Supplement 1 (of 2) to: Characterizing Dye Bias in Microarray Experiments

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

ADDITIONAL INTRODUCTORY MATERIAL

Review of previous studies of dye bias

To our knowledge, there have been no previous comprehensive studies of dye bias with adequate sample sizes to characterize dye bias in gene-specific models, nor studies of these issues for indirect-labeled experiments. Preliminary studies were presented by Tseng et al. (2001), who assessed gene-specific dye bias by looking at correlations amongst dye swap technical replicates, by Kerr et al. (2002), who analyzed a dataset of two samples under an assumption of common error variance across all genes, and by Dobbin et al. (2003b), who analyzed a small dataset gene-by-gene, with separate analysis for each gene that did not assume common error structure across genes. These studies suggested the existence of dye bias but were too small to be conclusive, and did not attempt to fully characterize the phenomenon. Dombkowski et al. (2004) found gene-and-sample-specific dye bias for some genes but did not quantify the bias in a meaningful way, nor accurately discuss the implications for statistical design and inference. Rosenzweig et al. (2004) presented results of a small study that were similar to the first two papers mentioned. In contrast to earlier studies, we analyze both cDNA and

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

olignoucleotide platforms, and, also in contrast to earlier studies, in both cases the experiments have adequate degrees of freedom to accurately characterize dye bias via separate models for each gene, as seen in the ANOVA tables in the supplement. In addition, we treat both types of dye bias, which have previously been considered only in isolation, and we discuss in detail the implications that these different types of dye biases have with respect to experimental design and statistical analysis.

Review of common misconceptions about dye bias

There are many common misconceptions about dye bias. For instance

1. Dye bias is an artifact of the normalization and background correction procedures used.
2. Dye bias can only be corrected by dye-swapping every array.
3. Dye bias is so complex that we will never be able to properly understand it or correct for it.
4. Dye bias is highly dependent on the type of statistical analysis or software used.
5. Dye bias has no relevance for single channel arrays, such as Affymetrix arrays.

Regarding 1, while normalization and background correction methods may have some marginal impact on dye bias, there is no evidence that these methods are the cause of dye bias, and the dye bias phenomenon appears persistent across the different methods in common use; regarding 2, as we discussed extensively in previous publications (Dobbin et al., 2003a, 2003b), gene-specific dye bias can be eliminated without dye swapping every array, and as we discussed earlier in this introduction, gene-and-sample-specific dye bias will remain even if every array is dye swapped, so the statement is false;

regarding 3, gene-specific dye bias is conceptually quite simple and methods for dealing with this type of phenomenon were developed in the statistical experimental design literature over a half century ago (Cochran and Cox, 1992); gene-and-sample-specific dye bias is more complex, but still explainable in simple intuitive terms; potential complexity in the mechanistic explanation of dye bias should not be confused with its relatively simple representation in a mathematical model; regarding 4, as with point 1, these issues will have only marginal effects on dye bias when reasonable procedures are used; regarding 5, as described in the Conclusion section below, dye bias does have implications for single-channel arrays.

WHY WE USED GENERALIZED LEAST SQUARES

We first used ordinary least squares to fit an analysis of variance model to the cDNA data on a gene-by-gene basis. However, this appeared inadequate because the model residuals displayed different variances across the different cell lines for many genes, violating the model assumption of homoscedasticity. Also, the p-values associated with the F-test for the gene-and-sample-specific dye bias tended to be close to 1 rather than uniformly distributed, indicating lack of fit of the model. A simulation study in which the estimated heteroscedastic variances were used to generate data under the normal model of Equation (1) provided identical indications of lack of fit, suggesting that the problem was due to differences in within-cell-line variances. Therefore, we re-fit the model using generalized least squares (Carroll and Ruppert, 1982a, 1982b), which allows for different

variances in the different cell lines. This corrected the apparent lack of fit. Data from the oligonucleotide arrays displayed similar characteristics and were analyzed the same way. All analyses were performed in R version 1.9.0 using the generalized least squares function.

INTENSITY-DEPENDENT DYE BIAS AND AUTOFLUORESCENCE

While the focus of this paper is the statistical characterization of dye bias and its ramifications for statistical inference, it is natural to speculate as to the causes of dye bias – particularly in an indirect labeled experiment such as this one. A possible partial explanation of the observed relation between median intensity and gene-specific dye bias is that nucleic acid other than labeled sample fluoresces naturally and is skewed towards the wavelength similar to that used for Cy3 (532nm) (Eisinger and Shulman, 1968; Onidas et al, 2002). That is, in addition to fluorescence attributable to the label tags, fluorescence attributable to the DNA itself, on the array and hybridized to the array, will be present and will fluoresce brighter at the shorter Cy3 wavelength than the longer Cy5 wavelength. This phenomenon is sometimes called autofluorescence. If there is a large amount of DNA binding to a spot, then some of this material may naturally fluoresce more intensely under the Cy3/green wavelength, resulting in unwanted enhancement of the Cy3/green channel reading (see, for example, Raghavachari et al., 2003, Figure 1). Because array normalization tends to center intensities around zero, autofluorescence is expected to produce spot-specific dye bias toward the green channel for spots with high concentrations of labeled DNA, and towards the red channel for spots with low concentrations of labeled DNA. Since gene-specific dye bias is a result of the cumulative

dye bias across all spots for a gene, the spot-specific dye bias will induce gene-specific dye bias towards the Cy3/green channel for high prevalence genes, which is the effect apparent in Figure 4. But other aspects of the experiment, such as spot size, printing order, and slide coating, can also play a role in determining the extent of this effect (Raghavachari et al., 2003; Mary-Huard et al., 2004).

ADDITIONAL CONCLUSIONS

It may appear counterintuitive that such a high proportion of genes exhibit statistically significant gene-specific dye bias, yet there was only disagreement on 7% of the genes in Table 2 – that is, when no correction for dye bias was made it affected statistical inference on only 7% of the genes. One partial explanation for this discrepancy is that, as noted earlier in the discussion of Table 2, each cell line is tagged nearly half the time with each dye, which means that even in the presence of gene-specific dye bias, the comparisons between the cell lines are nearly unbiased. In a less balanced design, such as Figure 1(b), the dye bias may be expected to have a larger effect on inference.

Although the data analyses of this paper only included indirect, amino-allyl labeled microarrays, the conclusions for statistical design and analysis should apply also to direct labeled microarrays. First, previous studies have found that there is statistically significant gene-specific dye bias present in direct labeled experiments for many genes, but that it tends to be relatively small, which is very similar to the nature of dye bias we have found in this paper for indirect labeled experiments. Second, although gene-and-

sample specific dye bias has not to our knowledge been previously studied in depth in direct labeled experiments, it is still the case that, if it exists, there is no valid statistical way to remove it. Hence, the existence of gene-and-sample-specific dye bias in direct labeled experiments might be of concern, but it does not have implications for experimental design different than those we have given here.

ADDITIONAL TABLES

Analysis of variance		
Source	Oligonucleotide Arrays DF	cDNA Arrays DF
Cell Line	5	5
Orientation	1	1
Cell Line by Orientation Interaction	5	5
Error	18	16
Total	29	27

Table 1: Analysis of variance degrees of freedom (DF) for oligonucleotide and cDNA microarray experiments.

	Total estimated cell line effects	Number with discrepancy larger than 1	Number with discrepancy larger than 2
MCF10a	8604	20 (0.2%)	0
LNCAP	8604	4 (0.05%)	0
L428	8604	42 (0.5%)	4 (0.05%)
SUDHL	8604	27 (0.3%)	0
OCILY3	8604	50 (0.6%)	5 (0.06%)
Jurkat	8604	22 (0.3%)	0

Table 2: cDNA experiment: Number of large discrepancies between effect size estimates based on the backward-only arrays compared to based on the forward-only arrays.

		All data: no dye effects	
		P-value < .001	P-value > .001
All data: dye effects	P-value < .001	10618 (67%)	925 (6%)
	P-value > .001	185 (1%)	4062 (26%)

Table 3: Oligonucleotide arrays. Observed agreement between models with and without gene-specific dye effects included.

		Loess 1	
		P-value<.001	P-value>.001
Loess 2	P-value <.001	2225 (26%)	135 (2%)
	P-value >.001	170 (2%)	6074 (71%)

		Loess 1	
		P-value <.001	P-value >.001
Loess 3	P-value <.001	2015 (23%)	416 (5%)
	P-value >.001	380 (4%)	5793 (67%)

		Loess 2	
		P-value <.001	P-value >.001
Loess 3	P-value <.001	1961 (23%)	470 (5%)
	P-value >.001	399 (5%)	5774 (67%)

Table 4: Comparison of loess gene-specific dye bias p-values under various loess normalization techniques. Loess 1 indicates a loess smooth in R version 1.9.0 with default span parameter setting 0.75. Loess 2 indicates a span parameter of 0.4. Loess 3 indicates a span parameter of 3.0. Note that there is ~90% or greater agreement in p-values across parameter settings, and roughly equal numbers of dye bias genes estimated under each parameter setting, indicating that the overall extent of dye bias is not affected by the span parameter.

Gene-specific dye bias P-values				
		Oligonucleotide array p-values		
		<.001	>.001	Total
cDNA array p-values	<.001	799	1539	2338
	>.001	1067	2651	3718
		1866	4190	6056
Cell line P-values				
		Oligonucleotide arrays		
		<.001	>.001	Total
cDNA array p-values	<.001	3163	518	3681
	>.001	1376	999	2375
		4539	1517	6056

Table 5: Tables used to generate kappa statistics in Table 6 of paper. P-values based on generalize F-test analysis of any dye bias effect, or any cell line effects, for each of the 6056 genes matched across platforms. Dye bias tests adjust for cell lines.

ADDITIONAL FIGURES

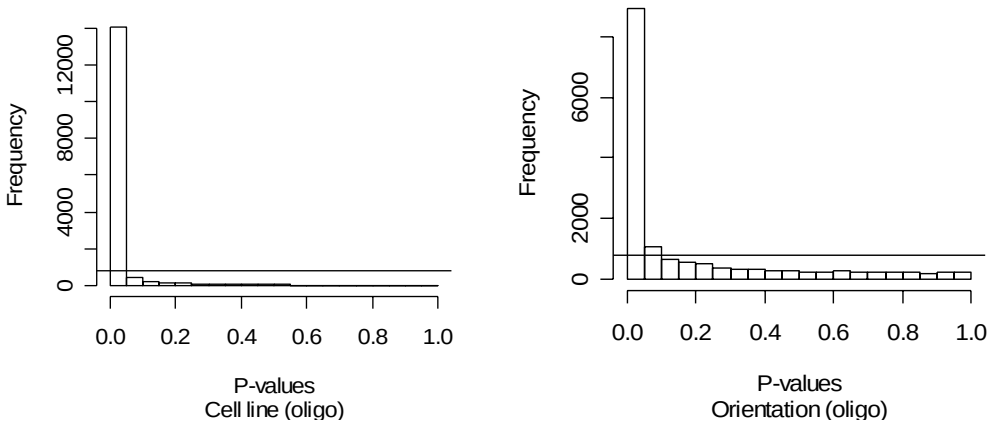

Figure 1: Oligonucleotide p-values. P-values for the F-tests of any effects due to (a) cell lines, (b) gene-specific dye bias.

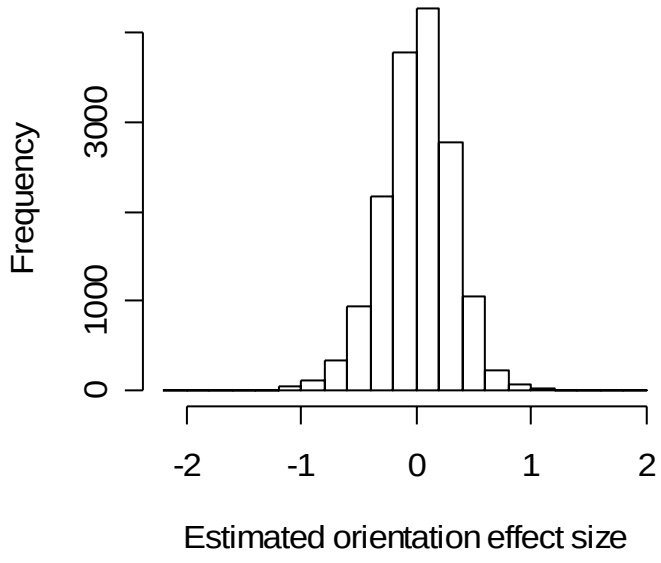

Figure 2: Oligonucleotide experiment estimated effect sizes of gene-specific dye bias for 15,790 genes.

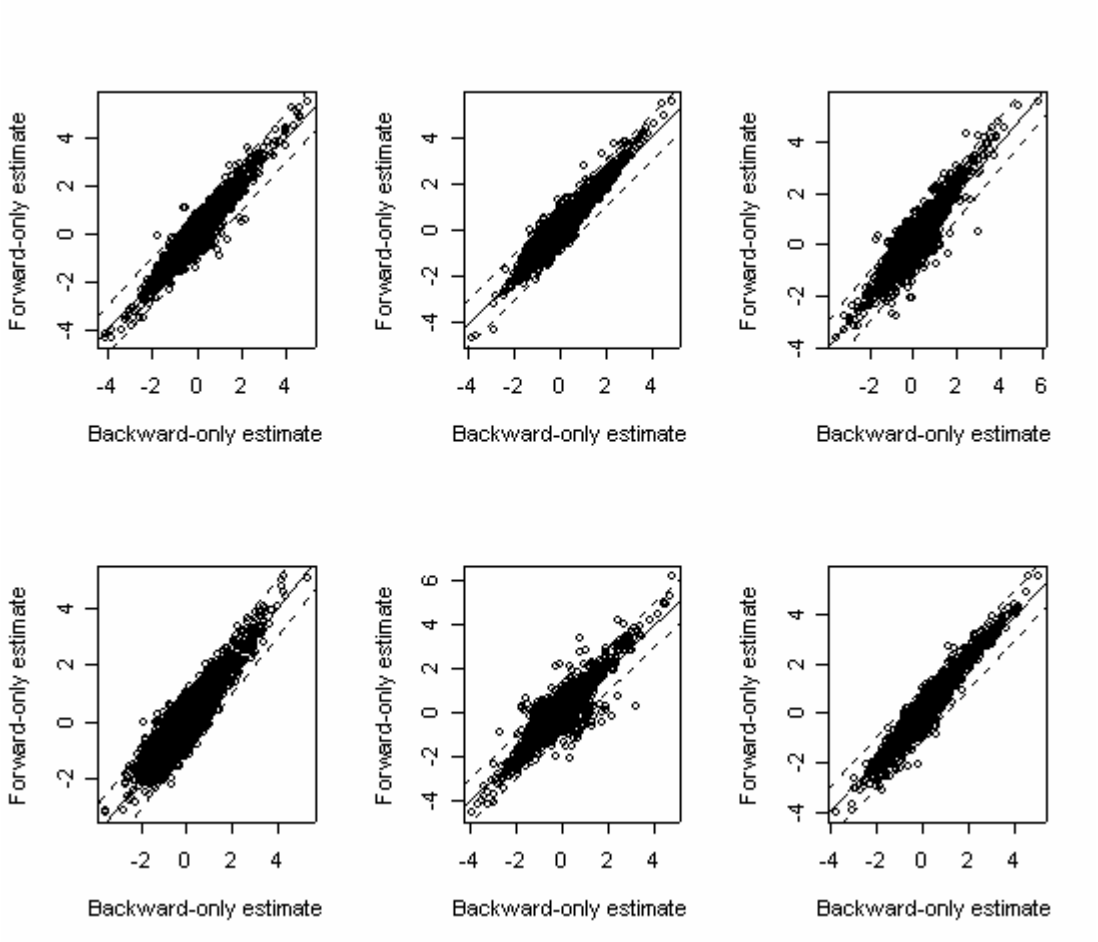

Figure 3: Comparison of cell line effect estimates from cDNA experiment using only the forward arrays versus using only the backward arrays. Drawn on the plots are a 45 degree line through the origin and lines +/- 1 above and below this line. Top row is MCF10a, LNCAP and L428 (left to right), and bottom row is SUDHL, OCILY3 and Jurkat. Correlations are .91, .92, .87, .84, .86 and .94 respectively.