11-28-2005

A New Approach to Intensity-Dependent Normalization of Two-Channel Microarrays

Alan R. Dabney
University of Washington, adabney@u.washington.edu

John D. Storey
University of Washington, jstorey@u.washington.edu

A New Approach to Intensity-Dependent Normalization of Two-Channel Microarrays

Alan R. Dabney and John D. Storey
Department of Biostatistics
University of Washington, Seattle, WA 98195
adabney@u.washington.edu, jstorey@u.washington.edu

Abstract

A two-channel microarray measures the relative expression levels of thousands of genes from a pair of biological samples. In order to reliably compare gene expression levels between and within arrays, it is necessary to remove systematic errors that distort the biological signal of interest. The standard for accomplishing this is smoothing "MA-plots" to remove intensity-dependent dye bias and array-specific effects. However, MA methods require strong assumptions. We review these assumptions and derive several practical scenarios in which they fail. The "dye-swap" normalization method has been much less frequently used because it requires two arrays per pair of samples. We show that a dye-swap is accurate under general assumptions, even under intensity-dependent dye bias, and that a dye-swap provides the *minimal* information required for removing dye bias from a pair of samples in general. Based on a flexible model of the relationship between mRNA amount and single channel fluorescence intensity, we demonstrate the general applicability of a dye-swap approach. We then propose a common array dye-swap (CADS) method for the normalization of two-channel microarrays. We show that CADS removes both dye-bias and array-specific effects, and preserves the true differential expression signal for every gene. Finally, we discuss some possible extensions of CADS that circumvent the need to use two arrays per pair of samples.

1 Introduction

A two-channel microarray simultaneously measures the expression levels of thousands of genes from a pair of biological samples (Schena et al. 1995, Hughes et al. 2001). This system has existed in the form of spotted cDNA arrays since the beginning of high-throughput gene expression technology (Schena et al. 1995). With the advent of oligo-based two-channel microarrays using new ink-jet printing technologies (e.g., Agilent Laboratories in Palo Alto, California), the two-channel platform continues to be of much interest.

In a two-channel microarray, each sample is labeled with a particular dye, Cy3 (green) or Cy5 (red), and both are competitively hybridized to the same array. The relative dye intensity for each gene is used to quantify expression. Due to various sources of systemic effects, the relative expression levels need to be adjusted to accurately reflect the underlying amount of mRNA from each sample. For example, the two dyes incorporate into samples at different rates, creating "dye bias" (Tseng et al. 2001, Yang et al. 2001, Yang et al. 2002b, Yang et al. 2002a, Yue et al. 2001, Dobbin et al. 2005). In addition to dye bias, there may be systematic differences between arrays (Tseng et al. 2001, Yang et al. 2002b). It is therefore necessary to remove dye bias and any other systematic effects from two-channel microarray measurements. This process is often referred to as "normalization." However, the general goal is not necessarily to make expression measurements conform to a standard (the usual meaning of normalization), but rather to preserve the biological signal of interest so that accurate statistical comparisons between and within arrays can be made.

To address this problem, we investigate the dye-swap design, showing that it provides the *minimal* information necessary to remove dye bias in general from a pair of samples, an observation related to others previously made (Kerr et al. 2000, Kerr & Churchill 2001, Fang et al. 2003). We show that while the original justification of the dye-swap (Kerr et al. 2000) assumed a constant dye effect, the dye-swap is accurate more generally, even in the case of intensity-dependent dye bias. We develop a model that treats dye and array effects as flexible, nonlinear functions of the amount of mRNA present in each sample (i.e., sample-specific intensity). Within this framework, we show that a simple dye-swap average removes dye bias without affecting signal and preserves the ordering of true expression means. We then propose the *common array dye-swap* (CADS) method, a two-stage procedure for removing dye-bias and array-specific effects. The first step is the usual dye-swap average. The second step centers each array around a "common array", i.e., an estimated model of the relationship between mRNA amount and fluorescence intensity when all array effects have been removed.

The most important goal in preprocessing microarray data is to maintain the true type of differential expression within and between arrays. We show that CADS preserves differential expression relationships in general; that is, in terms of the population quantities that are tested in a statistical analysis, CADS preserves whether there is no differential expression, over-expression, or under-expression among the *true* mRNA levels. Furthermore, CADS achieves this accuracy without requiring any of the strict assumptions of "MA-plot" (Tseng et al. 2001, Yang et al. 2002b) or global (Chen et al. 1997) normalization methods. We highlight the general applicability of CADS through several simple examples in which the assumptions behind MA-plot and global normalization methods are violated. Importantly, in each of these examples, it is shown that MA-plot and global methods may cause every null, non-differentially expressed gene to become spuriously differentially expressed. At the same time, the truly differentially expressed genes may have the degree and direction of their differential expression altered. Furthermore, without performing a dye-swap, it is not possible to verify that the assumptions behind MA-plot and global methods hold.

2 Currently Available Normalization Methods

Many normalization methods have previously been proposed (Quackenbush 2002, Bilban et al. 2002). Global normalization methods generally subtract some constant from all log expression ratios on an array (Kerr et al. 2000, Kroll & Wolfl 2002, Kepler et al. 2002). These are rarely used since systematic effects tend to be more complicated than constant shifts at the array level. Others have proposed procedures based on observed deviations in "housekeeping" genes (genes whose expressions levels are known to remain constant) (Chen et al. 1997). While true housekeeping genes do not appear to exist (Suzuki et al. 2000), the ideas have inspired other methods that attempt to either create exogenous housekeeping genes (Yang et al. 2002b, Benes & Muckenthaler 2003) or identify and/or synthesize genes on an existing array that behave like housekeeping genes (Zien et al. 2001, Tseng et al. 2001). The former approach is dependent on one's ability to identify a set of control genes for each individual experiment. The latter approach necessarily involves a subjective choice of the rule for calling a gene unchanged. The rule cannot be verified and, if incorrect, could result in more systematic bias after its application than before.

The vast majority of normalization methods used today build off of three highly-influential papers (Tseng et al. 2001, Yang et al. 2001, Yang et al. 2002b). There, "MA-plots" were used to demonstrate that dye bias depends on intensity in an array-specific manner. An MA-plot is a scatterplot with log intensities (log of red times green, divided by two) on the x-axis and log ratios (log of red divided by green) on the y-axis. Figure 1 shows example "self-self" MA-plots from calibration experiments (Tseng et al. 2001), where the pair of samples hybridized to each array are biologically equivalent. Since the intensities of the two dyes should be equal, the trends away from the zero line are evidence of dye bias. We refer to any normalization method that is based on removing trends from MA-plots as "MA methods." MA methods fit a smooth curve to the MA-plot for each array, and then subtract this curve from the log ratios, reshaping the MA-plot to be centered on the horizontal zero line. The smoother used is often the robust locally-linear loess routine (Cleveland 1979). To be precise, such loess-based MA methods can be viewed as special cases of the general MA method. Many other methods have been developed as extensions of this general idea (Benes & Muckenthaler 2003, Wilson et al. 2003, Fang et al. 2003, Fan et al. 2004, Xiao et al. 2005).

3 The Assumptions Behind MA Methods

An MA-plot is a scatterplot of the relative gene expression levels of two samples versus their average intensities. For convenience, we call one sample the "target" and the other the "control" throughout the entire article, although the ideas presented here apply to any two varieties of samples. Suppose we have labeled targets with red dye and controls with green dye on a single array. Let Y_{Ti} and Y_{Ci} be the $\log_2$ intensities for gene i, $i = 1, 2, \ldots, m$, in targets and controls, respectively. An MA-plot is a scatterplot with the differences in expression $M_i = Y_{Ti} - Y_{Ci}$ on the y-axis and the average abundances $A_i = (Y_{Ti} + Y_{Ci})/2$ on the x-axis (Yang et al. 2002b). Equivalently, an MA-plot is a scatterplot with the Y_{Ti} on the y-axis and the Y_{Ci} on the x-axis that has been rotated clockwise by 45° and rescaled.

*** Figure 1 about here. ***

Two examples of MA-plots are shown in the top row of Figure 1, based on data from a set of calibration experiments (Tseng et al. 2001). A sample of *E. coli* was partitioned into aliquots of equal size, with one labeled with red dye and the other labeled with green dye. Two arrays were then formed, each comparing the expression intensity under the two dyes. Because these are "self-self" comparisons, there is no differential expression, and all expression differences should be centered at zero (the dashed lines) in the MA-plots. Instead, we see a systematic deviation from zero; this is evidence of dye bias. Similar conclusions have been drawn from other calibration experiments (Yue et al. 2001, Yang et al. 2002b, Yang et al. 2002a).

MA methods attempt to remove bias by smoothing away MA trends. A smooth curve (often loess (Cleveland 1979)) is fit through the MA-plot, and its predicted values are subtracted off, recentering the plot along the zero line. The smoother curve can be estimated using all of the data or some subset of control genes. Regardless, the goal is to reshape an MA curve to lie on the zero line. While Figure 1 gives empirical support to this idea, no theoretical justification has been given for its general use. Note that in "self-self" experiments, there is no biological signal of interest present because each gene is equivalently expressed. As it turns out, true biological differences in expression between the two samples can also produce trends in the MA-plots. Therefore, regressing away trends in MA-plots may change the biological signal of interest. Through a rigorous mathematical analysis of this approach, in addition to empirical evidence, it appears that this procedure can create artificial differential expression, destroy true differential expression, or both.

MA methods have been justified when "there are good reasons to expect that (i) only a relatively small proportion of the genes will vary significantly in expression between the two cohybridized mRNA samples, or (ii) there is symmetry in the expression levels of the up/down-regulated genes" (Yang et al. 2002b). In fact, it can be shown that the fundamental assumption behind MA methods is that all systematic MA trends are due exclusively to bias (*Supporting Appendix*). Equivalently, MA methods assume there is no biologically-driven trend to an MA plot. This assumption in turn implies three additional assumptions (also proved in the *Supporting Appendix*) that have non-trivial consequences.

1. Overall, differential expression between the two samples is symmetric about zero.
2. There is no relationship between differential expression and expression abundance.
3. The variation of expression across genes is the same in each sample.

While the symmetry assumption was discussed in the original MA papers, the other two assumptions do not appear to have been acknowledged. As the original authors stated, MA methods are only appropriate when their assumptions are met. In order to make these assumptions concrete, we give three plausible scenarios in which they do not hold, shown in Figure 2. Applying MA methods when their assumptions are violated can create or destroy differential expression signal. The key observation here is that assumptions must be made about the true *biological signal* before applying MA methods. The motivating datasets for the MA-plot methods were specialized in that there was little to no differential expression. In such a case, the assumptions for MA methods may be met. However, since "the expression profiles in biological samples are (frequently) more divergent in nature than in the examples investigated (here), . . . a control sample that spans the intensity range and exhibits a relatively constant expression level across biological samples is desirable" (Yang et al. 2002b). This observation was the motivation behind the physically constructed *microarray sample pool* (MSP) (Yang et al. 2001) and the rank-invariant MA method (Tseng et al. 2001).

*** Figure 2 about here. ***

Simulation 1: Asymmetric Differential Expression. Suppose that two groups are compared where there is more differential expression in one direction than in the other. For example, targets may be overexpressed more often than they are underexpressed relative to the controls. In this case, target measurements will tend to be greater than their control counterparts, and an MA-plot will be centered above the zero line. The first plot of Figure 2 shows a simulated example (see *Supporting Appendix* for simulation details of all examples). There is no dye bias, and smoothing the MA curve pushes the cloud of data down to the zero line. Overexpressed genes will appear less extreme, under-expressed genes will appear more extreme, and *all* equivalently expressed genes will now be artificially differentially expressed. Thus, smoothing the MA-plot destroys signal in some genes and creates signal in others. Note that asymmetry is not uncommon (Tusher et al. 2001, Hedenfalk et al. 2001); in fact, only in specialized circumstances would it be guaranteed that differential expression between two biological groups of interest be perfectly symmetric.

Simulation 2: Intensity-Dependent Differential Expression. Suppose now that differential expression is related to expression abundance; for example, at low abundance genes differential expression may tend to be in the direction of controls, and at high abundance genes in the direction of targets. Differential expression is often related to the level of activity and importance of a gene in a particular cell type. For example, comparing cancer tumor to healthy cells, the genes that are related to the increased replication rate will be most abundant and most different in expression. If nothing else, differential expression cannot occur among genes that are not active in a cell, which forms a relationship between differential expression and abundance. This scenario produces an MA trend like that in the second plot of Figure 2 even though there is again no dye bias or other systematic effects included in our simulation. Smoothing the MA-plot will again destroy signal in some genes and create signal in others. In particular, it will again create signal in all of the null, equivalently expressed genes wherever the MA smoother does not pass through the zero line.

Simulation 3: Unequal Variation in Expression Means. Unequal variances between comparison groups occur, for example, when comparing immune challenged to normal cells (Storey et al. 2005). In particular, resources in the cell are reallocated so that a large proportion of genes have decreased expression, while immune response genes have increased expression, creating a difference in the variation of expression across genes between the two immune challenged and normal cells. As another example, due to the amount of genetic mutation and cell cycle activities associated with the cancer process, cancer expression measurements would easily have much more variation than control measurements when comparing these two types of cells. The third plot in Figure 2 shows a simulated example where this property holds, and no dye bias or systematic effects have been included. There is again a substantial MA trend due to this true biological signal. Smoothing the MA-plot destroys the biological signal.

*** Figure 3 about here. ***

Real Single-Channel Examples. Figure 3 shows two real single-channel microarray experiments that display the properties used in the three examples; we include the single-channel examples because we know any MA trends are not due to dye bias. In one experiment, arrays were obtained from sporadic, BRCA1 mutation-positive, and BRCA2 mutation-positive breast cancer tumors (Hedenfalk et al. 2001). Figure 3a shows the $\log_2$ fold-change among 3220 genes between BRCA1 and BRCA2 arrays, where there is clear asymmetric differential expression as in Example 1. Figure 3c shows there is a relationship between overall expression abundance and differential expression, as in Example 2. Figure 3d shows boxplots of the expression data among the sporadic, BRCA1 and BRCA2 groups. It can be seen there that the variation of expression across genes increases from sporadic to BRCA1 to BRCA2, as in Example 3. The increase in standard deviation across genes is about 11% moving from one group to the next. In the other experiment, four human volunteers were treated with endotoxin and four with a placebo, then arrays were obtained on blood samples from each over a 24 hour time course (Storey et al. 2005). Figure 3b shows the the $\log_2$ fold-change at hour four among over 44,000 genes between the two groups, where there is again clear asymmetric differential expression, as in Example 1. The data for each gene in all of these plots are an average of the $\log_2$ expression among all individuals within a group, which helps to remove some of the variability and make the evidence more compelling than showing data from a single array.

*** Figure 3 about here. ***

A Real Two-Channel Example. A third example comes from a prostate development study involving mice, with the data kindly provided by the Peter S. Nelson laboratory at the Fred Hutchinson Cancer Research Center. Here, six two-channel microarrays were formed, each comparing the prostate of a separate 30-day-old mouse to 14-day post-conception embryonic controls. On half of the arrays, targets were labeled red and controls green, while on the other half the dye configuration was swapped. Note that this experimental design differs from the conventional dye-swap, since biological, rather than technical, replicates were used for the swapping.

*** Figure 4 about here. ***

Figure 4 shows MA plots for the six arrays. The top row shows the arrays with targets red and controls green, while the bottom row shows the dye-swapped arrays. The solid curves are smoothers fit to each array individually and would be the targets of MA-normalization. The dotted curves are smoothers fit to all data with the same dye configuration. Thus, there is a single curve in each plot in the top row, representing average differences between targets labeled red and controls labeled green. Similarly, there is a single curve in the bottom row, representing average differences between targets labeled green and controls labeled red. Intuitively, if only dye-bias were behind the observed trend away from the zero line, then these two curves should be symmetric reflections of one another. In fact, the two curves are far from symmetric, as is evident in Figure 5.

*** Figure 5 about here. ***

Here, each array has been recentered around the *average* of the two dotted curves. Thus, there is intensity-dependent differential expression in this example. MA-normalization would destroy this systematic signal among truly differentially-expressed genes and add signal elsewhere. This example suggests that, while reshaping of MA-plots can be justified, a flat curve following the horizontal zero line is not necessarily the “correct” shape.

Global Normalization of Microarrays. Global normalization of microarrays refers to the process of adjusting every M_i by a simple linear transformation. As a simple example, the mean expression value of the array may be subtracted from each M_i, and then the subsequent values may be divided by the standard deviation of the array. In terms of MA-plots, this is equivalent to fitting a flat line through the data rather than a flexible smoother, then reshaping. This procedure is susceptible to the case where there is asymmetric differential expression. Global normalization can then destroy signal for truly differentially expressed genes and create signal for all genes that are not truly differentially expressed. As we see below, global methods also suffer from the fact that unwanted systematic effects can depend on the underlying abundance of mRNA.

MA Methods in General. When the assumptions behind MA methods are not expected to be valid, some sort of control must be used. The *microarray sample pool* (MSP) is designed for this purpose (Yang et al. 2002b), although the MSP does not appear to have been used much in practice. The *rank-invariant* MA method (Tseng et al. 2001) is similarly motivated, attempting to use only genes whose expression is equal between comparison groups to form the MA smoother curve. Genes are included in the smoother fit if their ranks in the targets are approximately equal to their ranks in the controls. This assumes that equality in rank corresponds to equality in expression. In Examples 1 and 2, however, the distribution of target mRNA is shifted relative to the distribution of control mRNA. As a result, the rank-invariant method will include differentially expressed genes in the smoother, producing the same negative effect as above.

4 A Generalization of the Problem

Due to the extensive assumptions required for MA and global normalization methods, we seek a more generally applicable method. Although it has been shown that some systematic error can be captured in an MA-plot (Tseng et al. 2001), this does not mean that any relationship between the M and A quantities is due to unwanted systematic effects. Furthermore, systematic bias can be modeled more generally in terms of the underlying amount of mRNA present in each sample for each gene, rather than the A variable. Therefore, the models we develop for normalization are written as functions of the true amount of mRNA in each biological sample. The goal is to use these models to adjust the observed fluorescence intensities so that they accurately reflect the underlying levels of mRNA. We take into account the fact that relationships between mRNA amount and fluorescence intensity can be due to bias or true biological signal.

Dye-swap required for removing dye bias in general. As a first step towards this end, it is necessary to determine the type of observations that are needed to unequivocally parse the different sources of signal. For each gene, there is the possibility for at least two signals to affect the relative expression values. The first is actual biological signal, and the second is dye bias. (There is also the possibility of an array-specific effect, but we delay considering this for the moment.) With only a single observation of the M variable for each gene, it is impossible to separate these two signals. MA normalization tries to get around this by assuming a particular relationship between the M and A variables, freeing up a degree of freedom to estimate the biological signal. However, as we have seen, this assumed relationship is not necessarily true in general.

We have been able to conclusively show that a dye-swap provides the minimal information required for separating dye bias from biological signal from a single pair of samples (*Supporting Appendix*). Consider the observed expression intensities for a single gene. Ignoring random measurement error and assuming that any bias is due exclusively to the dyes used, there are four unknown quantities involved in these two observations: the true target and control mRNA amounts and two functions which translate the mRNA amounts into measurements on the red and green dyes. In other words, we observe two unknown functions evaluated at two unknown arguments, making it impossible to make any reliable statement about the relative relationship between the two samples. On the other hand, by adding just two more observations, we can compare target to control as measured by a common function. Specifically, we must also label target green and control red, performing a dye-swap. Averaging the two target observations and two control observations separately, we have target and control measurements labeled with an "average dye." The trick is that the mRNA amount for each respective sample remains constant across both dye assignments.

Assuming that each dye (with all sources of variation and systematic bias removed) provides a useful measure of expression for a given underlying mRNA amount, the “average dye” should work equally well.

Models of fluorescence intensity in terms of mRNA amount. As a second step, we formulate models of fluorescence intensity in terms of the true underlying mRNA amounts present in each sample. The calibration experiments summarized in Figure 1 provide some information for building such a model. As we have already seen, the top row of Figure 1 shows behavior of dye bias. The second row of plots in Figure 1 compares the *E. coli* samples labeled with the same dyes on different arrays and thus tells us about differences between arrays.

We first introduce a model only incorporating dye bias effects. Let Y_{TRij} denote an observed fluorescence intensity on the $\log_2$ scale for gene i on target sample j, using the red dye. Similarly, Y_{CGij} corresponds to fluorescence intensity obtained from the control sample with green dye (on the same array as Y_{TRij}), and Y_{TGij}, Y_{CRij} are the dye-swap analogs. Let X_{Tij} and X_{Cij} be the true mRNA amounts for gene i in target and control samples j (say, in units that count the number of mRNA molecules). Ideally, we would observe X_{Tij} and X_{Cij} directly. However, variation is introduced into these quantities during the sample preparation, hybridization, and scanning stages. The resulting quantities after all of this variation has been introduced are the observed fluorescence intensities.

We assume that the observed intensities Y_{ij} are noisy versions of a true underlying dye-specific fluorescence function (d_G or d_R) applied to the X_{ij}:

$$\begin{aligned} Y_{TRij} &= d_R(X_{Tij}) + \epsilon_{TRij} \\ Y_{CGij} &= d_G(X_{Cij}) + \epsilon_{CGij} \\ Y_{TGij} &= d_G(X_{Tij}) + \epsilon_{TGij} \\ Y_{CRij} &= d_R(X_{Cij}) + \epsilon_{CRij}, \end{aligned} \tag{1}$$

for $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, where the ϵ_{ij} are mean-zero error terms. Note that each function, d_G and d_R, describes the fluorescence intensity that is observed for a given mRNA amount with all remaining sources of variation removed. We have allowed for the dye functions to be different, which allows dye bias to be taken into account. We do not assume any particular form of d_G and

d_R, except that (i) they are smooth functions and (ii) they are increasing functions. If these two properties do not hold, then surely the technology itself is problematic and cannot be fixed through normalization.

There are distinct nonlinear trends to both plots in the second row of Figure 1, which represent array-specific effects. Let a_1 and a_2 be the array functions on the first and second (dye-swap) arrays, respectively. The above model can be updated to include these effects:

$$\begin{aligned}
Y_{TRij} &= d_R(X_{Tij}) + a_{1j}(X_{Tij}) + \epsilon_{TRij} \\
Y_{CGij} &= d_G(X_{Cij}) + a_{1j}(X_{Cij}) + \epsilon_{CGij} \\
Y_{TGij} &= d_G(X_{Tij}) + a_{2j}(X_{Tij}) + \epsilon_{TGij} \\
Y_{CRij} &= d_R(X_{Cij}) + a_{2j}(X_{Cij}) + \epsilon_{CRij},
\end{aligned} \tag{2}$$

$i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$. Again, these array effects are written as functions of the true mRNA amounts present in each sample. Thus, for each sample pair j, the four observed fluorescence intensities Y_{TRij}, Y_{CGij}, Y_{TGij}, and Y_{CRij}, $i = 1, 2, \ldots, m$, are modeled using four functions d_R, d_G, a_{1j}, and a_{2j}. Note that there may be other influences at play, such as array batch effects or spatial effects, that could be included in the model. We discuss this in the *Supporting Appendix*, but assume for now that only dye effects and array effects are present.

Because we model the underlying mRNA amounts for each individual separately, the ϵ_{ij} do not include biological variability and hence are comparable across genes. The X_{ij} are random variables in that they represent biological variability in gene expression from sample to sample. Let μ_{Ti} and μ_{Ci} be the respective target and control mean RNA amounts for gene i in the population where all sources of biological variability have been removed. We assume that $d(X_{ij}) = d(\mu_i) + \gamma_{ij}$ for both dye functions and $a_j(X_{ij}) = a_j(\mu_i) + \alpha_{ij}$ for all array functions, where the γ_{ij} and α_{ij} have expected value zero. This allows us to connect the above models across different pairs of samples.

5 Proposed Methods

The fact that we do not know the dye functions, array functions, or the true underlying mRNA amounts makes it difficult to produce expression measures that do not include systematic biases. However, we now propose a method based on the above models that accomplishes exactly this in expectation under the requirement that dye-swap arrays have been obtained for each pair of samples.

Simple Dye-Swap Average. Assuming that the array effects can eventually be removed, there is still the problem of removing dye bias. Another way to think about this is to replace d_G and d_R with a single function that does not depend on which dye was used. There is no "correct" dye function; in fact, one would hope that d_G and d_R each produce reasonable measures of gene expression.

A simple way to remove the dye-specific effect would be to use the average of these two functions. A dye-swap does just that. The dye-swap average intensities can be written in terms of the models given in equation (1):

$$\begin{aligned} \tilde{Y}_{Tij} &= \frac{1}{2}\Big[Y_{TRij} + Y_{TGij}\Big] = \frac{1}{2}\Big[d_R(X_{Tij}) + d_G(X_{Tij})\Big] + \epsilon_{Tij} = d(X_{Tij}) + \epsilon_{Tij} \\ \tilde{Y}_{Cij} &= \frac{1}{2}\Big[Y_{CRij} + Y_{CGij}\Big] = \frac{1}{2}\Big[d_R(X_{Cij}) + d_G(X_{Cij})\Big] + \epsilon_{Cij} = d(X_{Cij}) + \epsilon_{Cij}, \end{aligned} \tag{3}$$

$i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, where d is the function $[(d_R + d_G)/2]$, and the ϵ_{ij} are the sums of the corresponding error terms in (1) divided by two. The $\tilde{Y}_{ij}$ are the data resulting from a simple dye-swap average. Thus, the average of the red and green channels from separate arrays can be seen as a single function applied to the separate mean expression levels. We refer to d as the *average dye* function. Since we assume that d_R and d_G are increasing, d is also increasing. Therefore, the true ordering of the X_{ij} will be preserved, in expectation, by simple averaging of dye-swap arrays.

Note that the dye-swap method has previously only been justified by assuming that any dye effect is constant (i.e., is not intensity dependent). However, it can be seen from the above calculation that the dye-swap does not make this assumption. In fact, it seems to offer the most general assumption about dye bias. By its very design, the underlying mRNA amounts are kept constant in each dye configuration, thereby automatically incorporating any intensity specific dye bias. This is why the average dye functions falls out so straightforwardly in the models of the dye-swap averaged expression measurements in equation (3).

Common Array Dye-Swap (CADS). Array effects represent the variability from array to array that is not due to biological sources of variation. It is possible that the array effects could be a function of the underlying mRNA abundance or the position of the probe. (The latter effect can be removed by applying a a regional smoother to the image file (Cui et al. 2003).) It is important to remove random array effects because they increase the variability and uncertainty of the measurements, and they can create dependence between genes that has adverse effects on significance calculations (Qiu et al. 2005). Including the array functions in (3), we have:

$$\begin{aligned} \tilde{Y}_{Tij} &= d(X_{Tij}) + a_j(X_{Tij}) + \epsilon_{Tij}, \\ \tilde{Y}_{Cij} &= d(X_{Cij}) + a_j(X_{Cij}) + \epsilon_{Cij}, \end{aligned} \tag{4}$$

$i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, where a_j is the function $[(a_{1j} + a_{2j})/2]$. Thus, the dye-swap simple average produces normalized data that still have array-specific features.

We propose a modification of the dye-swap simple average that removes array effects. This approach can be motivated under the assumption that each array function comes from a random distribution of curves that have expected value equal to the zero line. We show, however, that the method is equally effective under more general assumptions, which basically require that the array functions do not confound the relevant signal from the dye functions. If we form n dye-swap array pairs and subtract the dye-swap averages from each other, we have

$$\begin{aligned} \tilde{M}_{ij} &= \tilde{Y}_{Tij} - \tilde{Y}_{Cij} \\ &= d(X_{Tij}) - d(X_{Cij}) + a_j(X_{Tij}) - a_j(X_{Cij}) + \epsilon_{Tij} - \epsilon_{Cij}, \end{aligned}$$

$i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. Recall our assumption that $d(X_{ij}) = d(\mu_i) + \gamma_{ij}$, where γ_{ij} has mean zero. Then, averaging the $\tilde{Y}_{ij}$ and taking expectations with respect to arrays gives

$$\begin{aligned} \mathrm{E}\left[\frac{1}{n}\sum_{j=1}^{n}\tilde{M}_{ij}\right] &= \mathrm{E}\left\{\frac{1}{n}\sum_{j=1}^{n}\Big[d(X_{Tij}) - d(X_{Cij}) + a_j(X_{Tij}) - a_j(X_{Cij}) + \epsilon_{Tij} - \epsilon_{Cij}\Big]\right\} \\ &= d(\mu_{Ti}) - d(\mu_{Ci}), \end{aligned}$$

$i = 1, 2, \ldots, m$. This means that array effects should "average out" over n arrays; since the same dye functions are involved on each array, we can form an unbiased estimate of the dye functions by averaging over all arrays. We can then compare each array's profile to the estimated dye functions and interpret systematic differences between the two as array effects. Subtracting off these differences estimates the scenario where the target and control were labeled with the same dye and all comparisons were made on a common array.

We refer to the method as *Common Array Dye-Swap* (CADS). CADS can be summarized by the following algorithm:

1. Perform simple averaging of dye-swap arrays to form $\tilde{M}_{ij} = \frac{1}{2}(Y_{TRij} + Y_{TGij}) - \frac{1}{2}(Y_{CRij} + Y_{CGij})$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$. These quantities are random observations of target minus control, having labeled both with the same dye on n different pairs of arrays.

2. Average over the $\tilde{M}_{ij}$ to form $\tilde{\Delta}_i = \frac{1}{n}\sum_{i=1}^{n} \tilde{M}_{ij}$. This is an estimate of the expected difference between target and control under the average dye function. From our assumptions, it follows that $\mathrm{E}[\tilde{\Delta}_i] = d(\mu_{Ti}) - d(\mu_{Ci})$, $i = 1, 2, \ldots, m$.

3. Fit a smoother f (e.g. natural cubic spline (Green & Silverman 1993)) to the scatterplot with the $\tilde{\Delta}_i$ on the x-axis and $\tilde{Y}_{ij} - \tilde{\Delta}_i$ on the y-axis to form $\tilde{a}_{ij} = f(\tilde{\Delta}_i)$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$. Since we interpret any systematic difference between an array profile and the average dye function as an array effect, these are estimates of the array functions a_j, $j = 1, 2, \ldots, n$.

4. Form $\hat{M}_{ij}$ by subtracting off the estimated array functions: $\hat{M}_{ij} = \tilde{M}_{ij} - \tilde{a}_{ij}$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$. These quantities are random observations of target minus control, having labeled both with the same dye, and with all n comparisons being made on an estimated common array.

MA methods assume that every MA-plot should be centered on the zero line. If this were true, then reshaping each array accordingly would remove both dye bias and array effects. However, we have seen that this assumption does not always hold. Due to features like those discussed earlier, there may be a biological mechanism underlying an MA trend. CADS allows for this possibility. Furthermore, due to systematic differences between the arrays used, our method supposes that each array profile is composed of the true target-to-control comparison in addition to array effects. The idea behind CADS is to estimate the true target-to-control difference by averaging out array effects and center each array around the result.

CADS preserves differential expression relationships. CADS produces normalized data that preserve differential expression relationships in expectation. This means that null genes are null, overexpressed genes are overexpressed, and underexpressed genes are underexpressed. Specifically, under very general conditions, we can show that the CADS estimator is unbiased for the parameters of interest $d(\mu_{Ti}) - d(\mu_{Ci})$, $i = 1, 2, \ldots, m$. Since differential expression methods are almost always based on sample averages (Cui & Churchill 2003), it is sufficient to show that the sample average of the CADS-normalized arrays is unbiased. Therefore, we show that

$$\mathrm{E}\left(\frac{1}{n}\sum_{j=1}^{n}\hat{M}_{ij}\right) = d(\mu_{Ti}) - d(\mu_{Ci}),$$

$i = 1, 2, \ldots, m$ (*Supporting Appendix*). The proof follows from the fact that $\sum_{j=1}^{n}\hat{M}_{ij} = \sum_{j=1}^{n}\tilde{M}_{ij}$, $i = 1, 2, \ldots, m$. In fact, it is not necessary for the array functions to have expectation equal to the zero line. All that is required to preserve differential expression relationships is that the array functions average to some nondecreasing function (*Supporting Appendix*). In this case, CADS is not exactly unbiased, but it still preserves the presence and direction of differential expression, which is the relevant property given that the dye function d already distorts the true mRNA count.

Extensions of CADS. CADS can easily be applied when more than two groups are compared. A loop design with more than two nodes is required, but the method remains unchanged (*Supporting Appendix*). Also, the CADS model suggests a more efficient dye-swap design where only one array per pair of samples is required. With an even number of arrays, the first half are labeled with one dye configuration, and the second half are labeled with the swapped dye configuration. Thus, only one array is formed for each experimental unit. However, by swapping the dye configuration used on some arrays, the information provided by this design is similar to that provided by the traditional dye-swap (*Supporting Appendix*). Future work will be focused on making CADS applicable to general and more efficient experimental designs.

6 Examples

Simulations. We illustrate CADS on simulated examples that build off of those discussed earlier (details given in *Supporting Appendix*). The array functions represent array-specific error and are random curves centered at the zero line.

To compare the methods we have discussed, we formed t-statistics for each gene that test for equality of expression between target and control. In order to compare power, plots of the number of genes called significant versus the number of false discoveries are shown in Figure 6. Due to the MA assumptions being violated in these examples, MA methods do not perform well, with fewer genes called significant for any given number of false positives. CADS increases power over simple dye-swap averaging. This is expected, since array effects essentially represent extra variation. By removing them, we decrease the variation of each gene's observations, thereby increasing power. We also compared the number of expected false positives under the correct null distribution versus the number of observed false positives. The MA method produces more observed false positives than expected, implying that significance would be artificially inflated when performing significance tests on MA normalized data when, say, estimating a p-value or false discovery rate (Storey & Tibshirani 2003).

*** Figure 6 about here. ***

Prostate development. We now illustrate CADS on the prostate development study described earlier. In terms of our intensity-dependent models, the dotted curves in the top row of Figure 4 are estimates of the $d_R(\mu_{Ti}) - d_G(\mu_{Ci})$, and the dotted curves in the bottom row are estimates of the $d_G(\mu_{Ti}) - d_R(\mu_{Ci})$. The experimental design employed here is an example of the more efficient dye-swap design mentioned above, since dye-swapping was carried out on biological, rather than technical, replicates. Thus, the CADS method described above must be adjusted for this example. To this end, note that we can interpret differences between the solid and dotted curves in Figure 4 as array effects. Removing these, we are left with three arrays following the $d_R(\mu_{Ti}) - d_G(\mu_{Ci})$ and three following $d_G(\mu_{Ti}) - d_R(\mu_{Ci})$. The average of these curves are the $d(\mu_{Ti}) - d(\mu_{Ci})$, so reshaping each array around this average curve effectively removes dye bias. These ideas will be formulated more explicitly in a subsequent publication.

Figure 5 shows the CADS-normalized arrays. In this example, differential expression is intensity-dependent, apparent from the remaining MA trend in Figure 5. Because MA-normalization does not allow intensity-dependent differential expression, the signal at mid-range abundances will be destroyed. This is apparent in Figure 7, which compares p-values (from two-sample t-tests and perumutation null distributions) after normalization by an MA-method and by CADS.

*** Figure 7 about here. ***

Since CADS preserves the systematic differential expression at mid-range abundances, its p-values in this region tend to be much smaller than those for MA-normalization. Thus, in this example, many truly differentially expressed genes would be missed if MA-normalization were used.

7 Discussion

We have shown that the dye-swap provides the minimal information required to remove dye effects from a single pair of samples. We have described a normalization method (CADS) capable of removing all common biases under general assumptions. Using flexible models, we have also shown that CADS produces unbiased estimates of the relative expression levels of the pair of samples hybridized to each array. CADS requires dye-swap arrays, and this leaves the question of what to do in their absence. We are currently working on an alternative dye-swap design that requires only one array per sample pair.

Reference designs (Churchill 2002) are widely used due to their flexibility. It is also assumed that reference designs avoid the problem of dye bias, since targets and controls are labeled with the same dye. Inference can, however, change depending on dye orientation (Dombkowski et al. 2004). Also, the indirect comparisons of a reference design require the most measurements to be made on the quantity of least interest. Thus, direct comparisons are more efficient. It is straightforward to construct examples where indirect comparisons have variances that are several times those of direct comparisons (Yang & Speed 2002). CADS is not limited to dye and array effects. If the effects of other covariates such as array "batches" are to be removed, CADS could be applied separately to groups with common covariate values, leaving the covariate effect in place. The remaining covariate effect could then be handled in the inference stage. However, we intend to extend CADS to handle additional covariates directly in future work. We also plan to extend CADS to produce reliable expression measurements on general experimental designs and to characterize cases where CADS can be applied by using only one array per pair of samples.

References

Benes, V. & Muckenthaler, M. (2003). Standardization of protocols in cDNA microarray analysis, *Trends in Biochemical Sciences* **28**: 244–249.

Bilban, M., Buehler, L., Head, S., Desoye, G. & Quaranta, V. (2002). Normalizing DNA microarray data, *Current Issues in Molecular Biology* **4**: 57–64.

Chen, Y., Dougherty, E. & Bittner, M. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images, *Journal of Biomedical Optics* **24**: 364–374.

Churchill, G. (2002). Fundamentals of experimental design for cDNA microarrays, *Nature Genetics* **32**: 490–495.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**: 829–836.

Cui, X. & Churchill, G. (2003). Statistical tests for differential expression in cDNA microarray experiments, *Genome Biology* **4**: 210.

Cui, X., Kerr, M. & Churchill, G. (2003). Transformations for cDNA microarray data, *Statistical Applications in Genetics and Molecular Biology* **2**: Article 4.

Dobbin, K., Kawasaki, E., Petersen, D. & Simon, R. (2005). Characterizing dye bias in microarray experiments, *Bioinformatics* **21**: 2430–2437.

Dombkowski, A., Thibodeau, B., Starcevic, S. & Novak, R. (2004). Gene-specific dye bias in microarray reference designs, *FEBS Lett.* .

Fan, J., Tam, P., Woude, G. & Ren, Y. (2004). Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine, *Proceedings of the National Academy of Sciences* **101**: 1135–1140.

Fang, Y., Brass, A., Hoyle, D., Hayes, A., Bashein, A., Oliver, S., Waddington, D. & Rattray, M. (2003). A model-based analysis of microarray experimental error and normalisation, *Nucleic Acids Research* **31**: e96.

Green, P. & Silverman, B. (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O., Wilfond, B., Borg, A. & Trent, J. (2001). Gene expression profiles in hereditary breast cancer, *New England Journal of Medicine* **344**: 539–548.

Hughes, T., Mao, M., Jones, A., Burchard, J., Marton, M., Shannon, K., Lefkowitz, S., Ziman, M., Schelter, J., Meyer, M., Kobayashi, S., Davis, C., Dai, H., He, Y., Stephaniants, S., Cavet, G., Walker, W., West, A., Coffey, E., Shoemaker, D., Stoughton, R., Blanchard, A., Friend, S. & Linsley, P. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer, *Nature Biotechnology* **19**: 342–347.

Kepler, T., Crosby, L. & Morgan, K. (2002). Normalization and analysis of DNA microarray data by self-consistency and local regression, *Genome Biology* **3**: research0037.

Kerr, M. & Churchill, G. (2001). Experimental design for gene expression microarrays, *Biostatistics* **2**: 183–201.

Kerr, M., Martin, M. & Churchill, G. (2000). Analysis of variance for gene expression microarray data, *Journal of Computational Biology* **7**: 819–837.

Kroll, T. & Wolfl, S. (2002). Ranking: A closer look on globalisation methods for normalisation of gene expression arrays, *Nucleic Acids Research* **30**: e50.

Qiu, X., Brooks, A. I., Klebanov, L. & Yakovlev, N. (2005). The effects of normalization on the correlation structure of microarray data, *BMC Bioinformatics* **6**: 120.

Quackenbush, J. (2002). Microarray normalization and transformation, *Nature Genetics* **32**: 496–501.

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

Storey, J. D. & Tibshirani, R. (2003). Statistical significance for genome-wide studies, *Proceedings of the National Academy of Sciences* **100**: 9440–9445.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G. & Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences*, in press.

Suzuki, T., Higgins, P. & Crawford, D. (2000). Control selection for RNA quantitation, *Biotechniques* **29**: 332–337.

Tseng, G., Oh, M., Rohlin, L., Liao, J. & Wong, W. (2001). Issues in cDNA microarray analysis: Quality filtering, channel normalization, models of variations and assessment of gene effects, *Nucleic Acids Research* **29**: 2540–2557.

Tusher, V., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences* **98**: 5116–5124.

Wilson, D., Buckley, M., Helliwell, C. & Wilson, I. (2003). New normalization methods for cDNA microarray data, *Bioinformatics* **19**: 1325–1332.

Xiao, Y., Segal, M. & Yang, Y. (2005). Stepwise normalization of two-channel spotted microarrays, *Statistical Applications in Genetics and Molecular Biology* **4**: Article 4.

Yang, I., Chen, E., Hasseman, J., Liang, W., Frank, B., Wang, S., Sharov, V., Saeed, A., White, J., Li, J., Lee, N., Yeatman, T. & Quackenbush, J. (2002a). Within the fold: Assessing differential expression measures and reproducibility in microarray assays, *Genome Biology* **3**: research0062.1–0062.12.

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

Yang, Y. H., Dudoit, S., Luu, P. & Speed, T. P. (2001). *Microarrays: Optical Technologies and Informatics*, SPIE, San Jose, CA, chapter Normalization of cDNA microarrays.

Yang, Y. & Speed, T. (2002). Design issues for cDNA microarray experiments, *Nature Review Genetics* **3**: 579–588.

Yue, H., Eastman, P., Wang, B., Minor, J., Doctolero, M., Nuttall, R., Stack, R., Becker, J., Montgomery, J., Vainer, M. & Johnston, R. (2001). An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression, *Nucleic Acids Research* **29**: e41.

Zien, A., Aigner, T., Zimmer, R. & Lengauer, T. (2001). Centralization: A new method for the normalization of gene expression data, *Bioinformatica* **1**: 1–9.

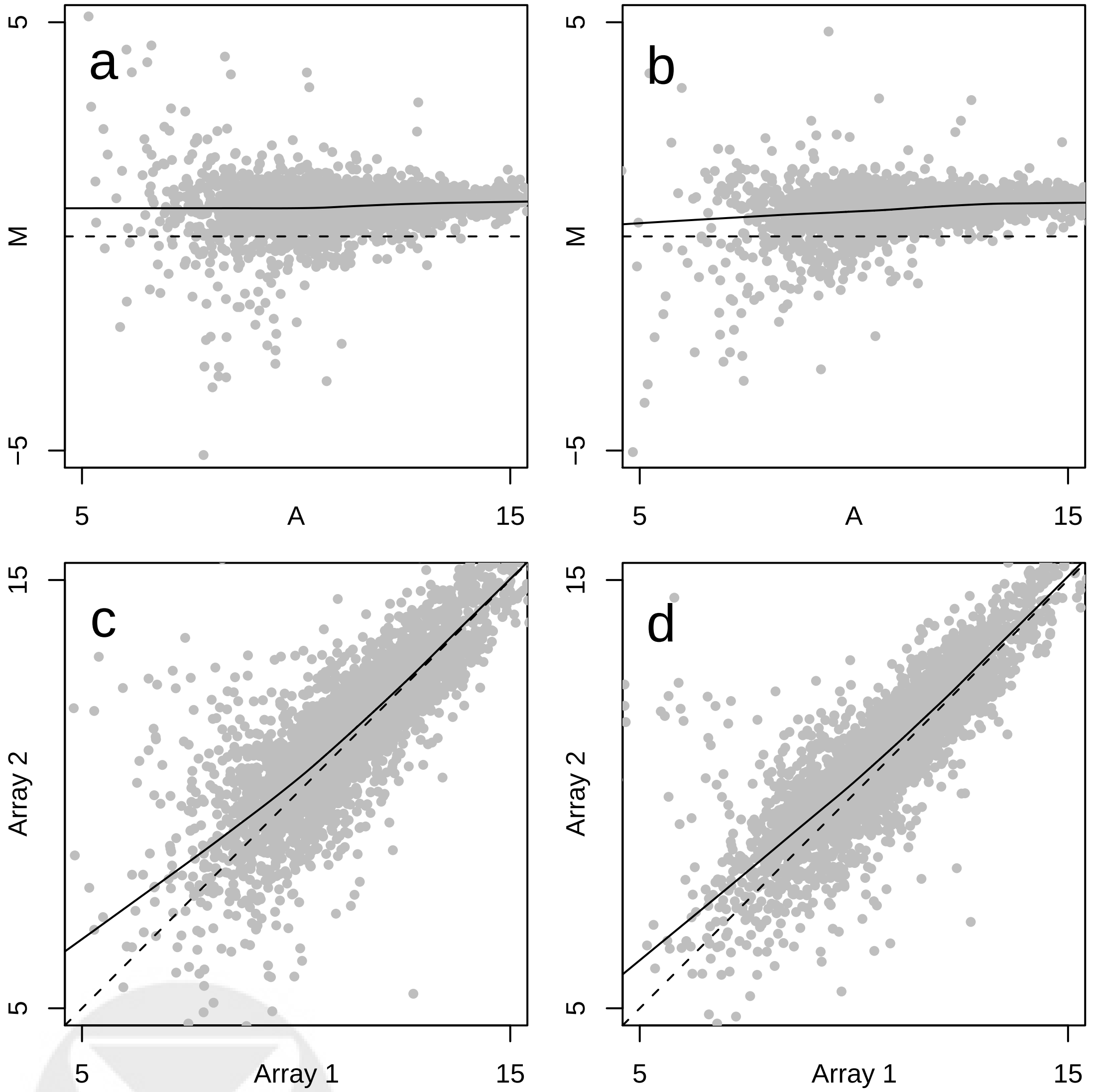

Figure 1: Self-self calibration experiments (Tseng *et al* (2001)). **(a & b)** MA-plots showing dye differences on the same sample. Without dye-bias, the point would be centered along dashed zero line. **(c & d)** Scatterplots of the same dyes applied to different arrays. Without array-effects, the points would be centered along dashed line of equality.

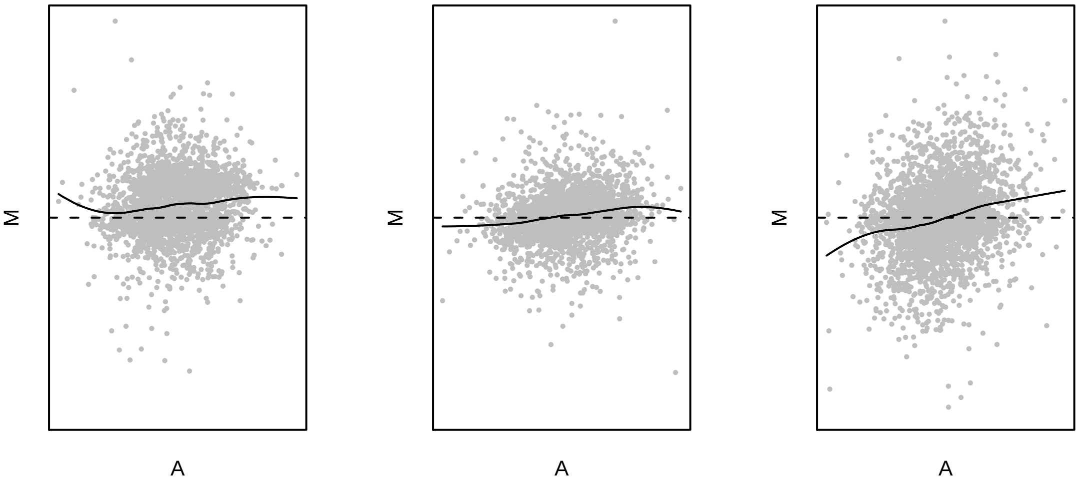

Figure 2: Three counterexamples for MA smoothing. From left to right, the plots illustrate MA trends due to asymmetric differential expression, asymmetric intensity-dependent differential expression, and unequal variation of expression means. All of the trends are due to biology, not sources of systematic bias. The solid lines are loess curves, showing that smoothing the MA-plots would create and/or destroy true differential expression signal.

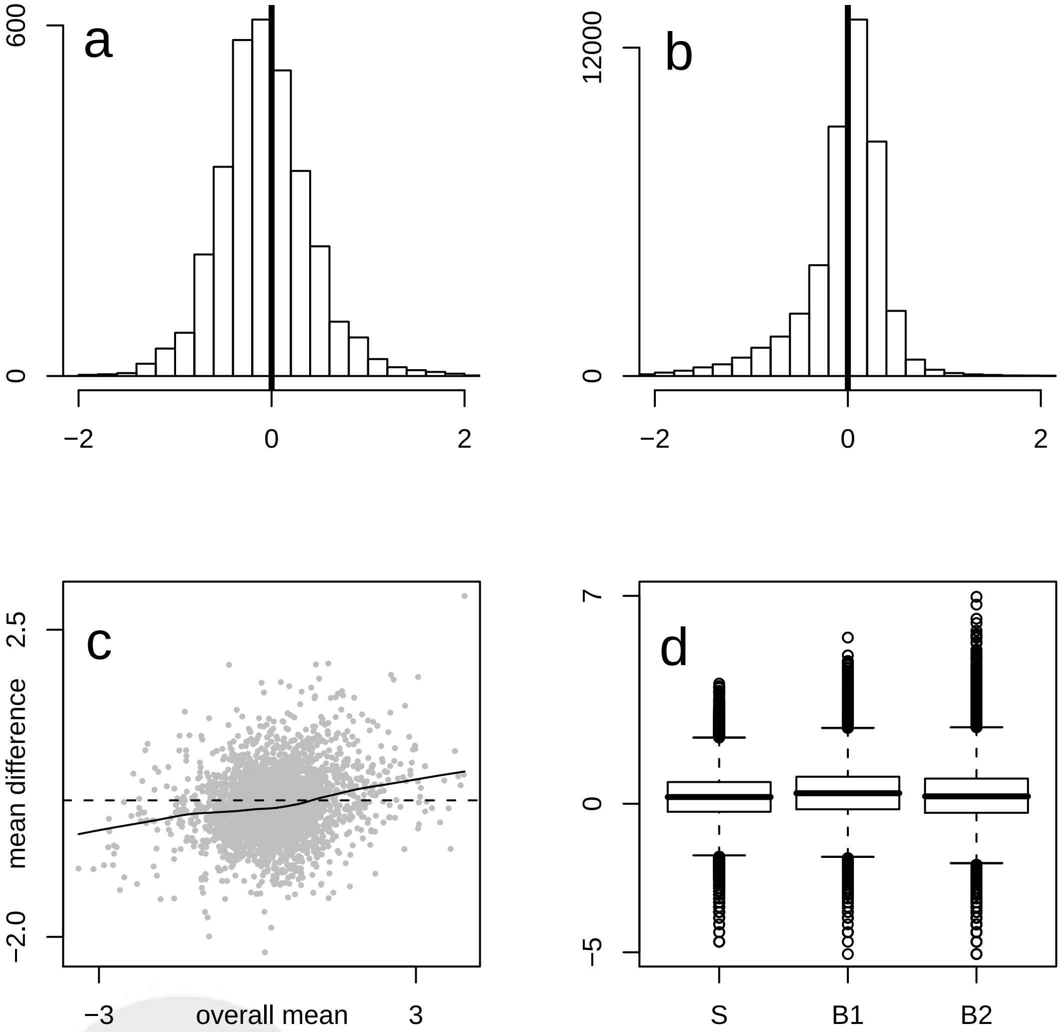

Figure 3: Examples of asymmetric differential expression, intensity dependent differential expression, and differences in variation of expression across samples. **(a)** Asymmetry among average $\log_2$ fold-change expression between BRCA1 and BRCA2 mutation positive tumors (Hedenfalk *et al* (2001)). **(b)** Asymmetry among average $\log_2$ fold-change expression between bloods cells of four endotoxin and four control individuals (Storey *et al* (2005)). **(c)** Average overall abundance versus $\log_2$ fold-change expression among BRCA1 and BRCA2 mutation positive tumors (Hedenfalk *et al* (2001)). **(d)** Boxplots of average $\log_2$ expression from tumors among Sporadic (S), BRCA1 (B1) and BRCA2 (B2) individuals.

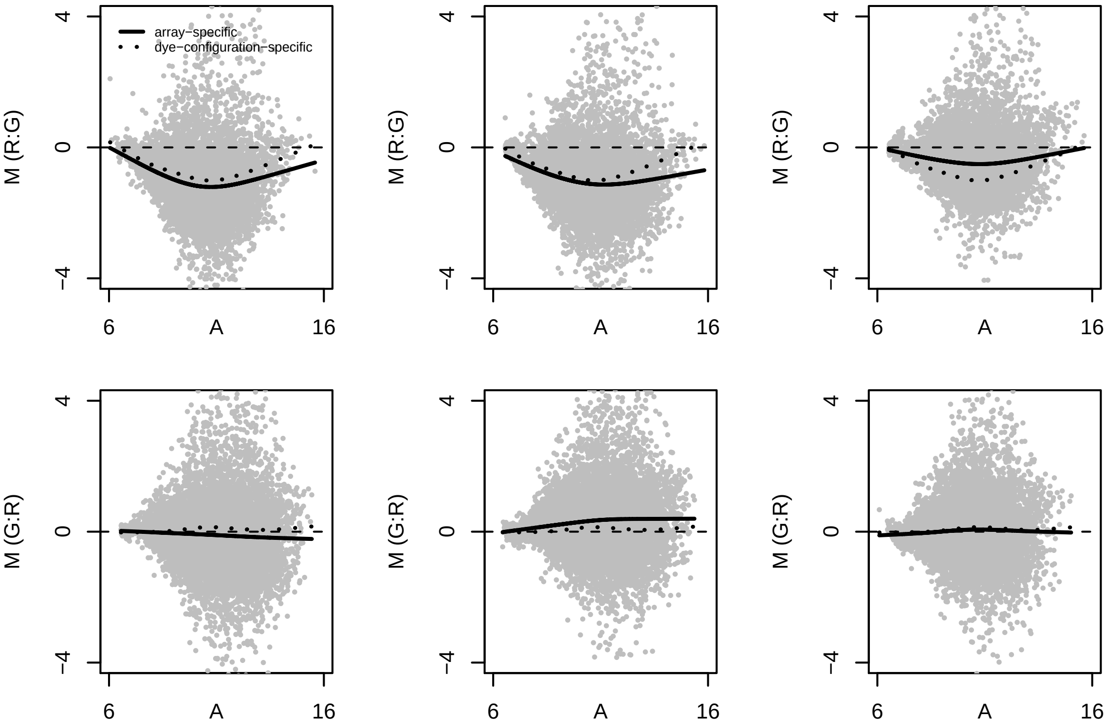

Figure 4: MA plots for six prostate samples. The top row is for arrays with target labeled red, control green. The bottom row is for arrays with target labeled green, control red. The solid lines are smoothers fit to each array individually (array-specific). The dotted lines are smoothers fit to all data within a particular dye configuration (dye-configuration-specific). For example, the dotted curve in the top row estimates the function $d_R(\mu_T) - d_G(\mu_C)$. Differences between the array-specific and dye-configuration-specific curves are estimates of array effects.

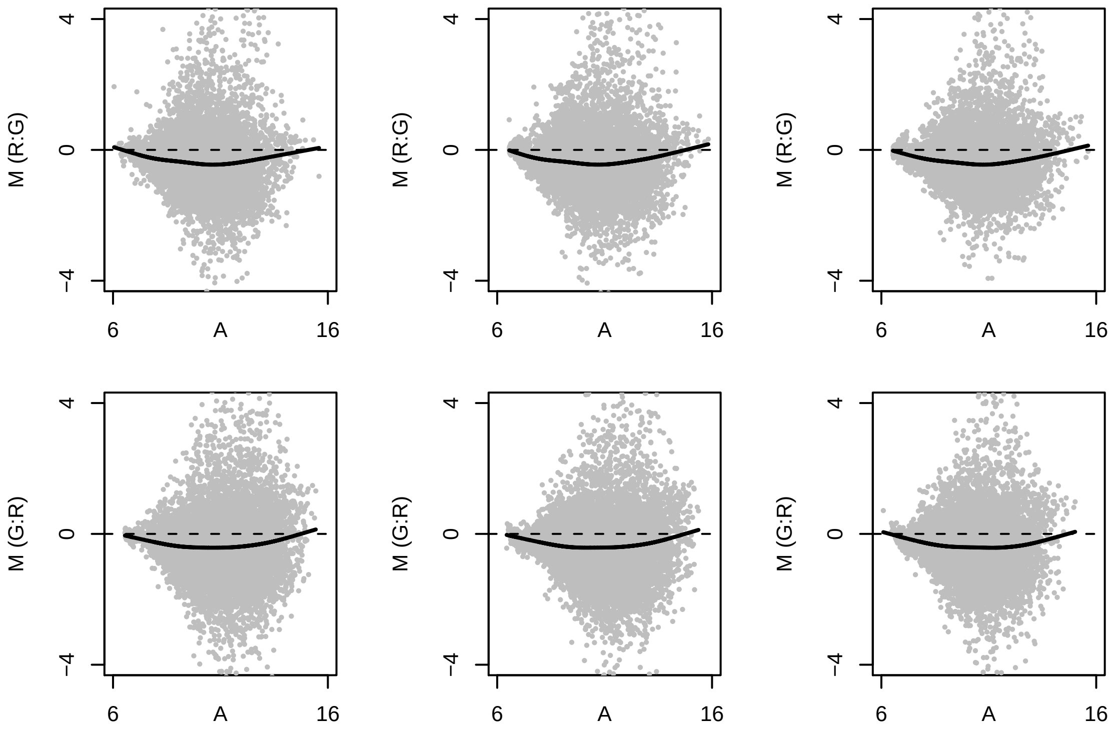

Figure 5: CADS-normalized arrays from the prostate example. Note the intensity-dependent trend that remains.

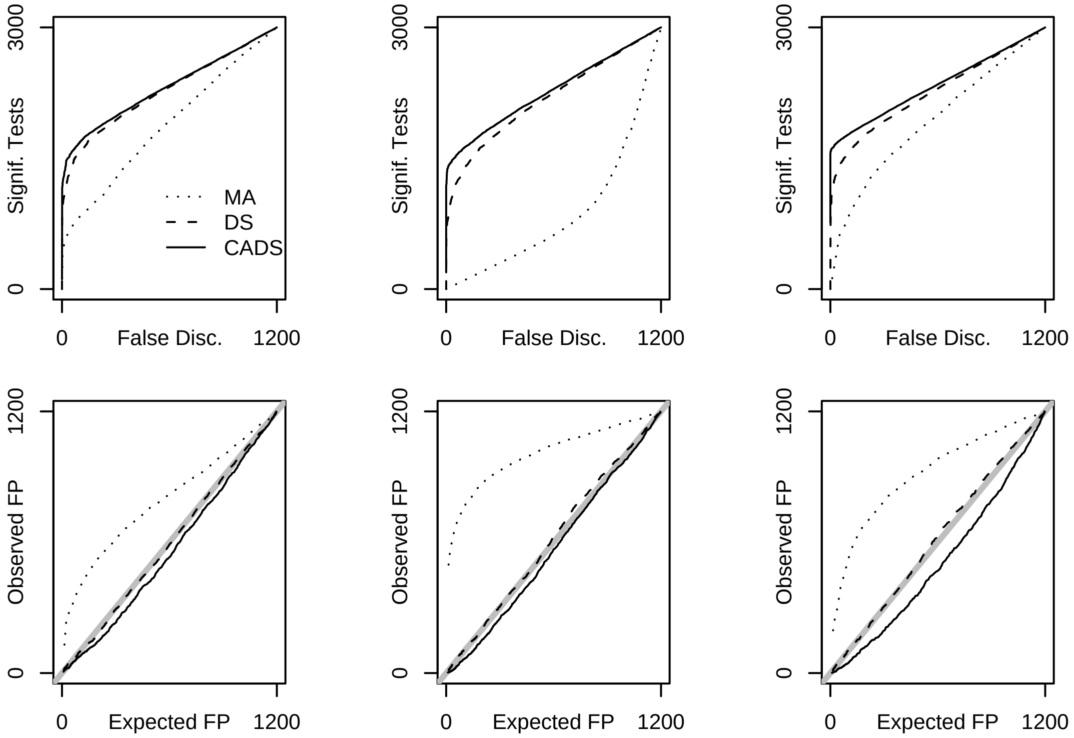

Figure 6: A comparison of MA, simple dye-swap average, and CADS normalization methods on the three simulated examples. Example number increases from left to right. The top row is a comparison based on false discoveries versus the number of significant tests. The bottom row are plots of the number of expected false positives (calculating under the correct null distribution) versus the number of observed false positives. The fact that the MA method produces more observed false positives than expected means that significance would be artificially inflated when performing significance tests on MA normalized data.

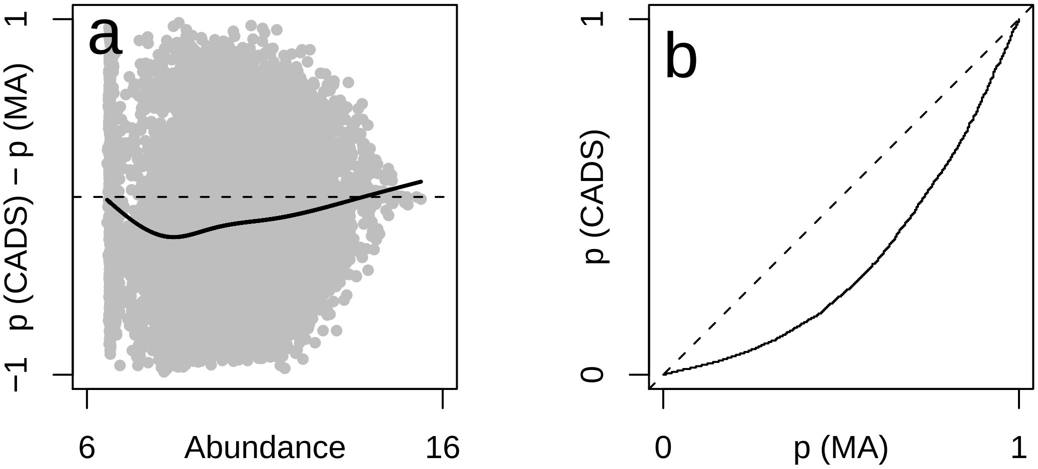

Figure 7: Comparison between p-values after CADS and MA-normalization, in the prostate example. **(a)** Differences between CADS and MA-based p-values, ordered by abundance. In this example, MA methods remove signal from genes with mid-range abundance. Thus, the p-values after CADS normalization are smaller than those after MA-normalization in this region. **(b)** QQ plot of the two sets of p-values, further illustrating that MA-normalization destroys signal in this example.

Supporting Appendix: A New Approach to Intensity-Dependent Normalization of Two-Channel Microarrays

Alan R. Dabney[*] and John D. Storey[*][†]

Contents

1 The Model and Assumptions Behind MA Methods 2

2 Simulation Details 3

3 Dye Swap Required for Removing Dye Bias in General 5

4 CADS Preserves Differential Expression Relationships 6

5 CADS Induces Negligible Dependence Between Arrays 9

6 Extensions to Other Experimental Designs 9

[*] Department of Biostatistics, University of Washington, Seattle, WA 98195

[†] To whom correspondence should be sent: `jstorey@u.washington.edu`

1 The Model and Assumptions Behind MA Methods

A two-channel microarray consists of target fluorescence intensities Y_{Ti} in one channel and control fluorescence intensities Y_{Ci} in the other, $i = 1, 2, \ldots, m$. MA methods form $M_i = Y_{Ti} - Y_{Ci}$ and $A_i = (Y_{Ti} + Y_{Ci})/2$, $i = 1, 2, \ldots, m$, plot these two quantities against each other, and force the points to be centered along the line $M = 0$. We now show that MA methods imply the following assumptions:

1. Overall, differential expression between the two samples is symmetric about zero.

2. There is no relationship between differential expression and expression abundance.

3. The variation of expression across genes is the same in each sample.

We begin by writing the model behind MA methods as

$$\mathrm{E}(M|A) = b(A) + s(A), \tag{5}$$

where b represents MA trends due to bias, and s represents MA trends due to biological signal. Since the goal of the MA method is to force $\mathrm{E}(M|A) = 0$, a smoother of M on A is subtracted from M. The claim is that this removes the bias $b(A)$. A fundamental assumption of MA methods is therefore that $s(A) = 0$. This implies assumption 2. We can also write model (5) as,

$$M_i = \delta_i + b(A_i) + s(A_i),$$

$i = 1, 2, \ldots, m$, where the $\delta_i + s(A_i)$ represent biological signal and noise. The model fit requires that $\sum \delta_i = 0$ (even locally, in the case of lowess smoothers), which in conjunction with the assumption that $s(A) = 0$ implies assumption 1.

The "variation" referred to in assumption 3 describes the spread of expression means *across all genes* under consideration. To be specific, let

$$\mathrm{Var}(\boldsymbol{Y_T}) = \frac{1}{m}\sum_{i=1}^{m}\left(\mathrm{E}(Y_{Ti}) - \frac{1}{m}\sum_{j=1}^{m}\mathrm{E}(Y_{Tj})\right)^2, \quad \text{and}$$

$$\mathrm{Var}(\boldsymbol{Y_C}) = \frac{1}{m}\sum_{i=1}^{m}\left(\mathrm{E}(Y_{Ci}) - \frac{1}{m}\sum_{j=1}^{m}\mathrm{E}(Y_{Cj})\right)^2.$$

Then $\mathrm{Var}(\boldsymbol{Y_T})$ describes the variation of expression across genes in targets, and $\mathrm{Var}(\boldsymbol{Y_C})$ is the analagous quantity for controls. Using similar notation, an MA trend occurs when there is covariation between the $Y_{Ti} - Y_{Ci}$ and the $Y_{Ti} + Y_{Ci}$:

$$\begin{aligned} \operatorname{Cov}(\boldsymbol{M},\ \boldsymbol{A}) &= \operatorname{Cov}(\boldsymbol{Y_T} - \boldsymbol{Y_C},\ \frac{1}{2}(\boldsymbol{Y_T} + \boldsymbol{Y_C})) \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(\mathrm{E}(Y_{Ti}) - \frac{1}{m}\sum_{j=1}^{m}\mathrm{E}(Y_{Tj})\right)\left(\mathrm{E}(Y_{Ci}) - \frac{1}{m}\sum_{j=1}^{m}\mathrm{E}(Y_{Cj})\right). \end{aligned} \tag{6}$$

In the absence of bias, MA methods assume that $\operatorname{Cov}(\boldsymbol{M},\ \boldsymbol{A}) = 0$. However, it follows from (6) that

$$\operatorname{Cov}(\boldsymbol{M},\ \boldsymbol{A}) = \frac{1}{2}\left(\operatorname{Var}(\boldsymbol{Y_T}) - \operatorname{Var}(\boldsymbol{Y_C})\right).$$

Thus, if the variation of expression across genes is not the same in each sample, there will be an MA trend, even in the absence of any bias. This implies assumption 3.

2 Simulation Details

Example 1: Asymmetric Differential Expression. Let $0 < \gamma < 100/3$. Suppose that $2\gamma\%$ of the genes have $\mu_T - \mu_C = \theta \neq 0$, another $\gamma\%$ have $\mu_T - \mu_C = -\theta 0$, and the remaining $1 - 3\gamma\%$ of the genes have equal expression. Let D_+ and D_- be the set of indices of the over-expressed and under-expressed genes, respectively. Similarly, D_0 is the set of indices of the null genes. Then

$$\mu_{Ti} - \mu_{Ci} = \begin{cases} \theta & i \in D_+, \\ -\theta & i \in D_-, \\ 0 & i \in D_0. \end{cases}$$

The overall average signal is

$$\frac{1}{m}\sum_{i=1}^{m}(\mu_{Ti} - \mu_{Ci}) = \frac{1}{m}\left(\frac{\gamma m\theta}{50} - \frac{\gamma m\theta}{100}\right) = \frac{\gamma\theta}{100} \neq 0,$$

so that $s(A) \neq 0$.

The data used in the section Counterexamples to MA Normalization Methods in the main paper were simulated as follows. First, we generated 3000 control means $\mu_{C1}, \mu_{C2}, \ldots,\ \mu_{C3000}$ from the $N(0, 4)$ distribution. We then randomly chose 40% of the genes to be over-expressed and 20% to be under-expressed. Target means were either control means plus two, minus 0.5, or unchanged, depending on whether they were in the 40% over-expressed, 20% under-expressed, or remaining 40% null groups. Sample variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_{3000}^2$ were generated from the $\chi^2(1)$ distribution. Finally, target and control observations for gene i were generated from the $N(\mu_{Ti}, \sigma_i^2)$

and $N(\mu_{Ci}, \sigma_i^2)$ distributions, respectively.

The simulation details for the data used in the section *Examples* in the main paper were similar, with the addition of artificial dye functions and array functions. Again, 3000 control means were generated first, this time from the $N(0, 25)$ distribution. Target means were either control means plus 0.75, minus 0.5, or unchanged, depending on whether they were in the 40% over-expressed, 20% under-expressed, or remaining 40% null groups. Sample variances were generated from the $\chi^2(1)$ distribution. Dye functions were chosen as nonlinear curves which did not deviate too greatly from the line of equality. Array functions were randomly generated as curves centered at the zero line. Figure 8 summarizes the simulation setup for Example 2, below. The dye functions used here are equivalent to those in the figure, while the array functions are similar, being random observations from the same distribution.

*** Figure 8 about here. ***

Example 2: Intensity-Dependent Differential Expression. Suppose that differential expression increases with increasing abundance. The differentially expressed genes can be modeled as $\mu_{Ti} = h(\mu_{Ci})$, where h describes a relationship between differential expression and intensity. Consider the scatterplot with observed abundance on the x-axis and true signal on the y-axis. For null genes, the true signal is zero. For differentially expressed genes, the true signal is $h(\mu_C) - \mu_C$. While the randomness of the observed abundances will prevent the observations on the y-axis from lining up perfectly, there will be a nonzero trend in general, so that $s(A) \neq 0$.

The data used in the section Smoothing MA Plots in the main paper were simulated as follows. First, we generated 3000 control means as in Example 1. We randomly chose 60% of the genes to be differentially expressed. We then formed the target mean for the differentially expressed gene i as a (symmetric about the line of equality) increasing function of μ_{Ci}, with $\mu_{Ti} = -0.1 + 0.3\mu_{Ci} + 2.7\mu_{Ci}^2 - 1.8\mu_{Ci}^3$. Target and control observations for each gene were generated using Normal distributions, as in Example 1. For the data in the section Examples in the main paper, control means were generated from the $N(0, 25)$ distribution, and a slightly less extreme functional relationship was used between the target and control means. Sample variances were generated from the $\chi^2(1)$ distribution. Dye and array functions were chosen as in Example 1. Again, see Figure 8 for a summary of the simulation for this example.

Example 3: Unequal Variation in Expression Means. The variation in expression means can be quantitated by $\sum_{i=1}^m \left(\mu_{Ti} - \frac{1}{m}\sum_{i=1}^m \mu_{Ti}\right)^2$ and $\sum_{i=1}^m \left(\mu_{Ci} - \frac{1}{m}\sum_{i=1}^m \mu_{Ci}\right)^2$ for targets and controls, respectively. Suppose that the former is greater than the latter. Consider the scatterplot with observed abundance on the x-axis and true signal on the y-axis. At extreme values of abundance, the target values will tend to be further away from their control counterparts. On the high end of abundance, the target values are likely much larger than controls. Similarly, on the low end of abundance, the target values are likely much smaller than controls. There will then be a trend away from the zero line, so that $s(A) \neq 0$.

The data used in the section Smoothing MA Plots in the main paper were simulated as follows. First, we generated 3000 control means as in Examples 1 and 2. We randomly chose 60% of the genes to be differentially expressed. The target mean for the differentially expressed gene i were generated from the $N(\mu_{Ci}, 4)$ distribution. Target and control observations for each gene were generated using Normal distributions, as in Examples 1 and 2. For the data in the section Examples in the main paper, control means were generated from the $N(0, 25)$ distribution, and the target means were then generated from the $N(\mu_{Ci}, 25)$ distributions, $i = 1, 2, \ldots, 3000$. Sample variances were generated from the $\chi^2(1)$ distribution. Dye and array functions were chosen as in Example 1.

3 Dye Swap Required for Removing Dye Bias in General

It can be shown that a dye swap provides the minimal information required for parsing dye bias from biological signal. (This claim holds for the case where dye bias is removed only using a single pair of biological samples.) Suppose that no dye-swap is performed and, say, the target is labeled with the red dye and control with the green dye. It then follows that

$$\begin{aligned} Y_{Ti} &= d_R(X_{Ti}) + \epsilon_{Ti}, \\ Y_{Ci} &= d_G(X_{Ci}) + \epsilon_{Ci} \\ &= d_R(X_{Ci}) + [d_G(X_{Ci}) - d_R(X_{Ci})] + \epsilon_{Ci}. \end{aligned} \tag{7}$$

Therefore, there are two functions that need to be estimated: d_R and $d_G - d_R$. The function d_R represents the biological signal, and the function $d_G - d_R$ represents the dye bias. (Here, we have chosen to call d_R the "true" dye function, although the arguments hold when calling any weighted average of the two the "true" dye function.) The implicit goal of normalization is to estimate these two functions. This problem can be made clearer by removing the noise:

$$\begin{aligned} Y_{Ti} &= d_R(X_{Ti}), \\ Y_{Ci} &= d_R(X_{Ci}) + [d_G(X_{Ci}) - d_R(X_{Ci})]. \end{aligned}$$

Even in this instance the two functions cannot be individually determined. The reason is that the arguments in the functions, X_{Ti} and X_{Ci}, are never observed. Therefore, for any fixed definition of d_R, it is possible to define an infinite number of functions $d_G - d_R$ where the above equalities hold by simply changing the values assumed for X_{Ti} and X_{Ci}. Therefore, it is impossible to determine d_R and $d_G - d_R$ with only one dye configuration.

Note that in the self-self experiments it is known that $X_{Ti} = X_{Ci}$, so it is easy to see how an MA method can be utilized there to uniquely determine d_R and $d_G - d_R$; specifically, $d_G - d_R = Y_C - Y_T$. The way that MA methods get around the case where there is no equivalent expression is that they assume a certain relationship between $d_G - d_R$ and the X_{Ti} and X_{Ci}, thereby reducing the allowable functions and permitting a unique solution to be found. However, we have shown that these constraints are not true in general.

We can show that adding one more unit of information (i.e., another set of array measurments) allows fluorescence intensities to be observed that do not depend on dye. Therefore, this is proof that a dye swap is the minimal required information for removing dye bias in general. With a dye swap we have:

$$\begin{aligned}
Y_{TRi} &= d_R(X_{Ti}) + \epsilon_{TRi}, \\
Y_{CGi} &= d_G(X_{Ci}) + \epsilon_{CGi}, \\
Y_{TGi} &= d_G(X_{Ti}) + \epsilon_{TGi}, \\
Y_{CRi} &= d_R(X_{Ci}) + \epsilon_{CRi}.
\end{aligned}$$

By performing the dye averages, we have:

$$\begin{aligned}
(Y_{TRi} + Y_{TGi})/2 &= d(X_{Ti}) + (\epsilon_{TRi} + \epsilon_{TGi})/2, \\
(Y_{CRi} + Y_{CGi})/2 &= d(X_{Ci}) + (\epsilon_{CRi} + \epsilon_{CGi})/2,
\end{aligned}$$

where $d = (d_R + d_G)/2$. Therefore, this extra information obviates the need to ever know the exact arguments X_{Ti} and X_{Ci} in extracting fluorescence intensities that do not depend on dye. The trick is that the function arguments remained constant when the dye assignments were exchanged.

4 CADS Preserves Differential Expression Relationships

CADS produces normalized data that preserve differential expression relationships in expectation. This means that, on average, null genes will be called null, over-expressed genes will be called over-expressed, and under-expressed genes will be called under-expressed. Specifically, under very general conditions, we show that the CADS estimator is unbiased for the parameters of interest $d(\mu_{Ti}) - d(\mu_{Ci})$, $i = 1, 2, \ldots, m$. Since statistical inference is almost always based on sample averages (Cui & Churchill 2003), it is sufficient to show that the sample average of the CADS-normalized arrays is unbiased.

Theorem 1 *Let $\hat{M}_{ij}$ be the CADS estimate for gene i and sample pair j, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$. Let μ_{Ti} and μ_{Ci} be the population average RNA amounts for gene i in targets and controls, respectively. With d_R and d_G the red and green dye functions, respectively, let $d = \frac{1}{2}(d_R + d_G)$ be the average dye function. Finally, with $a_1, a_2, \ldots, a_n$ independent random curves representing array effects, let $a_0 = Ea_j$, $j = 1, 2, \ldots, n$. Then, $E\left(\frac{1}{n}\sum_{j=1}^{n}\hat{M}_{ij}\right) = d(\mu_{Ti}) - d(\mu_{Ci}) + a_0(\mu_{Ti}) - a_0(\mu_{Ci})$, $i = 1, 2, \ldots, m$.*

Theorem 1 is proved below. Two results follow directly.

Lemma 1 *If a_0 is nondecreasing, then*

1. $d(\mu_{Ti}) = d(\mu_{Ci}) \Rightarrow E\left(\frac{1}{n}\sum_{j=1}^{n}\hat{M}_{ij}\right) = 0$,
2. $d(\mu_{Ti}) > d(\mu_{Ci}) \Rightarrow E\left(\frac{1}{n}\sum_{j=1}^{n}\hat{M}_{ij}\right) \geq d(\mu_{Ti}) - d(\mu_{Ci})$,
3. $d(\mu_{Ti}) < d(\mu_{Ci}) \Rightarrow E\left(\frac{1}{n}\sum_{j=1}^{n}\hat{M}_{ij}\right) \leq d(\mu_{Ti}) - d(\mu_{Ci})$,

$i = 1, 2, \ldots, m$.

Thus, if the array functions are nondecreasing on average, null genes remain null and the sign of differential expression is preserved, in expectation. Note that the case where the array functions average to a constant is included here. However, when the array functions average to the constant zero, we have an even stronger result.

Lemma 2 *If a_0 equals the zero line, then $E\left(\frac{1}{n}\sum_{j=1}^{n}\hat{M}_{ij}\right) = d(\mu_{Ti}) - d(\mu_{Ci})$, $i = 1, 2, \ldots, m$.*

That is, CADS is unbiased for the parameters of interest.

Proof of Theorem 1. The CADS model can be written in vector form as

$$\begin{aligned}
\boldsymbol{Y_{TRj}} &= \boldsymbol{d_R}(\boldsymbol{X_{Tj}}) + \boldsymbol{a_{1j}}(\boldsymbol{X_{Tj}}) + \boldsymbol{\epsilon_{TRj}}, \\
\boldsymbol{Y_{CGj}} &= \boldsymbol{d_G}(\boldsymbol{X_{Cj}}) + \boldsymbol{a_{1j}}(\boldsymbol{X_{Cj}}) + \boldsymbol{\epsilon_{CGj}}, \\
\boldsymbol{Y_{TGj}} &= \boldsymbol{d_G}(\boldsymbol{X_{Tj}}) + \boldsymbol{a_{2j}}(\boldsymbol{X_{Tj}}) + \boldsymbol{\epsilon_{TGj}}, \\
\boldsymbol{Y_{CRj}} &= \boldsymbol{d_R}(\boldsymbol{X_{Cj}}) + \boldsymbol{a_{2j}}(\boldsymbol{X_{Cj}}) + \boldsymbol{\epsilon_{CRj}},
\end{aligned}$$

$j = 1, 2, \ldots, n$. Each function is now vector-valued, so that, for example, $\boldsymbol{d_R}\left(\boldsymbol{X_{Tj}}\right)$ has ith component $d_R\left(X_{Tij}\right)$, $i = 1, 2, \ldots, m$. Averaging the two dye swap arrays gives

$$\tilde{\boldsymbol{Y}}_{\boldsymbol{Tj}} = \frac{1}{2}\Big(\boldsymbol{Y_{TRj}} + \boldsymbol{Y_{TGj}}\Big) = \boldsymbol{d}\left(\boldsymbol{X_{Tj}}\right) + \boldsymbol{a_j}\left(\boldsymbol{X_{Tj}}\right) + \boldsymbol{\epsilon_{Tj}},$$
$$\tilde{\boldsymbol{Y}}_{\boldsymbol{Cj}} = \frac{1}{2}\Big(\boldsymbol{Y_{CRj}} + \boldsymbol{Y_{CGj}}\Big) = \boldsymbol{d}\left(\boldsymbol{X_{Cj}}\right) + \boldsymbol{a_j}\left(\boldsymbol{X_{Cj}}\right) + \boldsymbol{\epsilon_{Cj}}.$$

Finally, subtracting these two quantities is equivalent to normalizing by simple dye-swap averaging, producing $\tilde{\boldsymbol{M}}_{\boldsymbol{j}} = \tilde{\boldsymbol{Y}}_{\boldsymbol{Tj}} - \tilde{\boldsymbol{Y}}_{\boldsymbol{Cj}}$. For each array j, CADS fits the model $\tilde{\boldsymbol{M}}_{\boldsymbol{j}} - \tilde{\boldsymbol{\Delta}} = \boldsymbol{f_j}\left(\tilde{\boldsymbol{\Delta}}\right) + \boldsymbol{\epsilon_j}$, where $\boldsymbol{f_j}$ is a smooth function, and $\tilde{\boldsymbol{\Delta}} = \frac{1}{n}\sum_{i=1}^{m} \tilde{\boldsymbol{M}}_{\boldsymbol{j}}$. To fix ideas, let us assume that $\boldsymbol{f_j}$ can be represented by $\boldsymbol{B_\Delta \beta_j}$, where $\boldsymbol{B_\Delta}$ is a known basis matrix evaluated at $\boldsymbol{d}\left(\boldsymbol{\mu_T}\right) - \boldsymbol{d}\left(\boldsymbol{\mu_C}\right)$; note that this implies $\boldsymbol{a_j}\left(\boldsymbol{\mu_{Ti}}\right) - \boldsymbol{a_j}\left(\boldsymbol{\mu_{Ci}}\right) = \boldsymbol{B_\Delta \beta_j}$. Any polynomial or spline function could be represented this way. We then use least-squares to estimate

$$\tilde{\boldsymbol{a}}_{\boldsymbol{j}} = \hat{\boldsymbol{f}}_{\boldsymbol{j}}\left(\tilde{\boldsymbol{\Delta}}\right) = \boldsymbol{B}_{\tilde{\boldsymbol{\Delta}}}\left(\boldsymbol{B}_{\tilde{\boldsymbol{\Delta}}}{}^{T}\boldsymbol{B}_{\tilde{\boldsymbol{\Delta}}}\right)^{-1}\boldsymbol{B}_{\tilde{\boldsymbol{\Delta}}}{}^{T}\left(\tilde{\boldsymbol{M}}_{\boldsymbol{j}} - \tilde{\boldsymbol{\Delta}}\right) = \boldsymbol{H}_{\tilde{\boldsymbol{\Delta}}}\left(\tilde{\boldsymbol{M}}_{\boldsymbol{j}} - \tilde{\boldsymbol{\Delta}}\right),$$

where $\boldsymbol{H}_{\tilde{\boldsymbol{\Delta}}} = \boldsymbol{B}_{\tilde{\boldsymbol{\Delta}}}\left(\boldsymbol{B}_{\tilde{\boldsymbol{\Delta}}}{}^{T}\boldsymbol{B}_{\tilde{\boldsymbol{\Delta}}}\right)^{-1}\boldsymbol{B}_{\tilde{\boldsymbol{\Delta}}}{}^{T}$ is a “hat matrix.” Thus, CADS normalizes array j to form the estimate

$$\hat{\boldsymbol{M}}_{\boldsymbol{j}} = \tilde{\boldsymbol{M}}_{\boldsymbol{j}} - \hat{\boldsymbol{f}}_{\boldsymbol{j}}\left(\tilde{\boldsymbol{\Delta}}\right) = \boldsymbol{H}_{\tilde{\boldsymbol{\Delta}}}\left(\tilde{\boldsymbol{M}}_{\boldsymbol{j}} - \tilde{\boldsymbol{\Delta}}\right),$$

$j = 1, 2, \ldots, n$. Therefore,

$$\begin{aligned}
\mathrm{E}\left(\frac{1}{n}\sum_{j=1}^{n}\hat{\boldsymbol{M}}_{\boldsymbol{j}}\right) &= \mathrm{E}\left(\frac{1}{n}\sum_{j=1}^{n}\left(\tilde{\boldsymbol{M}}_{\boldsymbol{j}} - \hat{\boldsymbol{f}}_{\boldsymbol{j}}\left(\tilde{\boldsymbol{\Delta}}\right)\right)\right) \\
&= \mathrm{E}\left(\frac{1}{n}\sum_{j=1}^{n}\left(\tilde{\boldsymbol{M}}_{\boldsymbol{j}} - \boldsymbol{H}_{\tilde{\boldsymbol{\Delta}}}\left(\tilde{\boldsymbol{M}}_{\boldsymbol{j}} - \tilde{\boldsymbol{\Delta}}\right)\right)\right) \\
&= \mathrm{E}\left(\frac{1}{n}\sum_{j=1}^{n}\tilde{\boldsymbol{M}}_{\boldsymbol{j}}\right) - \mathrm{E}\left(\boldsymbol{H}_{\tilde{\boldsymbol{\Delta}}}\frac{1}{n}\sum_{j=1}^{n}\left(\tilde{\boldsymbol{M}}_{\boldsymbol{j}} - \tilde{\boldsymbol{\Delta}}\right)\right) \\
&= \boldsymbol{d}\left(\boldsymbol{\mu_T}\right) - \boldsymbol{d}\left(\boldsymbol{\mu_C}\right) + \boldsymbol{a_0}\left(\boldsymbol{\mu_T}\right) - \boldsymbol{a_0}\left(\boldsymbol{\mu_C}\right),
\end{aligned}$$

since $\sum_{j=1}^{n}\left(\tilde{\boldsymbol{M}}_{\boldsymbol{j}} - \tilde{\boldsymbol{\Delta}}\right) = 0$. □

We note from the proof that

Remark 1 $\sum_{j=1}^{n}\hat{M}_{ij} = \sum_{j=1}^{n}\tilde{M}_{ij}$, $i = 1, 2, \ldots, m$.

This fact makes the proof of Theorem 1 even simpler, since $\mathrm{E}\left(\tilde{M}_{ij}\right) = d\left(\mu_{Ti}\right) - d\left(\mu_{Ci}\right) + a_0\left(\mu_{Ti}\right) - a_0\left(\mu_{Ci}\right)$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$. The proofs of the Lemmas are trivial and not shown here.

5 CADS Induces Negligible Dependence Between Arrays

Since CADS is a multiple array normalization technique, there is a concern that the arrays are made dependent even when they are composed of independently sampled biological. By refitting each array around the estimated average dye function, we induce some amount of dependence, since all arrays were used to estimate the average dye. However, the dependence induced by this procedure is minimal because, in comparison to the number of observations among all genes and arrays, a very low number of degrees of freedom are used when removing array effects. In order to demonstrate this, we performed a simple simulation to examine the effect of induced dependence. We generated 500 null means, then used the model to translate these means into expression profiles on 10 dye-swap array pairs. Applying CADS to the simulated data, we computed two-sided p-values using t-statistics. Since the null hypothesis is true for all genes, the p-values should be uniformly distributed on the interval $(0, 1)$ (Lehmann 1997). If substantial dependence was induced, the p-value distribution would be distorted. To test this, we performed a Kolmogorov-Smirnov goodness-of-fit test comparing the observed p-value distribution to the uniform $(0, 1)$ distribution. This was repeated 100 times. The KS p-values were less than 0.05 five times, as expected. Thus, we argue that the dependence induced by CADS will have little effect on inference.

6 Extensions to Other Experimental Designs

Reference Designs. A reference design compares targets and controls to a common reference (Churchill 2002). Both targets and controls are labeled with the same dye, and all reference material is labeled with the other dye. Two arrays are formed for each target/control pair, one comparing target to reference and the other comparing control to reference. Targets and controls are compared to each other indirectly as the difference between the target/reference and control/reference differences. Since the same dye is used to lable targets and controls, dye bias is assumed to not be an issue; however, inference can change depending on dye orientation (Dombkowski et al. 2004). Furthermore, the reference design is easily extended to include more samples and/or more comparison groups.

However, the reference design is inefficient when compared to designs that directly compare targets and controls. This is because a reference design makes the most measurements on the quantity of least interest. It is straightforward to construct examples where observations from a reference design have variances that are several times those from a design making direct comparisons (Yang & Speed 2002).

More Than Two Comparison Groups. CADS can easily be extended to more than two comparison groups. Suppose, for example, that we would like to compare targets A, B, and C. The key is to design the experiment so that dye-swap information is obtained on all three targets, using as few arrays as possible. This can be accomplished with array triplets using a loop design (Kerr & Churchill 2001, Churchill 2002), as described in the following figure. Arrows represent arrays. Tails of arrows represent one dye, and heads of arrows represent the other.

$$\begin{tikzcd} & T_A \arrow[dl] & \\ T_B \arrow[rr] & & T_C \arrow[ul] \end{tikzcd}$$

For each of the three pairwise comparisons, we would separately implement CADS. For example, to compare T_A to T_B, we would form the quantities $\tilde{Y}_A = \frac{1}{2}\left[Y_{AR} + Y_{AG}\right]$ and $\tilde{Y}_B = \frac{1}{2}\left[Y_{BR} + Y_{BG}\right]$, then subtract these to form $\tilde{Y} = \tilde{Y}_A - \tilde{Y}_B$. There are now three array functions involved instead of two, but the same principles apply. Thus, proceeding from this point with CADS will remove both dye and array effects from our T_A/T_B comparison. The other two pairwise comparisons are normalized analogously.

A More Efficient Dye-Swap Design. In principle, the CADS model allows for a more efficient dye-swap design, where only n arrays are required instead of $2n$. Suppose that n is an even number. Then, on the first $n/2$ arrays, we label targets with red and controls with green, say. On the second $n/2$ arrays, we swap the dye orientation. The difference between this and a traditional dye swap is that each sample is represented on only one array here, whereas they would be traditionally represented on two arrays. Applying CADS to the target/control comparisons in each group separately would produce unbiased estimates of the functions $d_R(\mu_T) - d_G(\mu_C)$ and $d_G(\mu_T) - d_R(\mu_C)$. We could then treat group membership as a covariate to be adjusted for in the inference stage. That is, `Group` $= 0$ on the first $n/2$ arrays and `Group` $= 1$ on the second $n/2$ arrays. Adjusting for `Group` in the inference stage will effectively compare target to control under the same dye orientation. This "reduced dye swap" would accomplish the same thing as a "full dye swap", but with half the arrays. There would of course be a decrease in precision due to the lower number of independent observations. We intend to develop CADS for this design in future work.

References

Churchill, G. (2002). Fundamentals of experimental design for cDNA microarrays, *Nature Genetics* **32**: 490–495.

Cui, X. & Churchill, G. (2003). Statistical tests for differential expression in cDNA microarray experiments, *Genome Biology* **4**: 210.

Dombkowski, A., Thibodeau, B., Starcevic, S. & Novak, R. (2004). Gene-specific dye bias in microarray reference designs, *FEBS Lett.* .

Kerr, M. & Churchill, G. (2001). Experimental design for gene expression microarrays, *Biostatistics* **2**: 183–201.

Lehmann, E. (1997). *Testing Statistical Hypotheses*, Springer.

Yang, Y. & Speed, T. (2002). Design issues for cDNA microarray experiments, *Nature Review Genetics* **3**: 579–588.

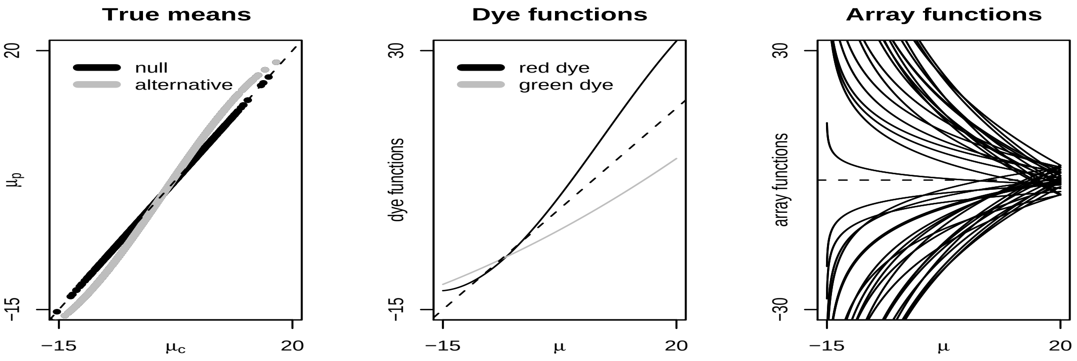

Figure 8: A summary of the simulation model used for example two in the section *Counterexamples to MA Normalization Methods* in the main paper. The other two examples were simulated analogously, with differences being evident in the plot of population means.