BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 10 2005, pages 2430–2437
doi:10.1093/bioinformatics/bti378

*Gene expression*

# Characterizing dye bias in microarray experiments

K. K. Dobbin[1,*], E. S. Kawasaki[2], D. W. Petersen[2] and R. M. Simon[1]

[1]Biometric Research Branch and [2]Advanced Technology Center, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA

Received on December 6, 2004; revised on March 3, 2005; accepted on March 5, 2005
Advance Access publication March 17, 2005

**ABSTRACT**

**Motivation:** Spot intensity serves as a proxy for gene expression in dual-label microarray experiments. Dye bias is defined as an intensity difference between samples labeled with different dyes attributable to the dyes instead of the gene expression in the samples. Dye bias that is not removed by array normalization can introduce bias into comparisons between samples of interest. But if the bias is consistent across samples for the same gene, it can be corrected by proper experimental design and analysis. If the dye bias is not consistent across samples for the same gene, but is different for different samples, then removing the bias becomes more problematic, perhaps indicating a technical limitation to the ability of fluorescent signals to accurately represent gene expression. Thus, it is important to characterize dye bias to determine: (1) whether it will be removed for all genes by array normalization, (2) whether it will not be removed by normalization but can be removed by proper experimental design and analysis and (3) whether dye bias correction is more problematic than either of these and is not easily removable.

**Results:** We analyzed two large (each $>$27 arrays) tissue culture experiments with extensive dye swap arrays to better characterize dye bias. Indirect, amino-allyl labeling was used in both experiments. We found that post-normalization dye bias that is consistent across samples does appear to exist for many genes, and that controlling and correcting for this type of dye bias in design and analysis is advisable. The extent of this type of dye bias remained unchanged under a wide range of normalization methods (median-centering, various loess normalizations) and statistical analysis techniques (parametric, rank based, permutation based, etc.). We also found dye bias related to the individual samples for a much smaller subset of genes. But these sample-specific dye biases appeared to have minimal impact on estimated gene-expression differences between the cell lines.

**Contact:** dobbinke@mail.nih.gov

**Availability:**

**Supplementary information:** http://linus.nci.nih.gov/~brb/TechReport.htm

## INTRODUCTION

In dual label microarray experiments, the fluorescent intensity of a dye in a spot on the microarray serves as a measure of the amount of the mRNA in the original sample resulting from transcription of the gene corresponding to the cDNA or oligonucleotide printed on that spot. But, in both direct and indirect labeled experiments, the fidelity of the intensity measurement to the underlying gene expression may be different for the two dyes.[1] To fix ideas, consider a series of self–self hybridization arrays, where the same RNA sample is tagged with both dyes during separate reverse transcriptions and hybridized to each array. For a particular gene, the Cy3/green channel may appear consistently brighter than the Cy5/red channel, despite the fact that there are no real differences in expression. This phenomenon is usually called dye bias (Tseng *et al*., 2001; Kerr *et al*., 2002; Dobbin *et al*., 2003a,b; Dombkowski *et al*., 2004; Rosenzweig *et al*., 2004), because it could potentially introduce bias into comparisons.

Dye bias can be subdivided into four different types: (1) dye bias that is the same for all genes on an array, causing one channel to appear brighter overall than the other; (2) dye bias that depends on the overall spot intensity, and is different for bright spots than for dim spots; (3) dye bias that is associated with some subset of genes, but is consistent for the same gene across samples; (4) dye bias that depends on a combination of characteristics of the sample as well as the gene. Dye bias of type (1) should be eliminated by the usual array normalization procedures (e.g. median centering of arrays, loess normalization), and loess normalization (Yang *et al*., 2002) is designed to eliminate bias of type (2). We will call type (3) a gene-specific dye bias because the bias is different for different genes, but the same for a given gene across all samples in an experiment. (Note that we are using the convention to refer to spots as genes, as in 'gene-specific dye bias,' although in fact not every spot is always associated with a unique gene, so that this would more properly be referred to as 'feature-specific dye bias'.) This type of bias can be eliminated by statistical design and analysis. We will call type (4) a gene- and sample-specific dye bias because it depends on both the gene and the sample being analyzed. This type of bias is more difficult to eliminate.

This paper investigates gene-specific dye bias and gene- and sample-specific dye bias. Gene-specific dye bias will not affect comparisons between samples or classes of samples labeled with the same dye, because the bias will cancel out of the comparisons. A 'reference design' is a design in which each array includes a common reference sample consistently labeled with the same dye. Even in

*To whom correspondence should be addressed.

[1]In direct labeled experiments, the efficiency of the incorporation of a dye during the reverse transcription of the mRNA may depend on the transcript's particular nucleotide sequence, and this incorporation efficiency may be different for the two dyes used in an experiment. Indirect labeling (Manduchi *et al*., 2002), which was used in the experiments presented here, lessens the effect of incorporation efficiency, but the quantum efficiencies and stabilities of the dyes are different, which can produce a phenomenon similar to differential incorporation efficiency.

(a)

| | Array 1 | Array 2 | Array 3 | Array 4 | Array 5 | Array 6 |
|---|---|---|---|---|---|---|
| Cy3 | Tumor | Tumor | Tumor | Normal | Normal | Normal |
| Cy5 | Reference | Reference | Reference | Reference | Reference | Reference |

(b)

| | Array 1 | Array 2 | Array 3 | Array 4 | Array 5 | Array 6 |
|---|---|---|---|---|---|---|
| Cy3 | Tumor | Tumor | Tumor | Tumor | Tumor | Tumor |
| Cy5 | Normal | Normal | Normal | Normal | Normal | Normal |

(c)

| | Array 1 | Array 2 | Array 3 | Array 4 | Array 5 | Array 6 |
|---|---|---|---|---|---|---|
| Cy3 | Tumor | Normal | Tumor | Normal | Tumor | Normal |
| Cy5 | Normal | Tumor | Normal | Tumor | Normal | Tumor |

**Fig. 1.** Examples of dual-label microarray designs. (**a**) A reference design comparing tumor tissue with normal tissue. (**b**) A design comparing tumor tissue with normal tissue. (**c**) A balanced block design comparing tumor tissue with normal tissue.

the presence of gene-specific dye bias, the reference design will not produce biased comparisons among classes of the (non-reference) samples. For example, in Figure 1a, comparisons between the tumor and normal tissues will not be biased.

Gene-specific dye bias may affect comparisons between samples or classes of samples labeled with different dyes, because the bias may not cancel out of the comparisons. For example, Figure 1b shows an experiment in which comparisons of the tumor tissue with the normal tissue will be affected by gene-specific dye bias. Since all the tumor tissues are labeled with Cy3 and all the normal tissues with Cy5, observed differences between the two tissue types may be attributable to either the gene-specific dye bias or to real differences in gene expression between the tumor and normal tissues. The design in Figure 1b is said to 'completely confound' (Cochran and Cox, 1992) the gene-specific dye bias with the tissue type distinction, because the two cannot be separated. Figure 1c shows a balanced block design in which half the tumor samples are labeled with Cy3 and the other half with Cy5, and the same for the normal samples. This labeling strategy removes the gene-specific dye bias from the comparisons of the tumor samples and the normal samples, and proper statistical analysis can result in a significant increase in efficiency compared with the reference design (Dobbin and Simon, 2002). Intuitively, the adjusted statistical analysis addresses the question: if these samples were all labeled the 'same way,' would there be significant differences between the gene-expression measurements? The 'same way' could be interpreted to mean that they were all labeled with Cy3 or with Cy5, or labeled with both the dyes and the average over the two dyes calculated. All three of these interpretations of 'same way' will produce identical statistical inference, which is why the dye bias is said to be eliminated from the analysis. Biases are not always so easy to eliminate. For instance, gene- and sample-specific dye bias is not subject to this type of statistical correction.

Gene- and sample-specific dye bias is more problematic. This type of dye bias is affected by characteristics of the sample as well as the gene, and was proposed by Dombkowski *et al*. (2004). In the presence of this type of dye bias, there is no straightforward way to analyze the data so as to eliminate the dye bias, as was the case with gene-specific dye bias. The reason being, that one can no longer answer in a general way the question: 'If all the samples were measured in the "same way," would there be significant differences?' because the answer will depend on how one defines the 'same way.' For instance, the answer will be different if one interprets 'same way' to mean all labeled with Cy3, than if one interprets it to mean all labeled with Cy5, or to mean all labeled with both dyes and the average of the two intensities calculated. Each of these three interpretations will produce different statistical inferences (for proof, see Supplement 2 of Supplementary data). But there is no a priori reason to choose one of these definitions over the others. Hence, there enters an arbitrary decision into the process which will affect the conclusions of the statistical analysis and truly valid, objective analysis is not possible. In particular, even dye-swapping every array will not allow one to perform valid statistical analyses free of gene- and sample-specific dye bias (although this was suggested by Dombowski, *et al*., 2004).

Previous preliminary findings related to gene-specific dye bias (Tseng *et al*., 2001; Kerr *et al*., 2002; Dobbin *et al*., 2003b; Rosenzweig *et al*., 2004) in direct labeled experiments have been highly tentative because of the small sample sizes used. Dombkowski *et al*. (2004) encountered gene- and sample-specific dye bias, but did not quantify the phenomenon adequately to assess its impact. This paper attempts to address the shortcomings of the previous studies by (1) analyzing larger datasets with sufficient replication to assure robust estimation and inference in gene-specific models, (2) considering both types of dye bias separately, (3) analyzing data from multiple platforms, and data that utilized indirect labeling technology and (4) explaining clearly the impact of these findings for statistical design and analysis of future microarray studies.

## MATERIALS AND METHODS

### Experimental description

*Preparation of cDNA and oligonucleotide arrays* Microarrays were manufactured at the NCI Microarray Facility, Advanced Technology Center, Gaithersburg, MD. Arrays with ∼10 000 cDNAs were prepared from ready to print UniGEM2 libraries obtained from Incyte, Inc. (Wilmington, DE). Human Genome Oligo Set Version 2.0 Oligo libraries containing ∼22 000 oligonucleotides of 70 bases in length were obtained from Operon, Inc. (Alameda, CA). Arrays were printed by standard protocols on Corning UltraGAPS II slides (Corning, NY) using a GeneMachine® (San Carlos, CA) OmniGrid 100 instrument. cDNAs were suspended at a concentration of 100 μg/ml and oligonucleotides at 25 μM in 3× SSC buffer, and the arrays printed using SMP3 pins from Telechem International (Sunnyvale, CA). The spotted nucleic acids were fixed to the slides and blocked with protocols supplied by the manufacturer.

*Cell lines and RNAs* Growth of cell lines and RNA isolation was done at the core Gene Expression Laboratory at NCI-Frederick. MCF10A (benign mammary epithelial), LNCAP (prostate carcinoma), Jurkat (T-cell lymphoma), SUDHL6 (germinal center B-cell like diffuse large B-cell lymphoma), OCI-Ly6 (activated B-cell like diffuse rare B-cell lymphoma) and L428 (Hodgkin's lymphoma) were grown under standard conditions (Chen *et al*., 1996), and RNA was isolated from the cells using TriReagent following the manufacturer's protocol (Molecular Research Center, Inc., Cincinnati, OH). The integrity of the RNA was confirmed by analysis with the Agilent 2100 Bioanalyzer (Palo Alto, CA) using the RNA 6000 LabChip®kit. As a control RNA, Human Universal Reference RNA (HUR RNA) was purchased from Stratagene (La Jolla, CA).

*Labeling and purification of targets* Labeled cDNA for the long oligonucleotide and cDNA arrays were synthesized and labeled by the indirect amino-allyl method using reagents and protocols supplied with the Stratagene FairPlay™ Microarray Labeling Kit. For cDNA synthesis, Stratascript reagents (Stratagene) were used, and Cy3/Cy5 fluorophore amino-allyl reagents were obtained from Amersham (Piscataway, NJ). For each synthesis, 20 μg of total RNA were used. Labeled cDNA targets were purified using Minelute purification kits (Qiagen, Valencia, CA).

*Hybridization and washing of arrays* The cDNA and long oligonucleotide microarrays were prehybridized in 40 μl of 5× SSC, 0.1%

**Table 1.** Six cell lines assayed in experiment

| Cell line | Number of oligonucleotide arrays (number with reference green/Cy3) | Number of cDNA arrays (number with reference green/Cy3) | Cell line description |
|---|---|---|---|
| MCF10a | 4 (2) | 4 (2) | Human mammary epithelial cell line |
| LNCAP | 4 (2) | 4 (2) | Human prostate cancer cell line |
| L428 | 9 (4) | 7 (4) | Hodgkin's disease cell line |
| SUDHL | 4 (2) | 4 (2) | Human lymphoma cell line |
| OCILY3 | 5 (3) | 5 (3) | Human lymphoma cell line |
| Jurkat | 4 (2) | 4 (2) | Human T lymphocyte acute T cell leukemia cell line |
| Total | 30 (15) | 28 (15) | |

SDS and 1% BSA at 42°C for 30 min. The prehybridization solution was removed and arrays were hybridized for 16 h at 42°C in 5× SSC buffer containing Cy3/Cy5 labeled targets, 25% formamide, 0.1% SDS, 1 μg Cot-1 DNA and 1 μg poly A RNA. The cDNA arrays were washed at room temperature in 2× SSC, 0.1% SDS for 2 min, 1× SSC for 2 min, 0.2× SSC for 2 min and 0.05× SSC for 1 min. The long oligonucleotide arrays were treated the same except for the omission of the last wash step. The slides were dried by spinning at 650 r.p.m for 3 min.

*Array scanning and image processing* Long oligonucleotide and cDNA arrays were scanned using Axon 4000B scanner at 10 μm resolution. Image processing and quantification of signal values of spotted arrays were performed using Genepix 3.0 software (Axon Instruments, Union City, CA). The Genepix result files including signal, background, standard deviation, pixel statistics and quality parameters of both channels were deposited in the microarray database (mAdb) maintained by NCI/CIT bioinformatics group (Greene *et al*., 2003).

## Data analysis

For two dual-label experiments, one with cDNA arrays and the other with printed oligonucleotide arrays, Stratagene universal human reference RNA was used as a standard for testing with RNA from cell lines MCF10a, LNCAP, L428, SUDHL, OCILY3 and Jurkat. All arrays were dye-swapped at least twice. There were a total of 28 cDNA arrays and 30 oligonucleotide arrays. Table 1 gives a description of the cell lines and experimental design.

Data were background corrected by subtracting the median background pixel intensity from the mean foreground intensity, because the median background subtraction makes the tiny dust particles less significant and the mean foreground is preferable to the median foreground for spots that lack signal in the center, called doughnuts. Signals <100 were truncated to 100. Spots flagged for poor quality were eliminated from the analysis and genes with missing data were eliminated. The reason we eliminated genes with missing data is that analyses with missing data may result in either inestimable parameters or significant power loss when compared with complete data. This left 8604 of the 9069 genes on the cDNA arrays for analysis, and 15 790 of the 21 794 genes on the oligonucleotide arrays for analysis. Normalizations using both median centering of arrays and loess smoothing (Yang *et al*., 2002) yielded very similar results. We present the median centering results here (with the exception of Figure 4).

For each gene, the general analysis of variance model for the data was

$$\mathrm{Log}_2\left[\frac{T_{\mathrm{cor}}}{R_{\mathrm{cor}}}\right] = \mu + C_c + O_o + CO_{\mathrm{co}} + \varepsilon_{\mathrm{cor}}, \qquad (1)$$

where $\mathrm{Log}_2$ is the base 2 logarithm, $T_{\mathrm{cor}}$ and $R_{\mathrm{cor}}$ are the background-corrected, normalized intensity in the target (cell line) channel and the reference channel, respectively; $\mu$ represents the overall mean log-ratio; $C_c$ are the cell line effects, for the six cell lines, representing differences in expression among the cell lines, $c = 1, 2, \ldots, 6$; $O_o$ are the orientation effects, representing gene-specific dye bias, for each dye orientation (e.g. target labeled with Cy3 or Cy5) $o = 1, 2$; $CO_{\mathrm{co}}$ are the cell line by orientation interactions, representing the gene- and sample-specific dye bias; and $\varepsilon_{\mathrm{cor}}$ is independent, normally distributed error. The usual parameter constraints ensure identifiability (Cochran and Cox, 1992).

Equation (1) models the log-ratios instead of the log-intensities, as is often done in microarray analysis of variance studies. But we have shown that the log-ratio and log-intensity models are equivalent, and provided a one-to-one mapping of the model parameters (Dobbin and Simon, 2005).

The analysis of variance tables for both experiments are given in Table 1 of the Supplementary materials. Importantly the table provides strong evidence that the sample sizes for this study are adequate, allowing 16 or more degrees of freedom for error in each case, in contrast to previous studies that had inadequate error degrees of freedom for robust gene specific analyses.

We believe that this dataset is most appropriately analyzed using generalized least squares (Carroll and Ruppert, 1982a,b; Pinheiro and Bates, 2000), because many genes displayed large heteroscedasticity of error variance for different cell lines (see Supplementary materials for further discussion and motivation). However, since our goal is to characterize dye bias in as broad a context as possible we have analyzed gene-specific dye bias using a wide range of parametric, rank-based, and permutation-based analysis methods. In addition, we have considered both median normalization and global loess normalization, and further considered a range of parameter settings for the loess normalization to ensure the robustness of these findings. In particular, the loess smoothing parameter alpha (Cleveland *et al*., 1992), which controls the degree of smoothing, was varied through the range 0.4–3.0, values outside this range appearing to grossly over- or under-smooth.

# RESULTS

We analyzed dual-label microarray data from both a cDNA experiment and an oligonucleotide experiment.

## Gene-specific dye bias

First we consider the cDNA experiment. We analyzed the normalized, background-corrected data separately for each gene. Table 2 shows the results of multiple analyses of the data. In assessing gene-specific dye bias, we considered three approaches:

(1) Make no adjustment for cell line heterogeneity;

(2) Make an adjustment only for differences in the mean or median expression in the different cell lines;

(3) Make adjustment for both differences in the mean expression in the different cell lines and for differences in the variances of expression in the different cell lines.

These analyses can be viewed as covering a range from most naïve (1) to least naïve (3). In Equation (1), the $P$-values in Table 2

**Table 2.** Analyses of gene-specific dye bias

| Cell line heterogeneity taken into account? | Cell line main effects adjustment | Significance test | Number of genes with gene-specific $P$-value $< 0.001$ | Median absolute value of dye bias for significant genes |
|---|---|---|---|---|
| No | None | Pooled $t$-test | 893 (10%) | 0.28 (1.2-fold) |
| | | Welch $t$-test | 869 (10%) | 0.28 (1.2-fold) |
| | | Wilcoxon rank–sum test | 901 (10%) | 0.43 (1.3-fold) |
| | | Permutation test[a] | 1014 (12%) | 0.67 (1.6-fold) |
| Location (mean/median) heterogeneity only | Mean-centering | Pooled $t$-test | 2035 (24%) | 0.57 (1.5-fold) |
| | | Welch $t$-test | 1932 (22%) | 0.58 (1.5-fold) |
| | | Wilcoxon rank–sum test | 2015 (23%) | 0.54 (1.5-fold) |
| | | Permutation test[a] | 2281 (27%) | 0.56 (1.5-fold) |
| | Median-centering | Pooled $t$-test | 1806 (21%) | 0.58 (1.5-fold) |
| | | Welch $t$-test | 1679 (20%) | 0.59 (1.5-fold) |
| | | Wilcoxon rank–sum test | 1758 (20%) | 0.48 (1.4-fold) |
| | | Permutation test[a] | 2229 (26%) | 0.56 (1.5-fold) |
| Location and scale (variance) heterogeneity | Mean-centering | Generalized least squares[b] | 3388 (39%) | 0.46 (1.4-fold) |
| | | Permutation test[c] | 3310 (38%) | 0.47 (1.4-fold) |
| Expected by chance | | | 9 (0.1%) | |

Using notation from Equation (1), let $Y_{\text{cor}} = \log_2[T_{\text{cor}}/R_{\text{cor}}]$. Mean-centering corresponds to the transformation $Z_{\text{cor}} = Y_{\text{cor}} - \bar{Y}_{c\bullet\bullet}$, where $\bar{Y}_{c\bullet\bullet}$ is the mean over the cell line; similarly, median-centering is $W_{\text{cor}} = Y_{\text{cor}} - \tilde{Y}_{c\bullet\bullet}$, where $\tilde{Y}_{c\bullet\bullet}$ is the median over the cell line. $T$-tests have the form $(\bar{Y}_{\bullet 1\bullet} - \bar{Y}_{\bullet 2\bullet})/\text{SD}$, where SD is the estimated standard deviation of the numerator, and $W$ or $Z$ are inserted for $Y$ as appropriate. SD is estimated under the assumption of equal variance for the pooled $t$-test, and unequal variances for the Welch $t$-test. Wilcoxon rank–sum test indicates the Wilcoxon two-sample test performed on the $Y$s, $W$s or $Z$s as appropriate. Permutation tests are based on pooled $t$-statistics with 10 000 permutations except as noted.

[a] Permutation test based on 10 000 permutations of the dye labels.

[b] Generalized least squares model fit with different error variance for each cell line stratum. $P$-values calculated via restricted maximum likelihood used to fit the model and conditional $F$-tests used to assess significance. Likelihood-ratio tests with maximum-likelihood estimates (not shown) produced virtually identical results.

[c] Permutation test based on 10 000 permutations of the dye labels within cell lines. Test statistic used is weighted sum of $t$-test numerators, with weights equal to inverse estimated variance for cell line $t$-test numerator.

corresponds to the statistical hypothesis test that each of the $O_o$ orientation effects terms is zero.

First, note that in all the analyses in Table 2, the number of genes which display statistically significant dye bias is much greater than the number expected by chance; the number of genes range from 869 to 3388, whereas only 9 are expected by chance. Second, note that for a given approach to cell line heterogeneity adjustment (i.e. either no adjustment, adjustment for location differences only or adjustment for both location and scale differences), the extent of the dye bias is extremely similar across a range of analytic techniques, from t-tests to rank-sum tests to permutation tests. If no cell line adjustment is made dye bias is observed in 10–12% of genes; if only a location cell line adjustment is made dye bias is observed in 20–27% of genes; if both location and scale cell line adjustment is made dye bias is observed in 38–39% of genes. In all cases, the percentage of gene-specific dye bias genes greatly exceeds the percentage expected by chance.[2]

Results as to the overall extent of dye bias were very similar for loess normalization under a range of different parameter settings for the loess fit (Table 4 of Supplementary material).

Table 2 also presents data on the size of the gene-specific dye bias for the statistically significant genes (rightmost column). In Equation (1), these correspond to the estimated $\hat{O}_o$ values from the model fit. The median effect size of the dye bias for genes which display dye bias ranges from 1.4-fold to 1.5-fold for the non-naïve analyses (with location only or location and scale adjustment). In the generalized least squares analysis, 125 (1.5%) of the genes have an estimated gene-specific dye bias >2-fold, and one gene has an estimated gene-specific dye bias >4-fold.

**Table 3.** cDNA agreement between models with and without gene-specific dye bias adjustments included

| | | All data: no dye bias adjustment $P$-value $< 0.001$ | All data: no dye bias adjustment $P$-value $> 0.001$ |
|---|---|---|---|
| All data: dye bias adjustment | $P$-value $< 0.001$ | 4801 (56%) | 559 (6%) |
| | $P$-value $> 0.001$ | 81 (1%) | 3163 (37%) |

$P$-values are for the $F$-test of no differential expression among any of the six cell lines. 'Dye bias adjustment' $P$-values are from fitting the model $\text{Log}_2[T_{\text{cor}}/R_{\text{cor}}] = C_c + O_o + \varepsilon_{\text{cor}}$ and 'no dye bias adjustment' $P$-values are from fitting the model $\text{Log}_2[T_{\text{cor}}/R_{\text{cor}}] = C_c + \varepsilon_{\text{cor}}$. The $P$-values are from the hypothesis test that each $C_c$ is zero.

To assess the impact of gene-specific dye bias on inferences about the cell lines, we fit a model that only accounted for differences between the cell lines (ignoring gene-specific dye bias), and a model that accounted for both differences between the cell lines and gene-specific dye bias, to compare the results. In particular, we used generalized least squares to fit the model $\text{Log}_2[T_{\text{cor}}/R_{\text{cor}}] = C_c + \varepsilon_{\text{cor}}$ and the model $\text{Log}_2[T_{\text{cor}}/R_{\text{cor}}] = C_c + O_o + \varepsilon_{\text{cor}}$. If the gene-specific dye biases (represented by $O_o$) are trivial, then the two models should lead to nearly identical statistical inference. Agreement between

[2] That is, the number expected to be observed at this significance level if, in fact, no genes are differentially expressed. This expected number generally depends on the number of genes on the array and the statistical test assumptions, but not on the correlation structure among the genes. Refer Dobbin and Simon (2005) and Supplementary section 2 for an example of how correlation does not impact the calculation of expected number.

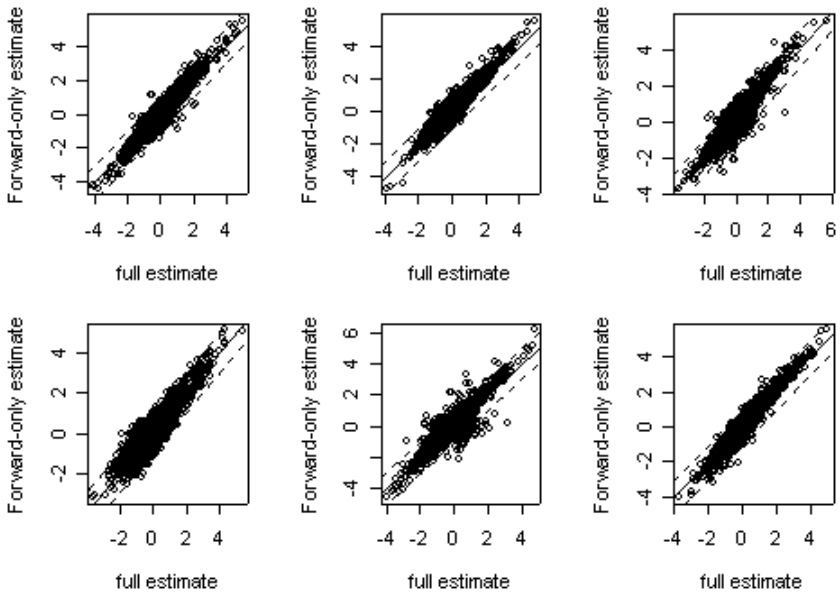


**Fig. 2.** Each plot shows the estimated expression effect sizes for all 8604 genes for one of the six cell lines. $x$-axis is the estimated effect sizes using all the arrays. $y$-axis is the estimated effect sizes using only the forward-labeled arrays. Drawn on the plots are a 45° line through the origin and lines ±1 above and below this line. Top row is MCF10a, LNCAP and L428 (left to right) and bottom row is SUDHL, OCILY3 and Jurkat. Pearson correlations are 0.91, 0.92, 0.87, 0.84, 0.86 and 0.94.

the $P$-values from the overall $F$-tests of any differential expression among the cell lines is shown in Table 3; each $F$-test tests the hypothesis that all the $C_c$ cell line main effects terms are zero. For 4801 genes (56%) both models indicated that gene expression varied among all lines. For 3163 genes (37%) both models indicated that gene expression did not vary across all lines. The models gave discrepant results for 640 genes (7%), i.e. one model found there was significant differential expression and the other found that there was no significant differential expression among the cell lines. For 559 genes (6%), the model with gene-specific dye bias found the genes significantly differentially expressed and the model without gene-specific dye bias found them not significant; so the dye bias appears to have masked the true differential expression. For 81 genes (1%), the discrepancy was in the opposite direction, so the dye bias appears to have led to 'false-positive' detection of differential expression for these genes. The observed imbalance in discrepancies is to be expected because for nearly balanced data like this, gene-specific dye bias will tend to mask true gene-expression differences rather than create 'false-positives.'

The results were very similar for the oligonucleotide arrays (see Supplementary material). While the higher proportion of filtered genes on the oligonucleotide arrays (∼28% versus ∼5% on the cDNA arrays) results in greater uncertainty as to the true extent of dye bias on this platform, the overall similarity of dye bias we observed on both platforms suggests that filtered genes may not systematically differ from unfiltered genes with regard to dye bias.

## Gene- and sample-specific dye bias

We next consider gene- and sample-specific dye bias. Based on the high concordance across analytic methods for the gene-specific dye bias, we restrict presentation to the generalized least squares analysis.

There appear to be far fewer genes with significant gene- and sample-specific dye bias than there were with gene-specific dye bias. But there do appear to be more genes with gene- and sample-specific dye bias than we would expect by chance. There were 1029 genes with $P$-values $< 0.05$, as compared with 430 expected by chance; and there were 150 with $P$-values $< 0.001$, as compared with 9 expected by chance.

Comparing the gene-specific dye bias to the gene- and sample-specific dye bias,[3] we find 3388 genes with gene-specific $P$-value $<$ 0.001 compared with 150 genes with gene- and sample specific $P$-value $< 0.001$. The relative sizes of the bias for the significant genes was similar; the 3388 genes with significant gene-specific dye bias $P$-value $< 0.001$ had bias with median absolute value 0.46 (1.4-fold), whereas the 150 genes with gene- and sample-specific dye bias had bias with median absolute value 0.27 (1.2-fold).

One test of the importance of the gene- and sample-specific dye bias on statistical inference is to compare the estimated differences in gene expression between the cell lines using all the dye swap arrays with those same estimates using only arrays with one labeling (e.g. with Stratagene labeled Cy3). Figure 2 shows plots of the estimated sizes of the differences in expression between each of the six cell lines and the overall average of the cell lines across the 8604 genes, using both the full dataset with all 28 arrays and using only the subset of 15 arrays that were all run with the same orientation (Stratagene labeled with Cy3/green dye). The estimates fall close to a 45° line through the origin, indicating good agreement between the dye swap estimates and the forward-only estimates. Table 4 shows the numbers of discrepancies that are large in estimated size when using the full dataset versus using only the forward-labeled arrays. In

[3]Gene-specific dye bias estimates represent the average amount by which one channel tends to show up brighter than the other. When gene- and sample-specific dye bias is also present, this overall trend in the average is supplemented by a sample-specific trend, so that the dye bias may be different in size or direction for a particular cell line.

**Table 4.** Numbers of large discrepancies between cell line effect size estimates based on the full dataset compared with estimates based on the forward-only arrays

| | Total estimated cell line effects | Number with discrepancy >1 | Number with discrepancy >2 |
|---|---|---|---|
| MCF10a | 8604 | 20 (0.2%) | 0 |
| LNCAP | 8604 | 4 (0.05%) | 0 |
| L428 | 8604 | 42 (0.5%) | 4 (0.05%) |
| SUDHL | 8604 | 29 (0.3%) | 0 |
| OCILY3 | 8604 | 50 (0.6%) | 5 (0.06%) |
| Jurkat | 8604 | 22 (0.3%) | 0 |

all cases, <1% of genes display estimated discrepancies >1 (2-fold). Very similar results were obtained when: (1) estimates derived from only the forward run arrays were compared with those derived from only the backward run arrays; (2) the oligonucleotide arrays were examined (Supplementary material).

In conclusion, although there is some evidence that gene- and sample-specific dye bias may exist for a subset of genes, the impact of this bias on estimated differences between gene expressions in the cell lines appears to be minor.

## Autofluorescence

We investigated the potential that dye bias was related to spot brightness by breaking down the dye bias estimates into groups based on median intensity of the Cy3/green dye. Since the experiments are nearly balanced, median intensity in the Cy3/green channel serves as a measure of the median amount of cDNA present across samples. Figure 3 shows the results. The significantly increasing trend suggests that a component of dye bias may be attributable to post-normalization median intensity-related effects. Interestingly, global loess normalization, which is designed to address intensity-dependent dye bias on an array-by-array basis, reduced this phenomenon slightly but did not eliminate it; in fact, loess resulted in a reversal of the direction of the apparent bias (Fig. 4), suggesting the loess methodology was overadjusting the data.

The observed relation between dye bias and spot intensity may be partly attributed to the phenomenon of autofluorescence. Autofluroescence is the tendency of unlabeled cDNA to fluoresce brighter at the lower Cy3/green frequency. Papers have been published describing this phenomenon (Eisinger and Shulman, 1968; Onidas *et al*., 2002; Raghavachari *et al*., 2003). Further discussion appears in the Supplementary section.

## Cross-platform comparison of gene-specific dye bias on cDNA and oligonucleotide arrays

Unigene identifiers were used to match genes across platforms. 6056 genes were matched in this way. When multiple oligonucleotide 70mers matched the same cDNA, the first match was used. Table 5 shows that there was minimal agreement across platforms as to the size and direction of gene-specific dye bias for different genes (correlation 0.12), whereas there was significant agreement as to the size and direction of cell line effects for different genes, indicating that the unigene identifiers are adequately matching corresponding genes on the different platforms. Similarly, agreement across platforms based on a *P*-value cutoff of 0.001 was much smaller for

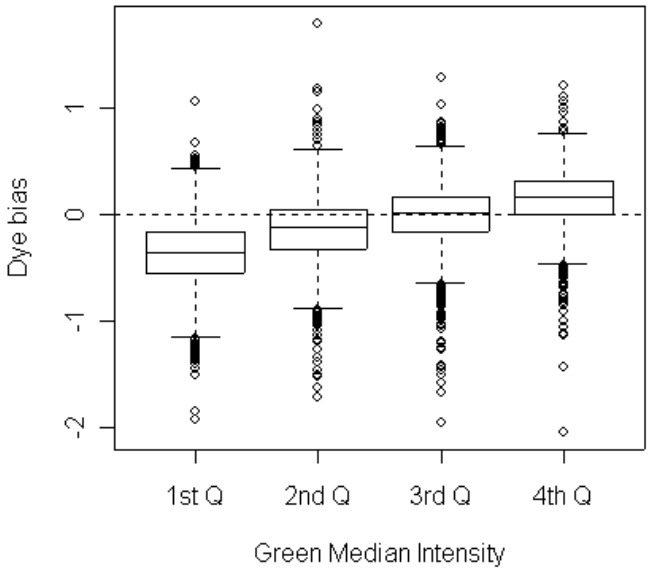


**Fig. 3.** Size of the estimated dye bias as a function of median intensity in the green (Cy3) channel. 1st Q indicates genes with median normalized intensity in the first (lowest) quartile of genes; 2nd Q indicates genes in the second quartile, etc. The brighter the median green channel intensity, the greater the gene-specific dye bias in the direction of the samples with target labeled green (Cy3)—which is the positive direction in this figure.

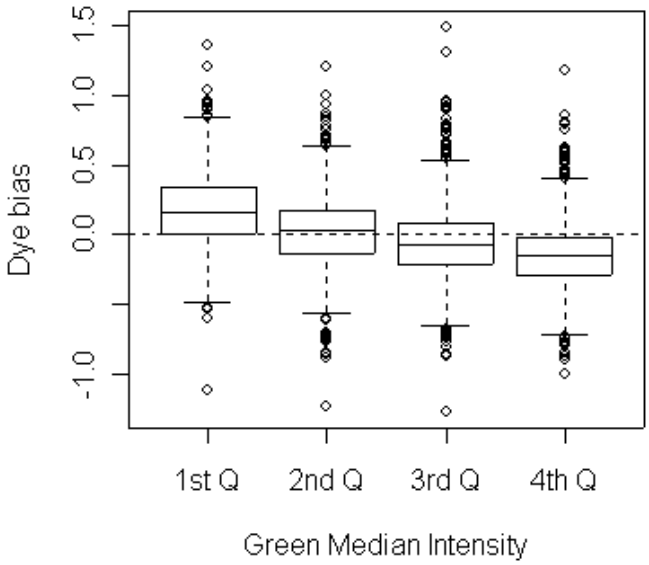


**Fig. 4.** Loess normalized data: size of the estimated dye bias as a function of median intensity quartile in the green (Cy3) channel. The pattern suggests that loess normalization does not remove the dye bias, and appears to over-adjust the data for intensity-dependent dye bias.

dye bias effects than for cell line effects. Cross-platform concordance of gene-specific dye bias might be expected to increase under a more sophisticated feature-matching methodology, e.g. one that uses oligonucleotide sequence information to verify the correct cDNA match (Mecham *et al*., 2004); but the relatively good concordance of cell lines indicates that overall gene-specific dye bias concordance

**Table 5.** Concordance across platform of gene-specific dye bias and of cell line effects

| Source of effect | Cross-platform correlation between average gene effects | Cohen's Kappa based on $2 \times 2$ table using 0.001 $P$-value cutoff |
|---|---|---|
| Dye bias | 0.12 | 0.06 (0.03, 0.08) |
| MCF10a | 0.74 | |
| LNCAP | 0.72 | |
| L428 | 0.66 | 0.30 (0.28, 0.32) |
| SUDHL | 0.69 | |
| OCILY3 | 0.64 | |
| JURKAT | 0.71 | |

Cross-platform agreement between gene-specific effects of dye bias, and each of the individual cell line effects, as measured by pairwise Pearson correlation of the effect estimates, and Cohen's Kappa statistic for $2 \times 2$ tables created by using $P$-value cutoff of 0.001 for $F$-tests. Agreement based on 6056 genes matched using unigene identifiers. For genes with multiple sequences represented on the oligonucleotide arrays, the first sequence in the file was used. Format of Cohen's kappas is: Kappa value (95% confidence interval). $2 \times 2$ tables appear in Table 5 of Supplementary material.

would probably remain minimal. In conclusion, there is some weak concordance across platforms of gene-specific dye bias.

## DISCUSSION

We have analyzed data from both an oligonucleotide and a cDNA microarray experiment to characterize dye bias. We have shown that many genes exhibit statistically significant gene-specific dye bias (39% with $P$-value $< 0.001$ on the cDNA arrays), and tend to appear brighter on average in one dye compared with the other. Gene-specific dye bias was small for the most part, but not insignificant, suggesting that when samples labeled with different dyes are being compared, statistical adjustment for this type of dye bias seems advisable. We showed that failure to adjust for dye bias does affect conclusions about differential expression.[4] In particular, designs, such as the reference design given in Figure 1a and the balanced block design given in Figure 1c appear superior to designs, such as that given in Figure 1b, because the design of Figure 1b makes it impossible to correct for gene-specific dye bias. The other two designs produce class comparisons free of gene-specific dye bias. More examples of designs that allow one to correct for this type of dye bias can be found in Dobbin *et al*. (2003a).

Gene- and sample-specific dye bias appeared statistically significant for a much smaller proportion of genes (2% with $P$-value $<$ 0.001), although still higher proportion than would be expected by chance. Estimated gene-expression differences between the cell lines produced by analysis of only the forward arrays (with reference labeled Cy3), only the backward arrays (with reference labeled Cy5) and all dye swapped arrays, were very similar in direction and magnitude, indicating that gene- and sample-specific dye biases have a minor impact on these estimates. Thus, gene- and sample-specific dye bias appeared to be of little practical concern. We also noted that no experimental design or statistical analysis will enable one to remove this type of dye bias (as discussed in the Results section). Instead, gene- and sample-specific dye bias, if it exists, indicates a limitation of the accuracy of the technology for a subset of genes. In particular, gene- and sample specific dye bias does not justify systematically dye swapping all arrays in an experiment, because such a design will not enable one to eliminate the bias. This type of design has also been shown to be inefficient (Dobbin *et al*., 2003a).

While we have established the existence of gene-specific dye bias and to a lesser extent, gene- and sample-specific dye bias, the causes of these phenomena remain unclear. For instance, what aspect of the gene-sequence spotted on an array causes the dye bias? Is it the actual sequence of the nucleotides, the order in which the spot was printed (Mary-Huard *et al*., 2004), the size or shape (morphology) of the spot, autofluorescence, an inadequacy of the linear additive model used to approximate the data or something else? If dye bias is chiefly related to how the arrays were constructed (location of spots, printing order, size, etc.), then one would expect that the bias would be consistent across a set of arrays with the same construction, which would result in gene-specific dye bias. The fact that the gene-specific dye bias showed so little concordance across the two platforms that we analyzed suggests that dye bias may be chiefly related to aspects of the array construction and therefore, it is important to use a homogeneous set of arrays for any microarray experiment, and to make dye bias correction within array type if different types or versions of arrays are used. Dye bias related to aspects of the original RNA samples would result in gene- and sample-specific dye bias. The fact that we observed such a small level of this type of bias suggests that most dye bias is not attributable to aspects of the original RNA samples.

These results have implications for single-label array experiments, such as Affymetrix arrays, if the labeling and scanning technology is the same as or similar to that used here. Gene-specific dye biases will not affect inference in single-label experiments for the same reason that they do not affect inference in reference design experiments when comparing non-reference samples. But gene- and sample-specific dye biases will affect inference in both dual-label and single-label systems. The problematic nature of removing the gene- and sample-specific bias is not improved under a single-label system. The fact that gene- and sample-specific biases are more difficult to detect in single-label systems should not be taken as evidence of the superiority of single-label systems with regard to gene- and sample-specific dye bias. The potential problem of gene- and sample-specific dye bias is still there, although, as we have shown, it appears to have a relatively minor impact on estimates of interest.

[4]See the supplemental material for some discussion of why the high proportion of genes with gene-specific dye bias produced so little discordance in differential expression calls.

## REFERENCES


Carroll,R.J. and Ruppert,D. (1982a) A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. *J. Am. Stat. Assoc.*, **77**, 878–882.

Carroll,R.J. and Ruppert,D. (1982b) Robust estimation in heteroscedastic linear models. *Ann. Statistics*, **10**, 429–441.

Cleveland,W.S., Grosse,E. and Shyu,W.M. (1992) Local regression models. In Chambers,J.M. and Hastie,T.J. (eds.), *Statistical Models in S*. Wadsworth & Brooks/Cole Advanced Books & Software, pp. 309–376.

Chen,S.L. *et al*. (1996) Isolation and characterization of a novel gene expressed in multiple cancers. *Oncogene*, **12**, 741–751.

Cochran,W.G. and Cox,G.M. (1992) *Experimental Designs, 2nd edn*. John Wiley & Sons, Inc. New York, NY.

Dobbin,K. and Simon,R. (2002) Comparison of microarray designs for class comparison and class discovery. *Bioinformatics*, **18**, 1438–1445.

Dobbin,K. *et al*. (2003a) Statistical design of reverse dye microarrays. *Bioinformatics*, **19**, 803–810.

Dobbin,K. *et al*. (2003b) Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. *J. Natl Cancer Inst. USA*, **95**, 1362–1369.

Dobbin,K. and Simon,R. (2005) Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, **6**, 27–38.

Dombkowski,A.A. *et al*. (2004) Gene-specific dye bias in microarray reference designs. *FEBS Lett.*, **560**, 120–124.

Eisinger,J. and Shulman,R.G. (1968) Excited electronic states of DNA. *Science*, **161**, 1311–1319.

Greene,J.M. *et al*. (2003) The NCI/CIT microarray database (mAdb) system—bioinformatics for the management and analysis of Affymetrix and spotted gene expression microarrays. *Proc. AMIA Annu. Fall Symp.*, 1066.

Kerr,M.K. *et al*. (2002) Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica*, **12**, 203–217.

Manduchi,E. *et al*. (2002) Comparison of different labeling methods for two-channel high-density microarray experiments. *Physiol. Genomics*, **10**, 169–179.

Mecham,B.H. *et al*. (2004) Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res.*, **32**, e74.

Mary-Huard,T. *et al*. (2004) Spotting effect in microarray experiments. *BMC Bioinformatics*, **5**, 63.

Onidas,D. *et al*. (2002) Fluorescence properties of DNA nucleosides and nucleotides: a refined steady-state and femtosecond investigation. *J. Phys. Chem. B*, **106**, 11367–11374.

Pinheiro,J.C. and Bates,D.M. (2000) *Mixed-Effects Models in S and S-plus.* Springer, New York, NY.

Raghavachari,N. *et al*. (2003) Reduction of autofluorescence on DNA microarrays and slide surfaces by treatment with sodium borohydride. *Anal. Biochem.*, **312**, 101–105.

Rosenzweig,B.A. *et al*. (2004) Dye-bias correction in dual-labeled cDNA microarray gene expression measurements. *Environ. Health Perspect.*, **112**, 480–487.

Tseng,G.C. *et al*. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.*, **29**, 2549–2557.

Yang,Y.H. *et al*. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.