948	1347	−0.812
1229	0.956	401	−0.816
1347	−1.014	1038	−0.887
1038	−1.028	1229	0.950
1560	−1.249	1560	−1.349

biased estimates of differential expression and, therefore, can misclassify some moderately important genes as unexpressed. We suggest that, for microarray studies with dye reversal, the pairs of values (*MM*, *AA*) be examined *jointly* as providing important but separate information about relative gene expression under two experimental samples. As a logical extension, we caution against the use of intensity-dependent normalization in the analysis of data from other microarray experiments. As a final remark, we note that the identification of this problem was made possible by the reversed-colour design embedded in this case-study design. We adhere to the advice of Kerr and Churchill (2001) and Kerr *et al*. (2001) that colour reversal should be incorporated in experimental designs for cDNA microarray studies.

Acknowledgements

This research was supported in parts by National Institutes of Health grants CA89756 and HG02510 (Lee), and the Natural Sciences and Engineering Research Council of Canada (Whitmore).

References

Dudoit, S., Yang Y.H., Callow, M.J. and Speed, T.P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist. Sinica*, **12**, 111–139.

Eisen, M.B. and Brown, P.O. (1999) DNA arrays for analysis of gene expression. In S.M. Weissman (ed.), *cDNA Preparation and Characterization*, Methods in Enzymology 303, pp. 179–205. San Diego, CA: Academic Press.

Kerr, M.K. and Churchill, G.A. (2001) Experimental design issues for gene expression microarrays. *Biostatistics*, **2**, 183–201.

Kerr, M.K., Martin, M. and Churchill, G.A. (2001) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Lee, M.-L.T., Lu, W., Whitmore, G.A. and Beier, D. (2002) Models for microarray gene expression data. *J. Biopharmaceutical Statist.*, **12**(1), 1–19.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J. and Speed, T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**(4), e15.

Received April 2003