

A new approach to intensity-dependent normalization of two-channel microarrays

ALAN R. DABNEY, JOHN D. STOREY*

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
jstorey@u.washington.edu

SUMMARY

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, which limit their general applicability. 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 removes dye bias from a single 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 under the assumptions of the model.

Keywords: Differential expression; Expression arrays; Loess normalization; Nonparametric regression.

1. INTRODUCTION

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

In a two-channel microarray, each of two matched samples is labeled with a distinct dye, Cy3 (green) or Cy5 (red), and both are competitively hybridized to the same array. The relative dye intensity for each spot is used to quantify expression for its corresponding gene. Due to various sources of systemic effects, the relative expression levels need to be adjusted to accurately reflect the underlying amount of

*To whom correspondence should be addressed.

mRNA from each sample. For example, the two dyes incorporate into samples at different rates, creating “dye bias” (Tseng *and others*, 2001; Yang *and others*, 2001, 2002). In addition to dye bias, there may be systematic differences between arrays. It is therefore necessary to remove dye bias and any other systematic effects from two-channel microarray measurements.

To address this problem, we investigate the dye-swap design, showing that a dye-swap removes dye bias in general from a single pair of samples, an observation related to others previously made (Kerr *and others*, 2000; Kerr and Churchill, 2001; Fang *and others*, 2003). We show that while the original justification of the dye-swap (Kerr *and others*, 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. Within this framework, we show that a simple dye-swap average removes dye bias without affecting biological signal and preserves the ordering of true expression means. We then propose the “common array dye-swap” (CADS) method, a two-stage multi-array 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 approximation 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. Specifically, CADS preserves whether there is no differential expression, overexpression, or underexpression among the “true” population-level mRNA levels. Furthermore, CADS achieves this accuracy without requiring any of the strict assumptions of “MA-plot” (Tseng *and others*, 2001; Yang *and others*, 2002) or global (Chen *and others*, 1997) normalization methods. We highlight the general applicability of CADS through simple examples in which the assumptions behind MA and global normalization methods are violated. Importantly, in each of these examples, MA and global methods may cause every null, nondifferentially 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, we show it is not possible to verify that the assumptions behind MA and global methods hold.

2. A NEW MODEL FOR NORMALIZATION

Existing normalization methods based on global adjustments or smoothing “MA-plots” require a number of assumptions that limit their general applicability, as we detail in Section 4. We propose a flexible model that allows for normalization under more general assumptions. The model we develop expresses fluorescence intensities in terms of functions of the true amount of mRNA in each biological sample.

2.1 Dye-swap removes dye bias in general

The ultimate goal is to use the proposed model to adjust the observed fluorescence intensities so that they accurately reflect the underlying levels of mRNA. As a first step toward 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 three signals to affect the relative expression values: biological signal, dye bias, and array-specific effects. With only a single observation of red divided by green for each gene, it is impossible to separate these. Figure 1 shows a picture of these three sources of signal as a function of underlying mRNA amount. The method we propose (called CADS) estimates each one of these signals and preserves only the biological signal, removing both dye bias and array effects. Dye bias is estimated from the information given by the dye-swap design, and the array effects are estimated by borrowing strength across arrays.

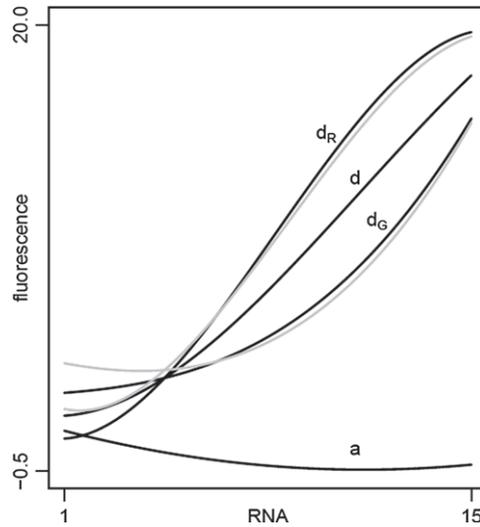


Fig. 1. A summary of the CADS model showing a hypothetical realization of a single array from the CADS model. The black curves show the individual and average dye functions as well as an example array function. The gray curves are the dye functions plus the common array effect.

It can be shown that a dye-swap separates dye bias from biological signal in a single pair of samples (Supplementary Material). 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 that 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 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 rationale is that, in principal, 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. Array effects represent systematic differences between fluorescence intensities on different arrays. These might arise, for example, when one array is inherently brighter than another. Simple averaging of two dye-swap arrays can not remove array effects in general. However, as we show below, array effects can be removed in dye-swap experiments by borrowing strength across multiple dye-swap array pairs.

2.2 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. 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_{Ti} and x_{Ci} be the true mRNA amounts for gene i in target and control samples j (say, in units that count the number of mRNA molecules). Similarly, let μ_{Ti} and μ_{Ci} be the population

average mRNA amounts for gene i in the comparison groups. Note that we use “target” and “control” only as arbitrary names to distinguish the two populations from which the biological samples come.

We assume that the observed intensities y_{ij} are noisy versions of the sum of a true underlying dye-specific fluorescence function (d_G or d_R) and an array-specific function (a_1 or a_2), with all functions applied to the x_{Tij} and x_{Cij} :

$$\begin{aligned} y_{TRij} &= d_R(x_{Tij}) + a_{1j}(x_{Tij}) + \epsilon_{TRij} \\ y_{CGij} &= d_G(x_{Cij}) + a_{1j}(x_{Cij}) + \epsilon_{CGij} \\ y_{TGij} &= d_G(x_{Tij}) + a_{2j}(x_{Tij}) + \epsilon_{TGij} \\ y_{CRIj} &= d_R(x_{Cij}) + a_{2j}(x_{Cij}) + \epsilon_{CRIj}, \end{aligned} \tag{2.1}$$

$i = 1, 2, \dots, m, j = 1, 2, \dots, n$, where the ϵ -variables 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.

Because we model the underlying mRNA amounts for each individual separately, the ϵ -variables 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. We assume that $E(d(x_{ij})) = d(\mu_i)$ for both dye functions and $E(a_j(x_{ij})) = a_j(\mu_i)$ for all array functions. This will be approximately true if x_{ij} equals μ_i plus random error and the functions d and a_j are locally linear. This allows us to connect the above models across different pairs of samples.

3. 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 (2.1) that accomplishes exactly this in expectation.

3.1 Simple dye-swap average

A simple way to remove the dye-specific effect would be to replace the individual dye functions with their average. A dye-swap does just that. The dye-swap average intensities can be written in terms of the models given in (2.1):

$$\begin{aligned} \tilde{y}_{Tij} &= \frac{1}{2}(y_{TRij} + y_{TGij}) = d(x_{Tij}) + a_j(x_{Tij}) + \epsilon_{Tij} \\ \tilde{y}_{Cij} &= \frac{1}{2}(y_{CRIj} + y_{CGij}) = d(x_{Cij}) + a_j(x_{Cij}) + \epsilon_{Cij}, \end{aligned} \tag{3.1}$$

$i = 1, 2, \dots, m, j = 1, 2, \dots, n$, where d is $[(d_R + d_G)/2]$, a_j is $[(a_{1j} + a_{2j})/2]$, and the ϵ -variables are the sums of the corresponding error terms in (2.1) divided by two. The \tilde{y} -variables are the data resulting from a simple dye-swap average. Assuming there are no array effects ($a_j = 0$), the average of the red and green channels from separate arrays can be seen as a single function d applied to the separate mean expression levels. We refer to d as the “average dye” function. Since 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. More generally, when array effects are present, they are not removed by simple dye-swap averaging.

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.

3.2 Common array dye-swap

Array effects represent the variability from array to array that is not due to biological sources of variation. 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 have adverse effects on significance calculations (Qiu *and others*, 2005).

We propose a modification of the simple dye-swap average that removes array effects. The method is described here for a direct comparison between “target” and “control”; more general experimental designs are considered in a subsequent manuscript (Dabney and Storey, 2006). Our 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. Suppose that we form n dye-swap array pairs and subtract the dye-swap averages from each other to form $\tilde{M}_{ij} = \tilde{y}_{Ti j} - \tilde{y}_{Ci j}$, $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, n$. Averaging the \tilde{M}_{ij} and taking expectations with respect to arrays gives

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

$i = 1, 2, \dots, m$. 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 approximates the scenario where the target and control were labeled with the same dye and all comparisons were made on a common array.

This CADS method can be summarized by the following algorithm:

1. Perform simple averaging of dye-swap arrays to form $\tilde{M}_{ij} = \tilde{y}_{Ti j} - \tilde{y}_{Ci j}$, $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$. 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 $\tilde{\Delta}_i = \frac{1}{n} \sum_{j=1}^n \tilde{M}_{ij}$. This is an estimate of the expected difference between target and control under the average dye function. From our assumptions, it follows that $E[\tilde{\Delta}_i] = d(\mu_{Ti}) - d(\mu_{Ci})$, $i = 1, 2, \dots, m$.
3. Fit a smoother f to the scatterplot with the $\tilde{\Delta}_i$ on the x -axis and $\tilde{M}_{ij} - \tilde{\Delta}_i$ on the y -axis to form $\tilde{a}_{ij} = f(\tilde{\Delta}_i)$, $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$. 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 , $j = 1, 2, \dots, n$.
4. Form \hat{M}_{ij} by subtracting off the estimated array functions: $\hat{M}_{ij} = \tilde{M}_{ij} - \tilde{a}_{ij}$, $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$. 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.

Note that, after Step 2, there are n dye-swap profiles that have had dye bias removed but not array effects. In Step 3, each of these profiles is compared individually to the “common array” to remove the remaining array effects. In terms of Figure 1, CADS compares each averaged dye-swap profile to an estimate of the average dye function d . A systematic difference between the two is regarded as an array effect a and is subtracted off. Figure 6 in the Supplementary Material illustrates the estimation of array effects in Step 3 of the CADS algorithm and corresponds to one of the simulations described in Section 5. Note that the array effect is especially straightforward to estimate here because of our assumption about a direct comparison between target and control. For more complicated experimental designs (e.g. multiple treatments, batch effects), the common array effect must be estimated taking all relevant variables into account (Dabney and Storey, 2006).

The CADS model (2.1) assumes an additive relationship between bias terms on the \log_2 scale. This assumption is analogous to accepted analysis of variance (ANOVA) -based models (Kerr *and others*, 2000; Kerr and Churchill, 2001) for two-channel microarrays, as well as to models for the single-channel platform (Irizarry *and others*, 2003). We also assume the same array function applies to both the red and green channels. The calibration experiments of Tseng *and others* (2001) allowed for a comparison of the array effects on both channels and give support to this assumption (data not shown). There are connections between the multi-array nature of CADS and normalization methods for single-channel microarrays. For example, Li and Wong (2001) center each array around the “median array,” the array that corresponds to the median overall intensity.

3.3 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 the CADS model, we show that the CADS estimator is unbiased for the parameters of interest $d(\mu_{Ti}) - d(\mu_{Ci})$, $i = 1, 2, \dots, m$. Since differential expression methods are almost always based on sample averages (Cui and Churchill, 2003), it is sufficient to show that the sample average of the CADS-normalized arrays is unbiased. It can be shown that

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

$i = 1, 2, \dots, m$ (Supplementary Material). It is not strictly necessary for the array functions to have expectation equal to the zero line. All that is required is that the array functions average to some non-decreasing function (Supplementary Material). 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.

4. SMOOTHING MA PLOTS

We now compare and contrast our proposed model and normalization method with an existing standard. The vast majority of normalization methods used today build on two highly influential papers (Tseng *and others*, 2001; Yang *and others*, 2002), where “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 A (\log of red times green, divided by two) on the x -axis and \log ratios M (\log of red divided by green) on the y -axis. The left plot in Figure 2 is a “self-self” MA-plot from calibration experiments (Tseng *and others*, 2001), where the pair of samples hybridized to each array are biologically equivalent. Since the intensities of the

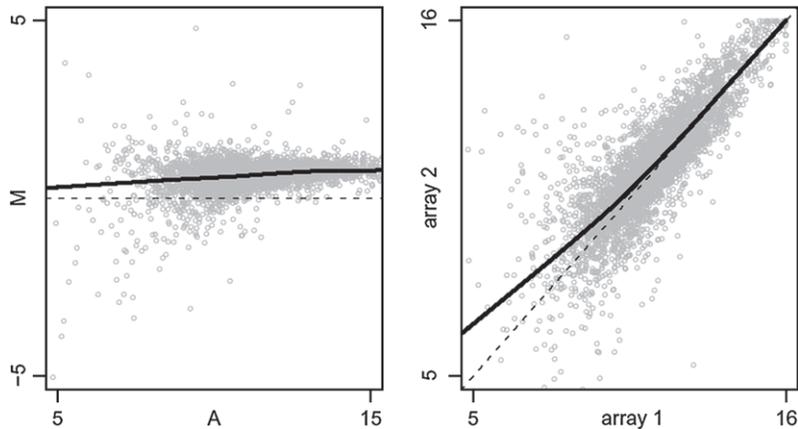


Fig. 2. Self-self calibration experiments (Tseng *and others*, 2001). Left plot is an MA-plot showing dye differences on the same sample. Without dye bias, the points would be centered along dashed zero line. Right plot is a scatterplot of the same dye applied to different arrays. Without array effects, the points would be centered along dashed line of equality.

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 attempt to remove bias by smoothing away MA trends. A smooth curve 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 the data or some subset of control genes. While Figure 2 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. 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. Our analysis below indicates that this procedure can create artificial differential expression, destroy true differential expression, or both.

4.1 *The assumptions behind MA methods*

MA methods were justified by Yang *and others* (2002) 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/downregulated genes.” We show that the fundamental assumption necessary to justify MA methods is that all systematic MA trends are due exclusively to bias (Supplementary Material). Necessary conditions to satisfy this assumption are (Supplementary Material)

1. 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, Assumptions 2 and 3 were not. As the original authors correctly stated, MA methods are only guaranteed to work properly when their assumptions are met. 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 data sets 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, “the expression profiles in biological samples are (frequently) more divergent in nature than in the examples investigated (here)” (Yang *and others*, 2002).

4.2 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, even in the absence of dye bias. Smoothing the MA curve pushes the cloud of data down to the zero line. Overexpressed genes will appear less extreme, underexpressed genes will appear more extreme, and all equivalently expressed genes will now be artificially differentially expressed. Smoothing the MA-plot reduces signal in some genes and creates bias in others. Note that asymmetry is common (Tusher *and others*, 2001; Hedenfalk *and others*, 2001); in fact, only in specialized circumstances would it be guaranteed that differential expression between two biological groups of interest be perfectly symmetric.

4.3 Intensity-dependent differential expression

Suppose 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 may 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 increasing from low- to high-abundance genes, even in the absence of dye bias or other systematic effects. Smoothing the MA-plot will again destroy signal in some genes and create bias in others. In particular, it will again create an apparent signal in all the null, equivalently expressed genes wherever the MA smoother does not pass through the zero line.

4.4 Unequal variation in expression means

Unequal variances between comparison groups occur, for example, when comparing immune challenged to normal cells (Storey *and others*, 2005). 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. This scenario also produces an MA trend increasing from low- to high-abundance genes, even in the absence of bias. Smoothing the MA-plot again destroys biological signal.

4.5 Real examples where assumptions are violated

Figure 3 shows a real microarray experiment in which all three MA assumptions are violated; a reference design was used in this example, so we know any MA trends are not due to dye bias. Arrays were obtained from sporadic, BRCA1 mutation-positive, and BRCA2 mutation-positive breast cancer tumors (Hedenfalk *and others*, 2001). The left plot shows the \log_2 fold-change among 3220 genes between BRCA1 and

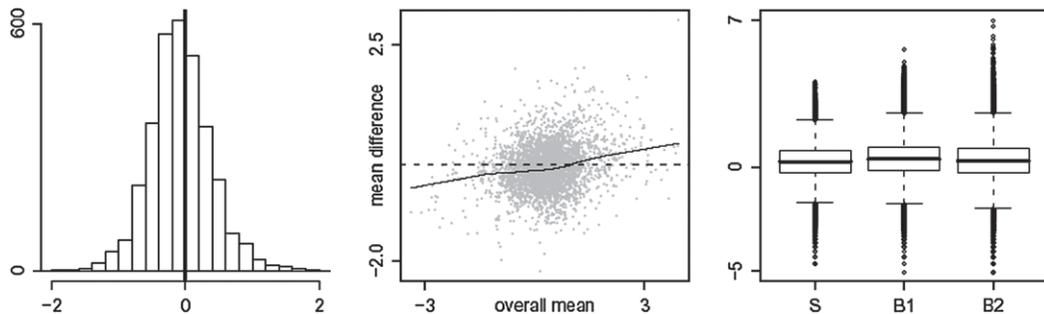


Fig. 3. Examples of asymmetric differential expression, intensity-dependent differential expression, and differences in variation of expression across samples in the BRCA data (Hedenfalk *and others*, 2001). The left plot shows asymmetry among average \log_2 fold-change expression between BRCA1 and BRCA2 mutation-positive tumors. The center plot shows average overall abundance versus \log_2 fold-change expression among BRCA1 and BRCA2 mutation-positive tumors. The right plot shows boxplots of average \log_2 expression from tumors among Sporadic (S), BRCA1 (B1), and BRCA2 (B2) individuals.

BRCA2 arrays, where there is clear asymmetric differential expression, violating Assumption 1. The center plot shows there is a relationship between overall expression abundance and differential expression, violating Assumption 2. The right plot shows boxplots of the expression data among the sporadic, BRCA1, and BRCA2 groups. It can be seen that the variation of expression across genes increases from sporadic to BRCA1 to BRCA2, violating Assumption 3. The increase in standard deviation across genes is about 11% moving from one group to the next.

4.6 Other normalization methods

In addition to MA methods, alternative normalization methods are surveyed in Quackenbush (2002). Global normalization methods subtract some constant from all log expression ratios on an array (Kerr *and others*, 2000). These are rarely used since systematic effects tend to be more complicated than constant shifts at the array level. Also, since a global procedure is equivalent to smoothing an MA plot with a flat curve, global methods suffer from the same limitations of MA methods discussed above. Others have proposed procedures based on observed deviations in “housekeeping” genes (genes whose expression levels are known to remain constant) (Chen *and others*, 1997). Still others attempt to either create exogenous housekeeping genes (Yang *and others*, 2002; Benes and Muckenthaler, 2003) or identify and/or synthesize genes on an existing array that behave like housekeeping genes (Zien *and others*, 2001; Tseng *and others*, 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.

4.7 The difference between CADS and MA methods

MA methods assume that every MA-plot should be centered on the zero line. The idea behind CADS is to estimate the true target-to-control difference by averaging out array effects and centering each array around the result. The most important difference between CADS and MA methods is that CADS removes bias due to dye and array effects under general assumptions, preserving differential expression relationships in expectation, while MA methods do not. There are also distinct operational differences between CADS and MA methods. In a dye-swap experiment, MA methods are typically applied by first smoothing

the MA-plot for each array, then averaging the dye-swaps. The first step of CADS is to average the dye-swaps, removing dye bias. The second step uses all arrays to form the “common array,” making CADS a multi-array method. Because CADS is a multi-array method, persistent trends that are not due to bias can be identified and retained.

5. SIMULATED EXAMPLES

We illustrate CADS on three simulated examples. Simulation 1 has asymmetric differential expression. Simulation 2 has intensity-dependent differential expression, where differential expression increases with increasing abundance. Simulation 3 has unequal variances in expression measurements between comparison groups. In each of 30 simulations, we generated expression measurements for 3000 genes in five target and five control samples. For each target/control sample pair, a dye-swap was carried out. Of the 3000 genes, 35% are differentially expressed. Nonlinear dye functions were incorporated, as were random array functions generated from a distribution with mean value equal to the zero line. All summaries reported below were averaged over the 30 simulations. For full details, see the (Supplementary Material). We note that in subsequent work (Dabney and Storey, 2006), we develop CADS further and use a real

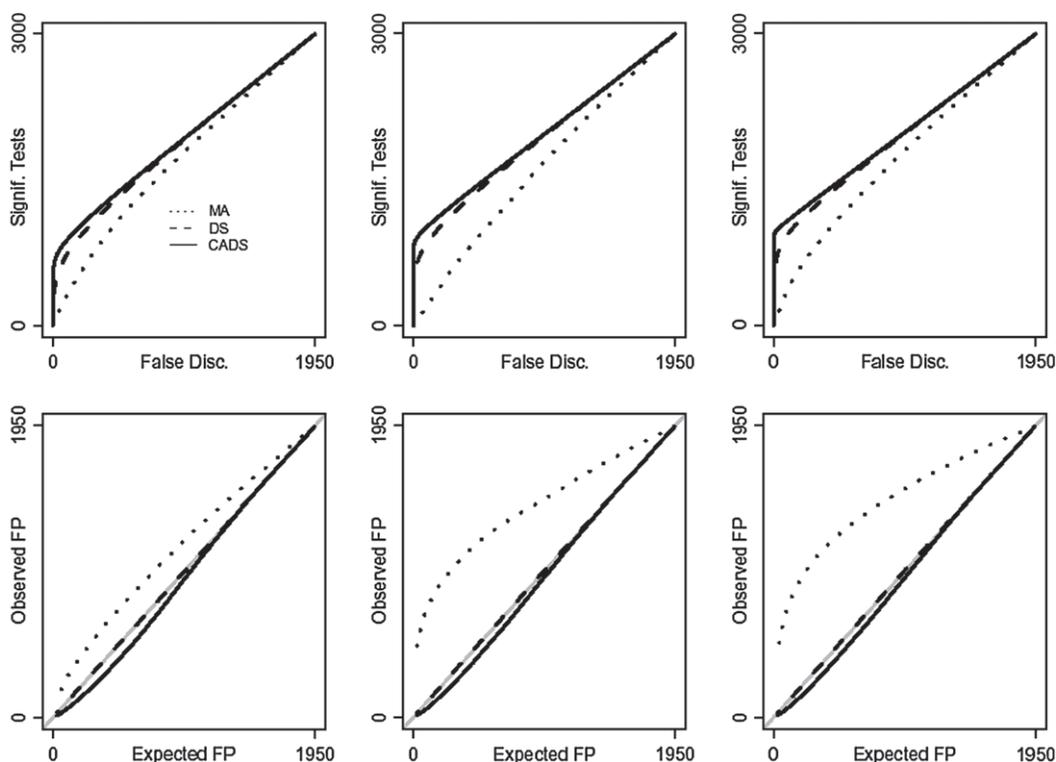


Fig. 4. A comparison of MA, simple dye-swap average, and CADS normalization methods on the three simulated examples. Example number increases from left to right. The top row is a comparison based on false discoveries versus the number of significant tests. The bottom row is a comparison of the number 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.

data set on prostate development, provided by the Pete Nelson laboratory at the Fred Hutchinson Cancer Research Center.

To compare MA-smoothing and CADS in the simulations, 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 4. 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. 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 and Tibshirani, 2003).

6. DISCUSSION

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 *and others*, 2002), although (based on a PubMed search) the MSP does not appear to have been used much in practice. The "rank-invariant" MA method (Tseng *and others*, 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 Simulations 1 and 2 of Section 5, 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.

CADS can easily be applied when more than two groups are compared. A loop design with more than two nodes is required, but the basic approach behind the method remains unchanged (Supplementary Material). Also, CADS can be more generally formulated as an ANOVA model in which factor terms have been replaced with functions of mRNA amount. Framed as an ANOVA model, it is simple to incorporate additional sources of bias or covariates. Furthermore, the CADS "functional" ANOVA model supports a more efficient dye-swap design in which only a single array is used for each sample pair. These extensions are developed in subsequent work (Dabney and Storey, 2006). These results together highlight important limitations in existing normalization methods and provide alternatives that are simple, flexible, and accurate under general assumptions.

ACKNOWLEDGMENTS

This research was supported in part by the Cancer-Epidemiology and Biostatistics Training Grant 5T32CA009168-29 and National Institutes of Health grant R01 HG002913. *Conflict of Interest:* None declared.

REFERENCES

- BENES, V. AND MUCKENTHALER, M. (2003). Standardization of protocols in cDNA microarray analysis. *Trends in Biochemical Sciences* **28**, 244–9.
- CHEN, Y., DOUGHERTY, E. AND BITTNER, M. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* **24**, 364–74.
- CUI, X. AND CHURCHILL, G. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* **4**, 210.

- DABNEY, A. AND STOREY, J. (2006). Functional ANOVA normalization of two-channel microarrays. *Technical Report 285*. Seattle, WA: Department of Biostatistics, University of Washington.
- FANG, Y., BRASS, A., HOYLE, D., HAYES, A., BASHEIN, A., OLIVER, S., WADDINGTON, D. AND RATTRAY, M. (2003). A model-based analysis of microarray experimental error and normalisation. *Nucleic Acids Research* **31**, e96.
- HEDENFALK, I., DUGGAN, D., CHEN, Y., RADMACHER, M., BITTNER, M., SIMON, R., MELTZER, P., GUSTERSON, B., ESTELLER, M., KALLIONIEMI, O. *and others* (2001). Gene expression profiles in hereditary breast cancer. *New England Journal of Medicine* **344**, 539–48.
- HUGHES, T., MAO, M., JONES, A., BURCHARD, J., MARTON, M., SHANNON, K., LEFKOWITZ, S., ZIMAN, M., SCHELTER, J., MEYER, M. *and others* (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology* **19**, 342–7.
- IRIZARRY, R., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y., ANTONELLIS, K., SCHERF, U. AND SPEED, T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–64.
- KERR, M. AND CHURCHILL, G. (2001). Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201.
- KERR, M., MARTIN, M. AND CHURCHILL, G. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–37.
- LI, C. AND WONG, W. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology* **2**, research0032.
- QIU, X., BROOKS, A. I., KLEBANOV, L. AND 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. AND BROWN, P. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–70.
- STOREY, J. D. AND TIBSHIRANI, R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–5.
- STOREY, J. D., XIAO, W., LEEK, J. T., TOMPKINS, R. G. AND DAVIS, R. W. (2005). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences* **102**, 12837–42.
- TSENG, G., OH, M., ROHLIN, L., LIAO, J. AND WONG, W. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research* **29**, 2540–57.
- TUSHER, V., TIBSHIRANI, R. AND CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98**, 5116–24.
- YANG, Y., DUDOIT, S., LUU, P., LIN, D., PENG, V., NGAI, J. AND SPEED, T. (2002). 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. AND SPEED, T. P. (2001). Normalization for cDNA microarray. data. In: Bittner, M. L., Chen, Y., Dorsel, A. N. and Dougherty, E. R. (editors), *Microarrays: Optical Technologies and Informatics*, Volume 4266, Proceedings of SPIE. San Jose, CA. pp 141–52.
- ZIEN, A., AIGNER, T., ZIMMER, R. AND LENGAUER, T. (2001). Centralization: a new method for the normalization of gene expression data. *Bioinformatica* **1**, 1–9.

Supporting Appendix:
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

April 2006

Contents

1	The Model and Assumptions Behind MA Methods	2
2	Simulation Details	3
3	Dye Swap Removes Dye Bias in General	4
4	CADS Preserves Differential Expression Relationships	5
5	CADS Algorithm Illustrated	7
6	CADS Induces Negligible Dependence Between Arrays	8
7	Extensions to Other Experimental Designs	8

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 \log_2 intensities y_{Ci} in the other, $i = 1, 2, \dots, m$. MA methods form $M_i = y_{Ti} - y_{Ci}$ and $A_i = (y_{Ti} + y_{Ci})/2$, $i = 1, 2, \dots, m$, fit a smooth curve to the scatterplot with the A_i on the x -axis and the M_i on the y -axis, and subtract off the fitted values, forcing the points to be centered along the line $M = 0$. We now show that MA methods imply the following assumptions:

1. 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

$$E(M|A) = b(A) + s(A), \tag{3}$$

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 $E(M|A) = 0$, 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 (3) as,

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

$i = 1, 2, \dots, m$, 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

$$\begin{aligned} \text{Var}(\mathbf{y}_T) &= \frac{1}{m} \sum_{i=1}^m \left(E(y_{Ti}) - \frac{1}{m} \sum_{j=1}^m E(y_{Tj}) \right)^2, \quad \text{and} \\ \text{Var}(\mathbf{y}_C) &= \frac{1}{m} \sum_{i=1}^m \left(E(y_{Ci}) - \frac{1}{m} \sum_{j=1}^m E(y_{Cj}) \right)^2. \end{aligned}$$

Then $\text{Var}(\mathbf{y}_T)$ describes the variation of expression across genes in targets, and $\text{Var}(\mathbf{y}_C)$ is the analogous quantity for controls. Using similar notation, an MA trend occurs when there is covari-

ation between the $y_{Ti} - y_{Ci}$ and the $y_{Ti} + y_{Ci}$:

$$\begin{aligned} \text{Cov}(\mathbf{M}, \mathbf{A}) &= \text{Cov}(\mathbf{y}_T - \mathbf{y}_C, \frac{1}{2}(\mathbf{y}_T + \mathbf{y}_C)) \\ &= \frac{1}{m} \sum_{i=1}^m \left(\text{E}(y_{Ti}) - \frac{1}{m} \sum_{j=1}^m \text{E}(y_{Tj}) \right) \left(\text{E}(y_{Ci}) - \frac{1}{m} \sum_{j=1}^m \text{E}(y_{Cj}) \right). \end{aligned} \quad (4)$$

In the absence of bias, MA methods assume that $\text{Cov}(\mathbf{M}, \mathbf{A}) = 0$. However, it follows from (4) that

$$\text{Cov}(\mathbf{M}, \mathbf{A}) = \frac{1}{2} (\text{Var}(\mathbf{y}_T) - \text{Var}(\mathbf{y}_C)).$$

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

For each of 30 simulations, expression measurements were generated for 3000 genes in 5 target samples and 5 control samples. For each target/control sample pair, a dye-swap was carried out. The dye functions used are nonlinear functions, as shown in the middle panel of Figure 5. The array functions are also nonlinear random functions generated from a distribution with expected value equal to the zero line, with examples shown in the right panel of Figure 5. All summary statistics reported in the main document were averaged across the 30 simulations.

Example 1: Asymmetric Differential Expression. We generated 3000 control means $\mu_{C1}, \mu_{C2}, \dots, \mu_{C3000}$ from the $N(0, 5^2)$ distribution. We then randomly chose 25% of the genes to be over-expressed and 10% to be under-expressed. Target means were either control means plus 1, minus 0.25, or unchanged, depending on which group they were in. Sample variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_{3000}^2$ were generated from the $\chi^2(1)$ distribution. Finally, target and control observations for gene i were generated according to the CADS model, with dye and array functions described in Figure 5. The random error terms were drawn from the $N(0, \sigma_i^2)$ distribution.

*** Figure 5 about here. ***

Example 2: Intensity-Dependent Differential Expression. We generated 3000 control means as in Example 1. We randomly chose 35% 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 $\mu_{Ti} = -0.1 + 0.3\mu_{Ci} + 2.7\mu_{Ci}^2 - 1.8\mu_{Ci}^3$. The left panel

of Figure 5 shows the simulated relationship between mean expression levels. Target and control observations for each gene were generated according to the CADS model, with random error terms drawn from normal distributions, as in Example 1.

Example 3: Unequal Variation in Expression Means. We generated 3000 control means as in Examples 1 and 2. We randomly chose 35% of the genes to be differentially expressed. Control means were generated from the $N(0, 5^2)$ distribution, and the target means were then generated from the $N(\mu_{Ci}, 5^2)$ distributions. Target and control observations for each gene were generated using Normal distributions, as in Examples 1 and 2.

3 Dye Swap Removes Dye Bias in General

It can be shown that a dye swap parses dye bias from biological signal when only a single pair of biological samples is obtained. 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

$$\begin{aligned}
 y_{Ti} &= d_R(x_{Ti}) + \epsilon_{Ti}, \\
 y_{Ci} &= d_G(x_{Ci}) + \epsilon_{Ci} \\
 &= d_R(x_{Ci}) + [d_G(x_{Ci}) - d_R(x_{Ci})] + \epsilon_{Ci}.
 \end{aligned}
 \tag{5}$$

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 $\mu_{Ti} = \mu_{Ci}$, 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 μ_{Ti} and μ_{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 measurements) 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 $d(\mu_{Ti}) - d(\mu_{Ci})$, $i = 1, 2, \dots, m$. 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 , $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$. 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, \dots, a_n independent random curves representing array effects, let $a_0 = Ea_j$, $j = 1, 2, \dots, n$. Then, $E\left(\frac{1}{n} \sum_{j=1}^n \hat{M}_{ij}\right) = d(\mu_{Ti}) - d(\mu_{Ci}) + a_0(\mu_{Ti}) - a_0(\mu_{Ci})$, $i = 1, 2, \dots, m$.

Theorem 1 is proved below. Two results follow directly.

Corollary 1 *If a_0 is nondecreasing, then*

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

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

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.

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

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

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

$$\begin{aligned} \mathbf{y}_{TRj} &= \mathbf{d}_R(\mathbf{x}_{Tj}) + \mathbf{a}_{1j}(\mathbf{x}_{Tj}) + \boldsymbol{\epsilon}_{TRj}, \\ \mathbf{y}_{CGj} &= \mathbf{d}_G(\mathbf{x}_{Cj}) + \mathbf{a}_{1j}(\mathbf{x}_{Cj}) + \boldsymbol{\epsilon}_{CGj}, \\ \mathbf{y}_{TGj} &= \mathbf{d}_G(\mathbf{x}_{Tj}) + \mathbf{a}_{2j}(\mathbf{x}_{Tj}) + \boldsymbol{\epsilon}_{TGj}, \\ \mathbf{y}_{CRj} &= \mathbf{d}_R(\mathbf{x}_{Cj}) + \mathbf{a}_{2j}(\mathbf{x}_{Cj}) + \boldsymbol{\epsilon}_{CRj}, \end{aligned}$$

$j = 1, 2, \dots, n$. Each function is now vector-valued, so that, for example, $\mathbf{d}_R(\mathbf{x}_{Tj})$ has i th component $d_R(x_{Tij})$, $i = 1, 2, \dots, m$. Averaging the two dye swap arrays gives

$$\begin{aligned} \tilde{\mathbf{y}}_{Tj} &= \frac{1}{2} \left(\mathbf{y}_{TRj} + \mathbf{y}_{TGj} \right) = \mathbf{d}(\mathbf{x}_{Tj}) + \mathbf{a}_j(\mathbf{x}_{Tj}) + \boldsymbol{\epsilon}_{Tj}, \\ \tilde{\mathbf{y}}_{Cj} &= \frac{1}{2} \left(\mathbf{y}_{CRj} + \mathbf{y}_{CGj} \right) = \mathbf{d}(\mathbf{x}_{Cj}) + \mathbf{a}_j(\mathbf{x}_{Cj}) + \boldsymbol{\epsilon}_{Cj}. \end{aligned}$$

Finally, subtracting these two quantities is equivalent to normalizing by simple dye-swap averaging, producing $\tilde{\mathbf{M}}_j = \tilde{\mathbf{y}}_{Tj} - \tilde{\mathbf{y}}_{Cj}$. For each array j , CADS fits the model $\tilde{\mathbf{M}}_j - \tilde{\boldsymbol{\Delta}} = \mathbf{f}_j(\tilde{\boldsymbol{\Delta}}) + \boldsymbol{\epsilon}_j$,

where f_j is a smooth function, and $\tilde{\Delta} = \frac{1}{n} \sum_{i=1}^m \tilde{M}_j$. To fix ideas, let us assume that f_j can be represented by $B_{\Delta} \beta_j$, where B_{Δ} is a known basis matrix evaluated at $d(\mu_T) - d(\mu_C)$; note that this implies $a_j(x_{Tj}) - a_j(x_{Cj}) = B_{\Delta} \beta_j$. Any polynomial or spline function could be represented this way. We then use least-squares to estimate

$$\tilde{a}_j = \hat{f}_j(\tilde{\Delta}) = B_{\tilde{\Delta}} (B_{\tilde{\Delta}}^T B_{\tilde{\Delta}})^{-1} B_{\tilde{\Delta}}^T (\tilde{M}_j - \tilde{\Delta}) = H_{\tilde{\Delta}} (\tilde{M}_j - \tilde{\Delta}),$$

where $H_{\tilde{\Delta}} = B_{\tilde{\Delta}} (B_{\tilde{\Delta}}^T B_{\tilde{\Delta}})^{-1} B_{\tilde{\Delta}}^T$ is a ‘‘hat matrix.’’ Thus, CADs normalizes array j to form the estimate

$$\hat{M}_j = \tilde{M}_j - \hat{f}_j(\tilde{\Delta}) = H_{\tilde{\Delta}} (\tilde{M}_j - \tilde{\Delta}),$$

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

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

since $\sum_{j=1}^n (\tilde{M}_j - \tilde{\Delta}) = 0$. \square

We note from the proof that

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

This fact makes the proof of Theorem 1 even simpler, since $E(\tilde{M}_{ij}) = d(\mu_{Ti}) - d(\mu_{Ci}) + a_0(\mu_{Ti}) - a_0(\mu_{Ci})$, $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$.

5 CADs Algorithm Illustrated

In step 2 of the CADs algorithm, dye-swap arrays are averaged. Thus, beginning with n sample pairs and $2n$ arrays (including the dye-swap arrays), step 2 averages the differences between targets and controls within each sample pair, resulting in n array profiles. Since we have averaged across the two dye configurations, we have removed dye bias, but array effects are still intact. Step 3 then forms the ‘‘common array’’ by averaging across the n profiles and compares the n individual profiles

to this average. Figure 6 illustrates this step on one simulated dataset from our simulation example 2. On the x -axis is the $\tilde{\Delta}$ variable, representing the common array. On the y -axis is one example of $\tilde{M}_j - \tilde{\Delta}$. The solid curve is a natural cubic spline with 5 degrees of freedom. As expected, the shape of the curve is analogous to the array functions shown in Figure 5 that were used to generate the data. In step 4, we subtract off the fitted values of the smoother, reshaping the scatterplot to be centered along the dashed horizontal zero line.

*** Figure 6 about here. ***

6 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.

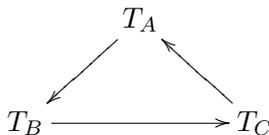
7 Extensions to Other Experimental Designs

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 label 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.



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 $\tilde{Y}_A = \frac{1}{2} [Y_{AR} + Y_{AG}]$ and $\tilde{Y}_B = \frac{1}{2} [Y_{BR} + Y_{BG}]$, 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.

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.

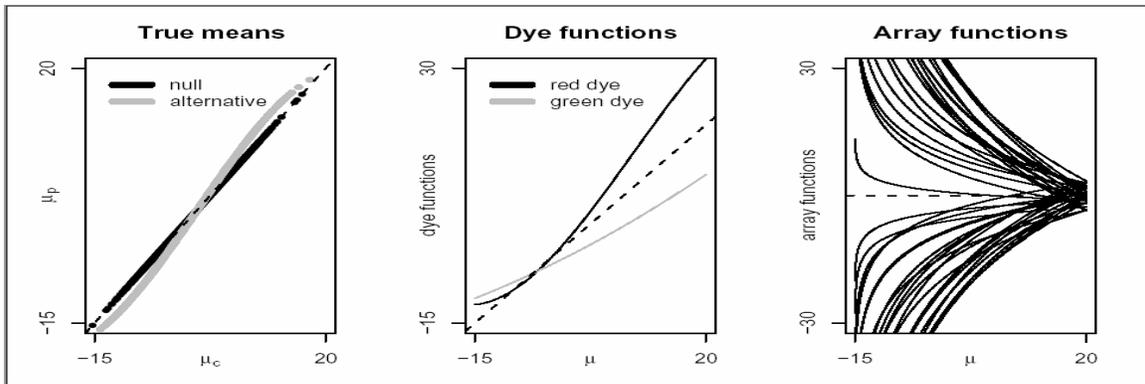


Figure 5: A summary of the simulation model used in the main paper. The left panel shows the simulated relationship between expression means in example 2. The middle panel shows the dye functions used. The right panel shows example array functions, drawn from a distribution with expected value equal to the zero line.

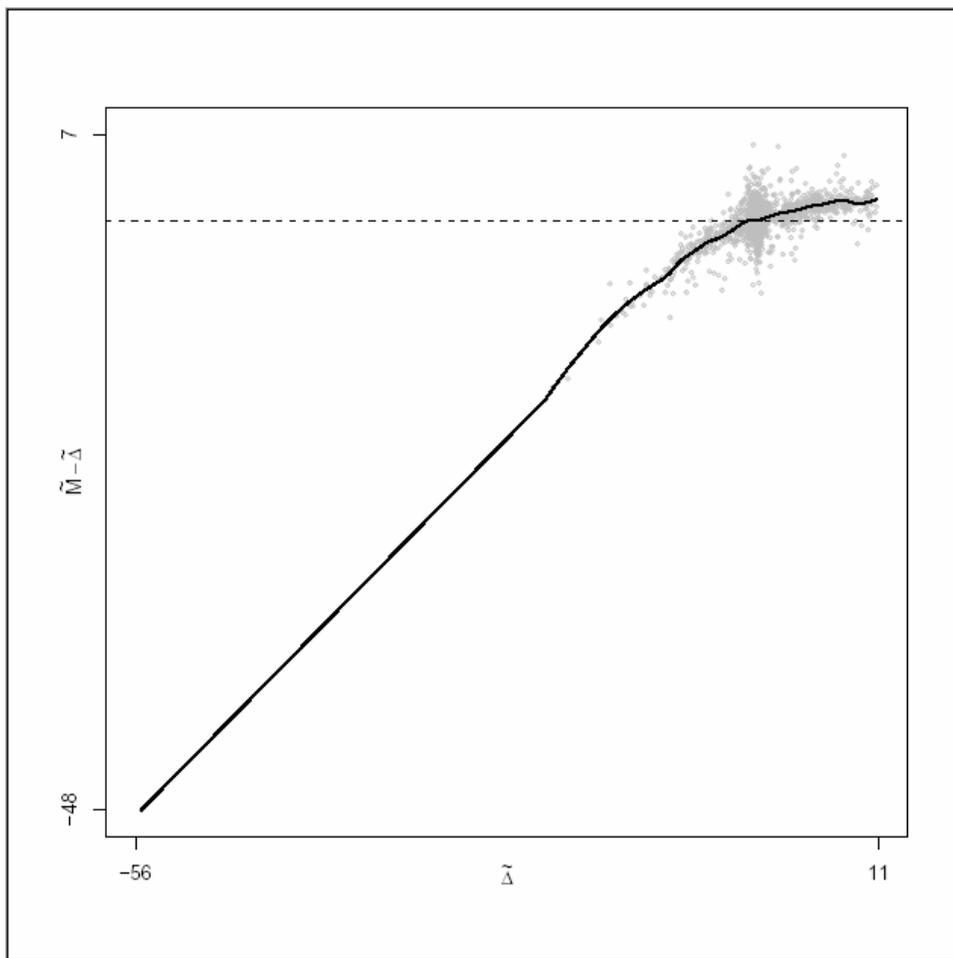


Figure 6: An illustration of step 3 of the CADS algorithm. The common array is on the x -axis, and the deviation from this common array for one simulated example is on the y -axis. To remove array effects, we subtract off the fitted values of the solid smoother curve.