# Supplement 2 (of 2) to: Characterizing dye bias in microarray experiments

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

**Proof that the existence of gene-and-sample specific dye bias implies that assumptions that cannot be verified from the microarray data alone are required for comparing samples**

## 1 Summary

In a single sentence the argument is that the model with three-way (dye by gene by sample) interactions require constraint equations for the interactions, but that these constraint equations implicitly make unverifiable assumptions about the nature of the gene-and-sample-specific dye bias, which thus makes valid statistical inference impossible.

## 2 Preliminary discussion

Dye-swapping individual arrays seems to be an intuitively appealing method of removing gene-and-sample-specific dye bias from estimates. And, if one fits and ANOVA model to the data with a gene-and-sample-specific dye bias effect in it, there will be no trouble estimating parameters and reaching conclusions. But the ANOVA approach relies on constraint equations to make parameters identifiable. Usually these constraint equations

[*] Biometric Research Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

[†] Advanced Technology Center, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

are of no concern, but in the case of gene-and-sample-specific dye bias, we argue, they are of concern. The reason the constraint equations matter is that, unlike in the gene-specific dye bias case, the estimate of gene expression differences between the samples will be different depending on whether Cy3, Cy5 or the average of the two is the scale on which comparisons are made. Intuitively, this is because, in switching from Cy3 scale to a Cy5 scale, one adds different amounts to each sample (by definition of sample-specific dye bias), and hence one reaches different conclusions on the Cy3 scale than on the Cy5 scale.

## 3 Notation and general model

Fix a particular gene. For sample $s$, on array $a$, tagged with dye $d$, on replicate $r$, let $Y_{sadr}$ be the normalized, background-adjusted, log-intensity with expectation:

$$E\left[Y_{sadr}\right] \quad = \quad \mu_{sad} = \mu + S_s + A_a + D_d + SD_{sd}.$$

Here $\mu$ is the grand mean, $S_s$ is the sample effect, $A_a$ is the “spot” effect, $D_d$ is the gene-specific dye bias, and $SD_{sd}$ is the gene-and-sample-specific dye bias.

## 4 Two samples in a single dye system

First, suppose we have two samples and one dye. In this case, the “spot” effects are no longer estimable, but suppose they are known.

$$\begin{aligned}
\mu_{1a_1d} &= \mu + S_1 + A_{a_1} + D_d + SD_{1d} \\
\mu_{2a_2d} &= \mu + S_2 + A_{a_2} + D_d + SD_{2d} \\
\mu_{1a_1d} - \mu_{2a_2d} &= A_{a_1} - A_{a_2} + S_1 - S_2 + SD_{1d} - SD_{2d}
\end{aligned}$$

Note that the last equation reveals a problem with the model. The sample specific dye bias is not eliminated from the difference. More generally, the model is not identifiable. To see this more clearly, note that we can rewrite the model,

$$\begin{aligned}\mu_{1ad} &= \mu + (S_1 + SD_{1d}) + A_a + D_d \\ \mu_{2ad} &= \mu + (S_2 + SD_{2d}) + A_a + D_d\end{aligned}$$

Then note that if there is only one dye, $S_1$ and $SD_{1d}$ are always observed together. And similarly, $S_2$ and $SD_{2d}$ are always observed together. And they do not cancel out of differences. So we can only hope to ever estimate the sums, $S_1 + SD_{1d}$ and $S_2 + SD_{2d}$. There is no way from observed data to estimate the quantities $S_1$ and $S_2$ without further model assumptions.

A reasonable assumption we might make in the case of just one dye is that, for the single dye $d = 1$, we have $SD_{11} = SD_{21} = 0$. This assumes that the gene does not exhibit sample-specific dye bias. More generally, one might assume that if a single dye is used, variation in the (normalized, background adjusted) log-intensities across arrays for a gene tagged with that dye accurately tracks variation of the underlying log-gene-expression. This is the assumption typically made when analyzing single dye systems, such as Affymetrix.

# 5 Two samples in a two-dye system

Now expand to a two-dye system. The mean equations are:

$$\begin{aligned}\mu_{1a1} &= \mu + S_1 + A_a + D_1 + SD_{11} \\ \mu_{1a2} &= \mu + S_1 + A_a + D_2 + SD_{12} \\ \mu_{2a1} &= \mu + S_2 + A_a + D_1 + SD_{21} \\ \mu_{2a2} &= \mu + S_2 + A_a + D_2 + SD_{22}\end{aligned}$$

This is the typical analysis of variance two-way layout. Note that the model is highly over-parameterized and not identifiable. Constraint equations for the parameters reduce the model dimension and result in an identifiable model.

Our argument is that constraints on the interaction terms necessitate assumptions about the relation between the underlying gene expression and the intensities that are unverifiable.

To simplify the situation, we will assume that $\mu$ and $A_a$ are all known.

First, consider the constraint equation on the interaction term we gave earlier, viz. $SD_{11} = SD_{21} = 0$. The constraints assumes that dye 1 accurately tracks gene expression variation but dye 2 may not. The usual model constraints on the samples and dyes are: $S_1 + S_2 = 0$, $D_1 + D_2 = 0$. Adding this to the previous two constraints on the interactions, gives a total of four constraints. There are now 8 unknowns $(S_1, S_2, D_1, D_2, SD_{11}, SD_{12}, SD_{21}, SD_{22})$, 4 constraints, and 4 equations. So there is a single unique solution. Call this model $M_1$:

$$\begin{aligned}
\mu_{1a1} &= \mu + S_1 + A_a + D_1 \\
\mu_{1a2} &= \mu + S_1 + A_a + D_2 + SD_{12} \\
\mu_{2a1} &= \mu + S_2 + A_a + D_1 \\
\mu_{2a2} &= \mu + S_2 + A_a + D_2 + SD_{22}
\end{aligned}$$

Now consider instead the constraint equations $0 = SD_{12} = SD_{22}$, and the same model constraints $S_1 + S_2 = 0$, $D_1 + D_2 = 0$. Here we assume that dye 2, instead of dye 1, accurately tracks with gene expression. Call this model $M_2$.

$$\begin{aligned}
\mu_{1a1} &= \mu + S_1 + A_a + D_1 + SD_{11} \\
\mu_{1a2} &= \mu + S_1 + A_a + D_2 \\
\mu_{2a1} &= \mu + S_2 + A_a + D_1 + SD_{21} \\
\mu_{2a2} &= \mu + S_2 + A_a + D_2
\end{aligned}$$

Since the $A_a$ and $\mu$ are known, the notation can be simplified further. Model $M_1$ becomes

$$k_{11} = \mu_{1a1} - \mu - A_a \quad = \quad S_1 + D_1$$

$$
\begin{aligned}
k_{12} = \mu_{1a2} - \mu - A_a &= S_1 + D_2 + SD_{12} \\
k_{21} = \mu_{2a1} - \mu - A_a &= S_2 + D_1 \\
k_{22} = \mu_{2a2} - \mu - A_a &= S_2 + D_2 + SD_{22}.
\end{aligned}
$$

Solving the equations:

$$
\begin{aligned}
S_1 &= \frac{1}{2}\left[k_{11} - k_{21}\right] \\
S_2 &= \frac{1}{2}\left[k_{21} - k_{11}\right] \\
D_1 &= \frac{1}{2}\left[k_{11} + k_{21}\right] \\
D_2 &= -\frac{1}{2}\left[k_{11} + k_{21}\right] \\
SD_{12} &= k_{12} + k_{21} \\
SD_{22} &= k_{11} + k_{22}
\end{aligned}
$$

Model $M_2$ becomes

$$
\begin{aligned}
k_{1a1} = \mu_{1a1} - \mu - A_a &= S_1 + D_1 + SD_{11} \\
k_{1a2} = \mu_{1a2} - \mu - A_a &= S_1 + D_2 \\
k_{2a1} = \mu_{2a1} - \mu - A_a &= S_2 + D_1 + SD_{21} \\
k_{2a2} = \mu_{2a2} - \mu - A_a &= S_2 + D_2.
\end{aligned}
$$

Solving the equations:

$$
\begin{aligned}
S_1 &= \frac{1}{2}\left[k_{12} - k_{22}\right] \\
S_2 &= \frac{1}{2}\left[k_{22} - k_{12}\right]
\end{aligned}
$$

$$\begin{aligned}
D_2 &= \frac{1}{2}\left[k_{12} + k_{22}\right] \\
D_1 &= -\frac{1}{2}\left[k_{12} + k_{22}\right] \\
SD_{11} &= k_{11} + k_{22} \\
SD_{21} &= k_{12} + k_{21}
\end{aligned}$$

Finally, noting that under model $M_1$ we have $S_1 - S_2 = k_{11} - k_{21}$, while under model $M_2$ we have $S_1 - S_2 = k_{12} - k_{22}$, we can see that the models are not equivalent.

Similarly, if instead of assuming one dye tracked best with gene expression, we assumed that the average over the two dyes tracked best with gene expression, then this would give rise to a different set of equations and a different set of solutions.

Therefore, the estimates of the differences between the samples will depend on the constraints used in the model. But there is no *a priori* reason to pick one set of constraints because we do not know the true relation between the underlying gene expression and the fluorescent intensity under each dye.