Hindawi Publishing Corporation
Comparative and Functional Genomics
Volume 2009, Article ID 950171, 11 pages
doi:10.1155/2009/950171

Research Article

Statistical Analysis of Microarray Data with Replicated Spots: A Case Study with *Synechococcus* WH8102

E. V. Thomas,[1] K. H. Phillippy,[2] B. Brahamsha,[3] D. M. Haaland,[4] J. A. Timlin,[4] L. D. H. Elbourne,[5] B. Palenik,[3] and I. T. Paulsen[5]

[1] *Department of Independent Surveillance Assessment and Statistics, Sandia National Laboratories, Albuquerque, NM 87185-0829, USA*
[2] *National Center for Biotechnology Information, National Library of Medicine, National Institute of Health, Bethesda, MD 20894, USA*
[3] *Scripps Institution of Oceanography, University of California at San Diego, La Jolla, CA 92093-0202, USA*
[4] *Department of Biomolecular Analysis and Imaging, Sandia National Laboratories, Albuquerque, NM 87185-0895, USA*
[5] *Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW 2109, Australia*

Correspondence should be addressed to E. V. Thomas, evthoma@sandia.gov

Received 25 September 2008; Revised 15 January 2009; Accepted 9 February 2009

Recommended by Antoine Danchin

Until recently microarray experiments often involved relatively few arrays with only a single representation of each gene on each array. A complete genome microarray with multiple spots per gene (spread out spatially across the array) was developed in order to compare the gene expression of a marine cyanobacterium and a knockout mutant strain in a defined artificial seawater medium. Statistical methods were developed for analysis in the special situation of this case study where there is gene replication within an array and where relatively few arrays are used, which can be the case with current array technology. Due in part to the replication within an array, it was possible to detect very small changes in the levels of expression between the wild type and mutant strains. One interesting biological outcome of this experiment is the indication of the extent to which the phosphorus regulatory system of this cyanobacterium affects the expression of multiple genes beyond those strictly involved in phosphorus acquisition.

1. Introduction

Microarray experiments provide high-throughput gene expression data required for elucidating networks and pathways occurring in organisms and for validating models derived from other experimental data. The quality of models and inference derived from microarray experiments obviously depends on the quality of the microarray data. For example, predictive models are hard to develop or validate if microarray data have high false positive and/or false negative rates for identifying differential gene expression. Thus, it is important to make results from microarray experiments as reproducible and reliable as possible. In addition, it is important to institute a process to monitor, assess, and ultimately improve the quality of the microarray data.

A number of researchers have identified a variety of sources of variation which affect the reproducibility of microarray data. Statistically designed microarray experiments that include replication have been critical to understanding, assessing, and improving the quality of microarray data [1–3]. In our own experience, through various statistically designed experiments, we have been able to identify and correct problems with the training of operators (scanner), inhomogeneous hybridizations, inadequate blocking of the poly-L-lysine coatings, print problems, and normalization procedures.

Along with others (see, e.g., [4, 5]), we have often observed effects of sources of variation that are manifested spatially. Frequently, these effects are most striking from the top to the bottom of an array. We have reduced these effects by modifying our hybridization processes to include a gentle rocking of the hybridization chamber (e.g., see also [6]). Nevertheless, even after this process modification, we have observed spatial effects that can result in apparent differences in relative expression of 30% or more across an array. Variation of this magnitude can be problematic

Table 1: Array assignment.

Slide	Cy3	Cy5
1	SYNW0947-sample no. 1	WH8102
2	WH8102	SYNW0947-sample no. 1
3	SYNW0947-sample no. 2	WH8102
4	WH8102	SYNW0947-sample no. 2

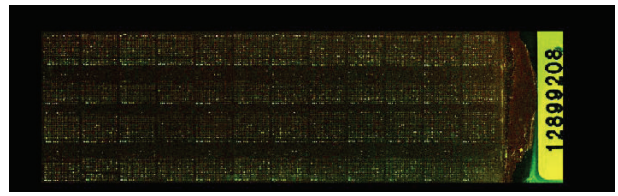

Figure 1: Full genome *Synechococcus* array.

when one is trying to identify genes that are weakly up- or downregulated. Thus, it is important to be able to easily monitor spatial effects.

The continuing effects of spatially-related sources of variation (including instances where printing or hybridization artifacts render a portion of an array completely unusable) have motivated the development of print designs that include replicate spots per gene that are spatially distributed over the array and printed with different pins. Combining this approach along with multiple technical and biological replicates is an effective way to provide the necessary data to enable a meaningful analysis that is able to separate the effects of multiple sources of variation and produce a more accurate assessment of a gene's true expression level.

In our study of gene expression in *Synechococcus* WH8102, we have constructed a complete genome microarray with multiple spots per gene spread out spatially across the array. This microarray is being used as a platform to compare various regulatory mutants of *Synechococcus* with the wild type under a variety of conditions and to study the effects of different sources of nitrogen or phosphorus for growth of the wild type [7]. Here we report a case study of the analysis of one of these experiments, comparing phosphorus metabolism of wild type and a strain in which a phosphorus-related response regulator gene has been inactivated.

Phosphorus can sometimes be a limiting nutrient in marine ecosystems (see, e.g., [8]). The availability of intracellular phosphorus for growth and the response of the cell to changing phosphorus levels are controlled in many bacteria by a two-component system including a histidine kinase (sensor) and response regulator (DNA-binding protein) pair, *PhoR* and *PhoB*, respectively, [9, 10]. In *Synechococcus* WH8102 the gene SYNW0947 is a *PhoB* homologue [11]. This gene was insertionally inactivated using the methods described in [12]. Gene expression of this mutant was then compared to that of wild type grown under standard conditions. This comparison along with other studies of cells grown under different phosphorus conditions will lead to an understanding of the phosphate regulon of these ecologically important microorganisms.

2. Materials and Methods

2.1. Experimental. The complete genome microarray for *Synechococcus* sp. strain WH8102 was used as the platform for a replicated dye-swap design [13] involving four slides (see Table 1). A single sample of the wild-type *Synechococcus* (WH8102) RNA was used as a control, while two samples of the mutant RNA were obtained for comparison. The microarray consists of a mixed population of PCR amplicons (2142 genes) and 70-mer oligonucleotides (389 genes). Unique PCR amplicons representing each gene are approximately 800 bp in size or smaller if the gene size is smaller. Unique 70-mer oligonucleotides were utilized for genes under 300 bp in size and for the two genes that we were unable to amplify by PCR. Six complete replicates of the 2531-member gene set were printed on aminosilane coated Corning ultraGAP glass slides using an Intelligent Automation Systems (IAS) high-precision microarray-printing robot with 48 pins for printing and irreversibly bound by UV-crosslinking at 250 mJ. Each array slide also includes a variety of negative controls (50% DMSO/50% deionized water) and positive controls (including a total mix of WH8102 PCR amplicons, spiked *Arabidopsis* PCR amplicons and 70-mer oligonucleotides).

The amplicons/oligonucleotides were split into two separate sets of 384-well plates with each amplicon/oligonucleotide in a different well position. This enabled us to develop a print pattern with each of the six replicate spots located in different blocks separated both horizontally and vertically across the slide.

The *Synechococcus* strains were grown in standard ocean water (SOW) medium, and total RNA was extracted using a Trizol-based method (Invitrogen) following manufacturers recommendations and purified using a mini RNeasy kit (Qiagen). The purity and yield of the RNA were determined spectrophotometrically by measuring optical density at wavelengths of 260 and 280 nm. An indirect labeling method was used to label cDNA, where cDNA was synthesized in the presence of a nucleoside triphosphate analog containing a reactive aminoallyl group to which the fluorescent dye molecule was coupled. Prior to hybridization, labeled cDNA was scanned spectrophotometrically to ensure optimal dye incorporation per sample for adequate signal intensity. A single sample of the wild-type *Synechococcus* (WH8102) RNA was used as a control, while two samples of the mutant RNA were obtained for comparison. Hybridizations were performed as previously described in [14], and slides were promptly scanned at a 10-μm resolution using an Axon 4000B scanner with GenePix 4.0 software.

Figure 1 displays the fluorescence image of a hybridized array. The array contains 19 200 spots in 48 blocks with 20 rows and columns in each block. Each of the genes appears in six different blocks within the array (and therefore is printed by six of the 48 different pins) and is assigned to a letter {A, B, C, D, E, F, G, or H}. For a given gene, the block positions are given by the position of its assigned letter in Figure 2. The position of a given gene within a block is consistent across

D	H	B	F	D	H	B	F	D	H	B	F
C	G	A	E	C	G	A	E	C	G	A	E
B	F	D	H	B	F	D	H	B	F	D	H
A	E	C	G	A	E	C	G	A	E	C	G

Figure 2: Full genome *Synechococcus* array (showing block positions of replicates).

its six replicates. In addition, the array contains a number of control spots, both positive and negative. Some control spots are used for alignment (e.g., see first column of the first few rows of each block), and others are used for quality control.

2.2. Data Preprocessing. TIGR's SPOTFINDER and MIDAS software [15] was used to process the four microarray images. This processing resulted nominally in a "4 arrays × 6 gene replicates × 2531 genes" data array consisting of the relative intensities, $I_{\text{Treatment}}/I_{\text{Control}}$, of each spot. The relatively few spots that were rejected were rejected only on the basis of poor visual quality. Spots with low intensity were not automatically rejected, resulting in quantitative representation of a vast majority of the genes over six spatially varying replicate spots on each array.

We use $\log_2(I_{\text{Treatment}}/I_{\text{Control}})$ as a basis for the quantitative analysis that follows.

2.3. Array Normalization. A two-step modeling process analogous to the approach used in [16] was used to normalize the data. However, unlike in [16], log(ratios) were used rather than log(intensities). First, the data were normalized by subtracting the slide-specific global average log-ratio. This adjusted for global effects (across all spots on a slide) due to the dye configuration (standard versus flipped) and/or the biological replicate. To formalize this, let Y_{gij} be the observed log-ratio associated with the gth gene, ith biological replicate, and jth *dye* configuration ($i = 1{:}2$ and $j = 1{:}2$). Then, the normalized expression data are given by $R_{gij} = Y_{gij} - \overline{Y}_{.ij}$, where $\overline{Y}_{.ij}$ represents the average expression level of the slide corresponding to the ith biological replicate and the jth *dye* configuration.

2.4. Variance Components Analysis. Following array normalization, a variance components analysis was used to partition the observed variability in expression level across replicate arrays. The purpose of this analysis was to help further understanding the relative magnitudes of the various sources of experimental variation. A model for the normalized expression data is given by $R_{gij} = G_g + (BG)_{gi} + (DG)_{gj} + \varepsilon_{gij}$, where R_{gij} is the observed normalized relative expression of the gth gene for the ith biological replicate and the jth *dye* configuration. G_g represents the true (but unknown) relative expression level of the gth gene, and $(BG)_{gi}$ and $(DG)_{gj}$ represent the random gene-specific effects associated with the biological replicate and the dye. The term ε_{gij} is representative of a nonspecific random effect that is unrelated to the biological replicate or the dye. The variances of these random effects are given by σ_b^2, σ_d^2, and σ_ε^2. The true expression level of a given gene is estimated as the average value of R over the four slides: $\hat{G}_g = (1/4)\cdot\sum_{i=1}^{2}\sum_{j=1}^{2} R_{gij}$.

One degree-of-freedom estimates for the three variance components can be obtained for *each gene* via an analysis of variance (ANOVA) of the values of R (see, e.g., [17]):

$$\hat{\sigma}_\varepsilon^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\left(R_{gij} - \overline{R}_{gi\cdot} - \overline{R}_{g\cdot j} + \hat{G}_g\right)^2,$$

$$\hat{\sigma}_b^2 = \max\left(0, \sum_{i=1}^{2}\left(\overline{R}_{gi\cdot} - \hat{G}_g\right)^2 - \frac{1}{2}\cdot\hat{\sigma}_\varepsilon^2\right), \tag{1}$$

$$\hat{\sigma}_d^2 = \max\left(0, \sum_{j=1}^{2}\left(\overline{R}_{g\cdot j} - \hat{G}_g\right)^2 - \frac{1}{2}\cdot\hat{\sigma}_\varepsilon^2\right),$$

where

$$\overline{R}_{gi\cdot} = \frac{1}{2}\cdot\sum_{j=1}^{2} R_{gij}, \quad \overline{R}_{g\cdot j} = \frac{1}{2}\cdot\sum_{i=1}^{2} R_{gij}. \tag{2}$$

Smoothed versions ("running 10%-trimmed means") of these summary statistics were also computed. That is, for each case, $(\hat{G}, \hat{\sigma})$ are ordered by the value of $\hat{G}$, resulting in $\{\hat{G}(1), \hat{G}(2), \ldots, \hat{G}(N)\}$ and $\{\hat{\sigma}(1), \hat{\sigma}(2), \ldots, \hat{\sigma}(N)\}$, where N is the number of genes considered. The left endpoint of each curve is given by the co-ordinates: $\text{median}_{i=1:100}(\hat{G}(i))$ and $\sqrt{\text{trimmed mean}_{i=1:100}(\hat{\sigma}^2(i))}$. In general the jth point of each curve is given by the coordinates: $\text{median}_{i=j:100+j-1}(\hat{G}(i))$ and $\sqrt{\text{trimmed mean}_{i=j:100+j-1}(\hat{\sigma}^2(i))}$. The trimmed mean is the average of the 100 observations with the smallest and largest 5 observations removed. In contrast to the noisy individual values of $\hat{\sigma}_d, \hat{\sigma}_b$, and $\hat{\sigma}_\varepsilon$ (which are each associated with a single degree of freedom), these curves provide a smooth visual perspective regarding the behavior of each of the variance components with varying levels of $\hat{G}$. In addition, statistics derived from these curves are used as a basis for making inference.

2.5. Standard Error of $\hat{G}_g$. Based on the gene-specific variance components estimates, a direct (but noisy) estimate of the standard error of $\hat{G}_g$ is given by

$$\hat{\sigma}_{\hat{G}_g} = \sqrt{\frac{\hat{\sigma}_d^2(\hat{G}_g)}{2} + \frac{\hat{\sigma}_b^2(\hat{G}_g)}{2} + \frac{\hat{\sigma}_\varepsilon^2(\hat{G}_g)}{4}}. \tag{3}$$

Alternatively, we can assume that the smooth versions of these variance components are more representative of the underlying true levels of the variance components and that these variance components are dependent only on the level of G_g. Denote these smooth curves by $\tilde{\sigma}_d(\hat{G})$, $\tilde{\sigma}_b(\hat{G})$, and $\tilde{\sigma}_\varepsilon(\hat{G})$. Based on these smooth curves, the estimated standard error of $\hat{G}$ is given by

$$\tilde{\sigma}_{\hat{G}} = \sqrt{\frac{\tilde{\sigma}_d^2(\hat{G})}{2} + \frac{\tilde{\sigma}_b^2(\hat{G})}{2} + \frac{\tilde{\sigma}_\varepsilon^2(\hat{G})}{4}}. \tag{4}$$

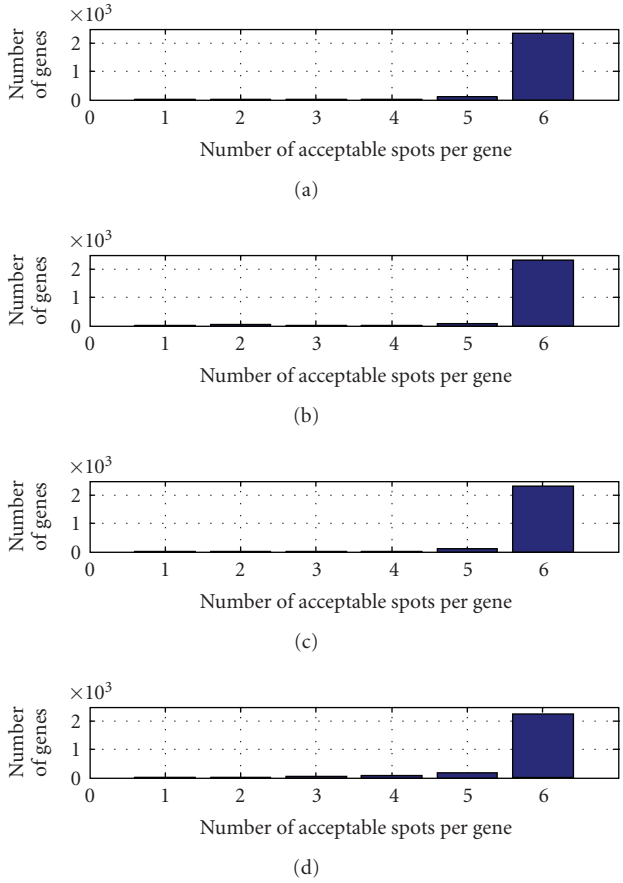

FIGURE 3: Number of genes with {1, 2, 3, 4, 5, or 6} acceptable spots per slide. Slides 1, 2, 3, and 4 are represented from top to bottom.

We are most interested in the constituent variance components and overall level of variability of $\hat{G}$ when $G = 0$ (corresponding to the case when the hypothetical treatment gene expression level is unchanged from the control). In practice, since we do not know what the true gene expression level (G) is, we are interested in the level of variability when $\hat{G} \approx 0$ (corresponding to the case where there is a relatively little observed change in the gene expression level). Evaluating $\tilde{\sigma}_{\hat{G}}$ at $\hat{G} = 0$, we computed

$$\tilde{\sigma}_0 = \sqrt{\frac{\tilde{\sigma}_d^2(0)}{2} + \frac{\tilde{\sigma}_b^2(0)}{2} + \frac{\tilde{\sigma}_\varepsilon^2(0)}{4}}. \quad (5)$$

2.6. Test Statistic. A test statistic was developed to form the basis for our assessment of whether a particular gene was significantly upregulated or downregulated. The test statistic is $S_g = \hat{G}_g / \hat{\sigma}_{\hat{G}_g(\text{com})}$, where $\hat{\sigma}_{\hat{G}_g}(\text{com}) = \max(\hat{\sigma}_{\hat{G}_g}, \tilde{\sigma}_0)$. The purpose of this *combined* estimate for the standard error of $\hat{G}_g$ is to prevent the computed statistic, S_g, from being too large (in absolute value) based on a chance small value of $\hat{\sigma}_{\hat{G}_g}$ that is not representative of the true value of $\sigma_{\hat{G}_g}$. Such nonrepresentative small values of $\hat{\sigma}_{\hat{G}_g}$ would not be uncommon due to the small sample size of 4 arrays. Note

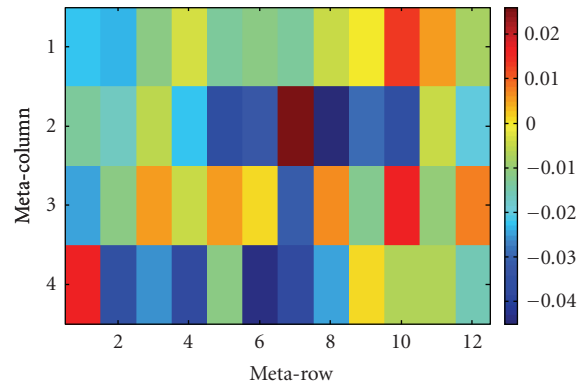

FIGURE 4: Median log-ratios within each block: slide no. 1.

that Cui and Churchill [18] discuss other modified t-tests used to assess differential expression. The floor of $\hat{\sigma}_{\hat{G}_g}(\text{com})$, $\tilde{\sigma}_0$, is analogous to the "fudge" term used in the widely used significance analysis of microarrays method (SAM) that was developed by Tusher et al. [19]. The distribution of this test statistic, when $G_g = 0$, is complicated and depends on assumptions about the random effects in the normalized gene expression model: $R_{gij} = G_g + (BG)_{gi} + (DG)_{gj} + \varepsilon_{gij}$.

If we assume that the random effects are normally distributed with zero mean and specified variances $(\sigma_d^2, \sigma_b^2, \sigma_\varepsilon^2)$, then selected percentiles of the null distribution of the test statistic can be estimated by simulating gene expression data via the model: $R_{ij} = G + B_i + D_j + \varepsilon_{ij}$ (i = 1:2 and j = 1:2) with $G = 0$. The simulation is set up to mimic the actual experiment: a replicated dye-swap design involving four slides and two biological samples. The experiment can be simulated many times with each realization resulting in a value for the test statistic, S_g. Selected order statistics from the distribution of S_g values obtained from the simulations provide approximate percentiles of the null distribution.

3. Results and Discussion

3.1. Assessment of Slide Quality and Identification of Anomalous Data. The four microarray images each containing six replicate representations of the 2531 genes were processed into a $4 \times 6 \times 2531$ data array of relative intensities. Spots were rejected solely on the basis of poor quality resulting in quantitative representation of a vast majority of the genes over six spatially varying replicate spots on each array. Figure 3 illustrates the distribution of acceptable spots per gene on each array. We recommend a graphic of this nature for experiments which have multiple spots per gene printed on each slide as it allows for a quick assessment of the relative quality of each slide in the study.

Here, due to the nature of the print design it is also possible to examine whether there are gross spatial effects within each slide. Note that the 48 blocks are arranged in a 12 meta-row by 4 meta-column configuration. About 300 genes are printed in each block. Figure 4 displays the median log-ratios of spots within each block for slide no. 1. Assuming that the typical gene is not differentially expressed, we expect

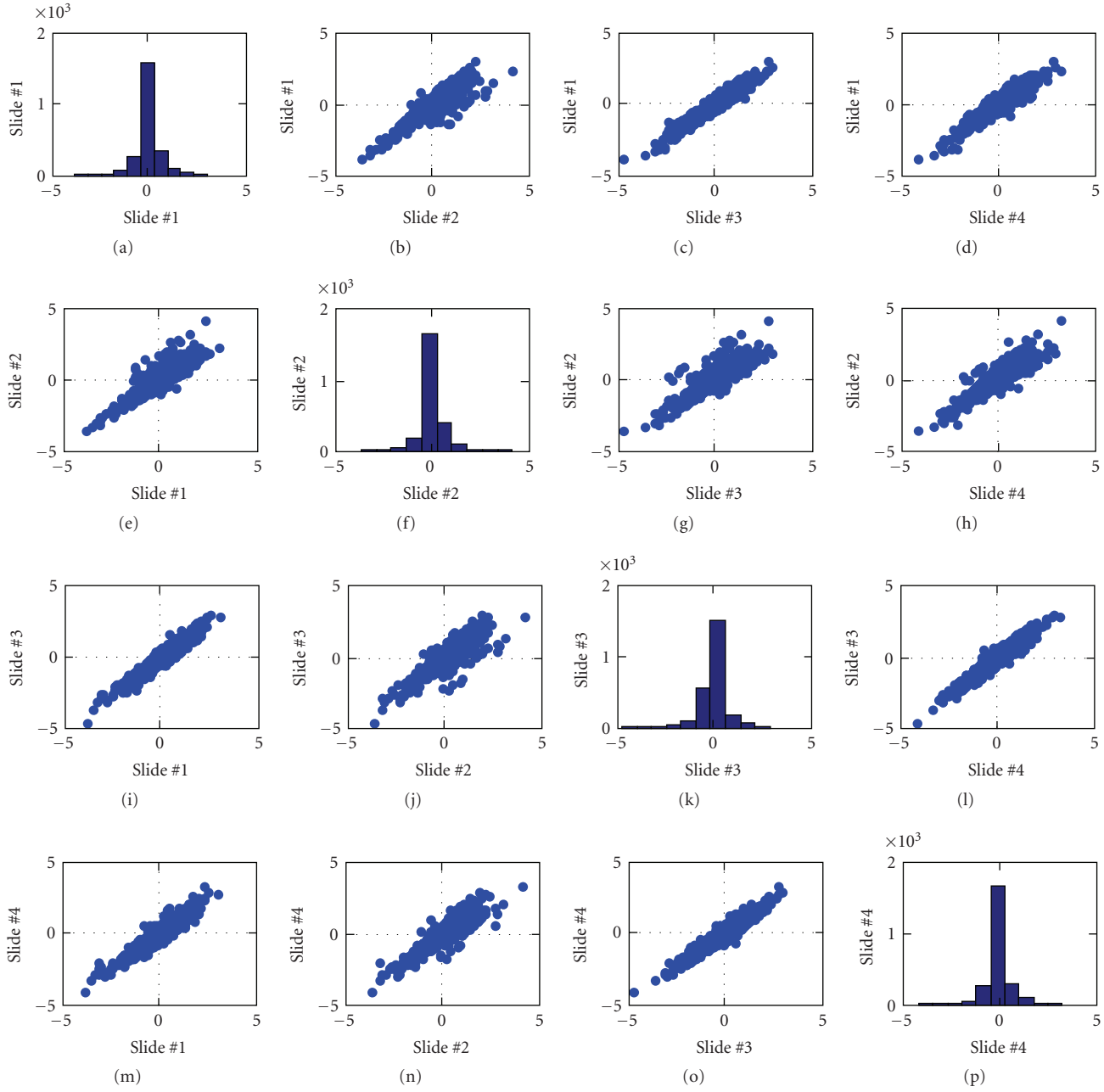

FIGURE 5: Scatterplot matrix of the median log-ratios. The expression distribution of each slide is represented along the diagonal of the scatterplot matrix.

that the median log-ratio for each block to be close to zero. Overall, the median log-ratios of slide no. 1 are slightly negative, but quite small in magnitude (effects span about 0.07 $\log_2$ units). However, as is the case with the other slides, no large block-to-block spatial effects are observed. Note that this is in contrast to earlier *Synechococcus* experiments that we conducted in which much larger spatial effects (spanning about 0.3 $\log_2$ units across slides) were observed but later improved by changing hybridization conditions. If such large effects were present in association with a traditional print design, the perceived expression level of genes with spots located only in the discrepant area would be inaccurate. In our print design, the influence of the spatial effects is minimized since affected genes are represented elsewhere in spatially distinct locations on the slide.

The results from the 2408 genes represented by at least 4 spots on each array of "acceptable" quality form the basis for further analysis and modeling. For each of these genes, we computed median($\log_2(I_{\text{Treatment}}/I_{\text{Control}})$) across the acceptable replicate spots within each slide. Figure 5 presents the relationship between values of median log-ratios across the four slides. For the most part, the median log-ratios are quite consistent across the four slides. However, there are a number of genes that produced atypically large

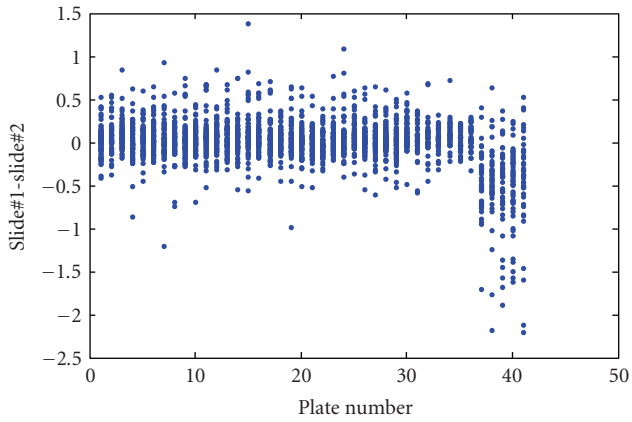

Figure 6: Difference in median log-ratios (slide no. 1-slide no. 2) versus plate number.

log-ratios for slide no. 2 (see scatter plots in the second row and the second column of Figure 5). A graphical analysis comparing slide no. 1 to slide no. 2 shows that these genes were associated with the last five print plates in the print run (see Figure 6). Although not confirmed, it is suspected that these effects are due to evaporation of the print solution. Figure 7 presents the relationship between values of median log-ratios across the four slides after excluding the 271 genes associated with the five suspect print plates.

3.2. Results of Array Normalization. The remaining data (involving 2137 genes) were normalized using the procedure described in Section 2.3. Figure 8 displays the values of $\overline{Y}_{\cdot ij}$ and hence illustrates the average effects of dye and biological replicate over the 4 slides. Notice that across a slide the average effect of the dye is about 0.05 $\log_2$ units, while the average effect due to the biological replicate is barely perceptible.

3.3. Results of Variance Components Analysis. As described in Section 2.4, one degree-of-freedom estimates of the three variance components ($\hat{\sigma}_d^2$, $\hat{\sigma}_b^2$, and $\hat{\sigma}_\varepsilon^2$) were obtained for *each gene* via an analysis of variance (ANOVA) of the values of R. These summary statistics ($\hat{G}_g$, $\hat{\sigma}_d^2$, $\hat{\sigma}_b^2$, and $\hat{\sigma}_\varepsilon^2$) were computed for *each gene* and are displayed in Figures 9–12. Figure 9 displays the empirical cumulative distribution of estimated gene expression levels ($\hat{G}_g$). For example, from this figure one can see that about 90% of the genes produced values of $|\hat{G}_g|$ that are less than one (or, exhibited less than a 2-fold change). Superimposed on the summary statistics in Figures 10–12 are the "curves" that represent the "running 10%-trimmed mean" of the summary statistics ($\hat{\sigma}_d$, $\hat{\sigma}_b$, and $\hat{\sigma}_\varepsilon$) versus $\hat{G}_g$.

From Figure 10, one can conclude that the magnitude of the gene-specific effects associated with the dye status does not depend strongly on the level of $\hat{G}$ as the curve is nearly flat. Conversely, Figures 11 and 12 show that the magnitudes of the "biological" and "nonspecific" sources of variation depend on the level of $\hat{G}$. As $|\hat{G}|$ increases, the magnitudes of the "biological" and "nonspecific" sources of variation increase. The asymmetry of the curves in Figures 11 and 12 is interesting. The data indicate the biological (and nonspecific) variation of positively expressed genes exceeds that of negatively expressed genes. It should be noted that in some of our other experiments, we have noted much more variation across biological replicates and in the future we hope to identify and minimize the underlying sources of the variation across biological replicates.

Table 2: Selected percentiles of S_g under assumption of no treatment effect based on 1 000 000 independent simulation realizations.

$\alpha/2$	$1-\alpha/2$ percentile
.01	2.16
.001	2.95
.0001	3.6
.000025	4.2

3.4. Identification of Up- and Downregulated Genes. The ultimate objective of this study is to discover differences between the wild type and mutant strains in their response to their growth environment. The assessment whether a particular gene is upregulated or downregulated in the mutant (compared to the wild-type) is based on the test statistic $S_g = \hat{G}_g/\hat{\sigma}_{\hat{G}_g(\text{com})}$, where $\hat{\sigma}_{\hat{G}_g}(\text{com}) = \max(\hat{\sigma}_{\hat{G}_g}, \tilde{\sigma}_0)$ as discussed in Sections 2.5 and 2.6. In the neighborhood around $\hat{G} = 0$, we find that $\tilde{\sigma}_d \approx 0.047$, $\tilde{\sigma}_b \approx 0.048$, $\tilde{\sigma}_\varepsilon \approx 0.067$, and thus

$$\tilde{\sigma}_0 = \sqrt{\frac{\tilde{\sigma}_d^2(0)}{2} + \frac{\tilde{\sigma}_b^2(0)}{2} + \frac{\tilde{\sigma}_\varepsilon^2(0)}{4}} = 0.058. \tag{6}$$

Selected percentiles of the test statistic given in Table 2 were obtained by simulating expression data (assuming that $\sigma_d = 0.047$, $\sigma_b = 0.048$, and $\sigma_\varepsilon = 0.067$) as described in Section 2.6. An individual gene is declared as being significantly expressed (either up or down relative to the control) if $|S_g| > 4.2$. This corresponds to a type-1 error of $\alpha = 0.00005$, meaning that the likelihood of incorrectly declaring a specific gene (i.e., in fact nondifferentially expressive) as being significantly expressive is about 0.00005. Using the very conservative Bonferroni correction for the simultaneous inference of about 2000 genes, we have a type-1 error of about 0.10. Figure 13 illustrates the set of 629 genes that were declared as being significantly expressed relative to the control. Note that the significance analysis of microarrays (SAMs) procedure developed by Tusher et al. [19]) was not used in this example due to the fact that it is not possible to create a good resampling distribution with the very restricted number of possible permutations available with only 4 slides (see, e.g., [20]).

A similar process was used to assess the expression level associated with the 271 genes whose slide no. 2 measurements were anomalous (see Figures 5 and 6). Again, we rely on the model $R_{ij} = G + B_i + D_j + \varepsilon_{ij}$ with specified levels of the random effects given by $\sigma_d = 0.047$, $\sigma_b = 0.048$,

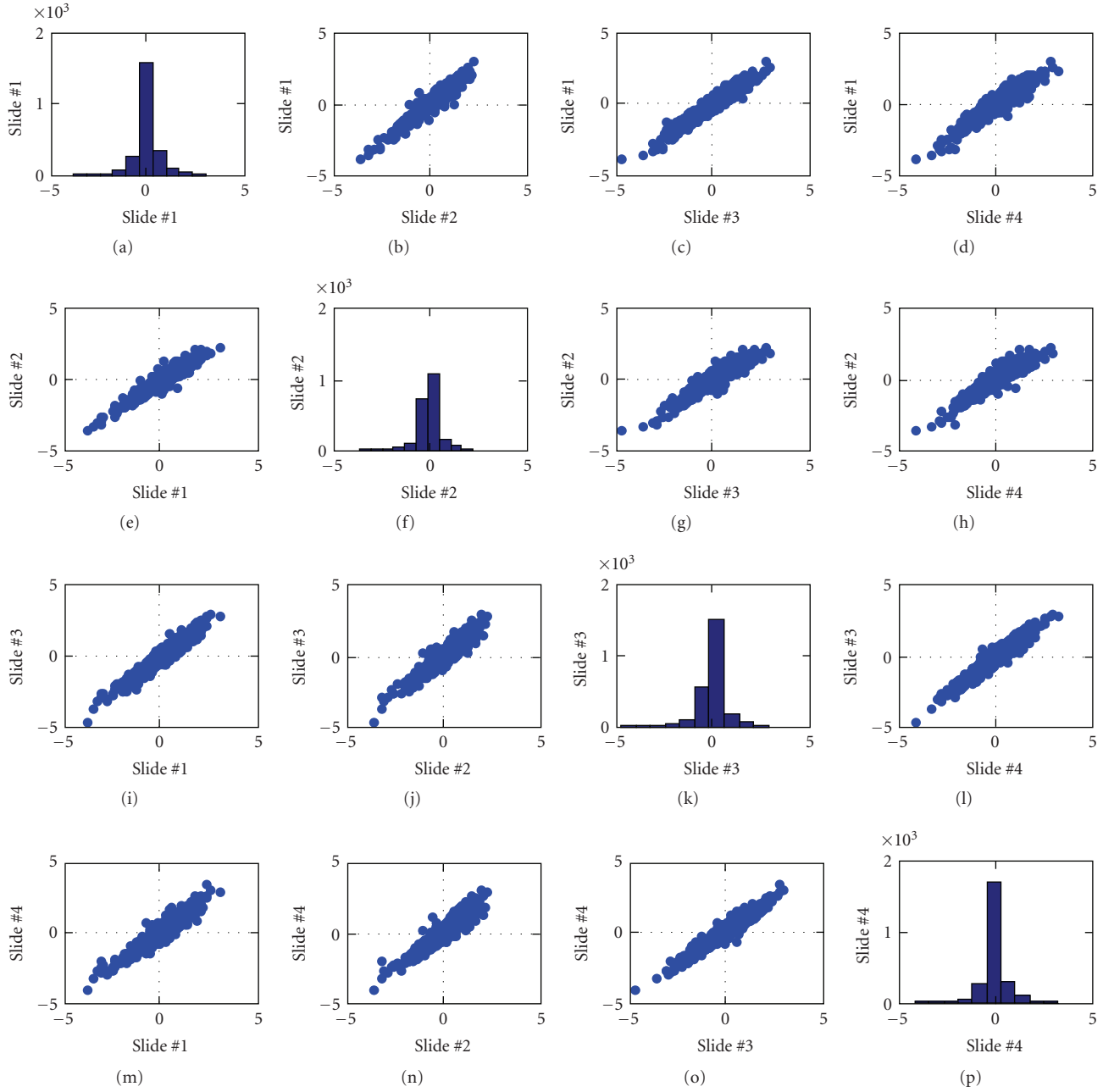

FIGURE 7: Scatterplot matrix of the median log-ratios (genes from 5 suspect plates removed). The expression distribution of each slide is represented along the diagonal of the scatterplot matrix.

and $\sigma_\varepsilon = 0.067$. Here, however, the simulation used to obtain the null distribution of the test statistic uses only three slides (since for these cases, results from three slides [rather than four slides] were used) and two biological samples. In this case, the test statistic is $S_g^* = \hat{G}_g^* / \hat{\sigma}_{\hat{G}_g^*(\mathrm{com})}$, where

$$\hat{G}_g^* = \frac{w_1 \cdot \overline{R}_{1\cdot} + w_2 \cdot R_{21}}{w_1 + w_2},$$

$$w_1 = \frac{1}{.5 \cdot \sigma_d^2 + .5 \cdot \sigma_\varepsilon^2 + \sigma_b^2}, \quad w_2 = \frac{1}{\sigma_d^2 + \sigma_\varepsilon^2 + \sigma_b^2},$$

$$\hat{\sigma}_{\hat{G}_g^*(\mathrm{com})} = \max\left(\hat{\sigma}_{\hat{G}_g}, 0.063\right), \quad \hat{\sigma}_{\hat{G}_g}^2 = \frac{1}{\hat{w}_1 + \hat{w}_2},$$

$$\hat{w}_1 = \frac{1}{.5 \cdot \hat{\sigma}_w^2 + \hat{\sigma}_b^2}, \quad \hat{w}_2 = \frac{1}{\hat{\sigma}_w^2 + \hat{\sigma}_b^2},$$

$$\hat{\sigma}_w^2 = \sum_{i=1}^{2} \left(R_{1i} - \overline{R}_{1\cdot}\right)^2,$$

$$\hat{\sigma}_b^2 = \max\left(0, \frac{\left(2 \cdot \left(\overline{R}_{1\cdot} - \overline{\overline{R}}\right)^2 + \left(R_{21} - \overline{\overline{R}}\right)^2\right) - \hat{\sigma}_w^2}{(3 - 5/3)}\right),$$

$$\overline{R}_{1\cdot} = .5 \cdot (R_{11} + R_{12}), \quad \overline{\overline{R}} = \frac{1}{3} \cdot (R_{11} + R_{12} + R_{21}). \tag{7}$$

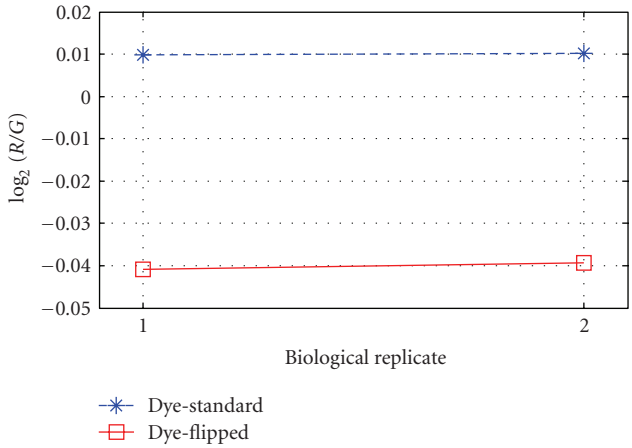

FIGURE 8: Average effects of dye and biological replicate (genes from suspect plates removed).

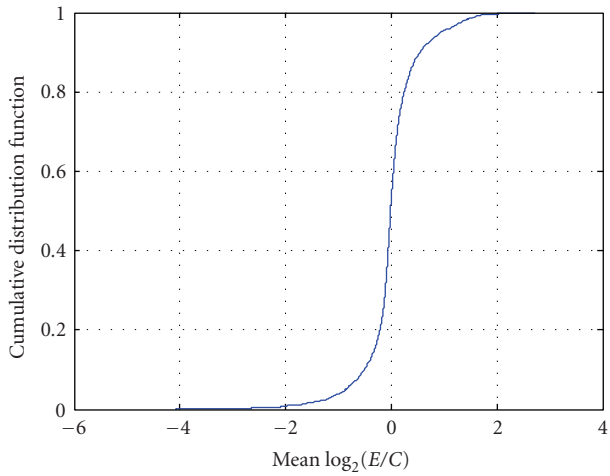

FIGURE 9: Cumulative distribution of estimated gene expression levels ($\hat{G}_g$).

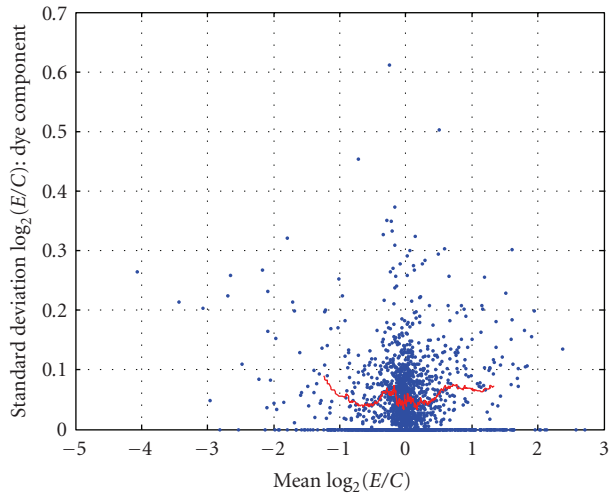

FIGURE 10: $\hat{\sigma}_d^2$ versus $\hat{G}_g$.

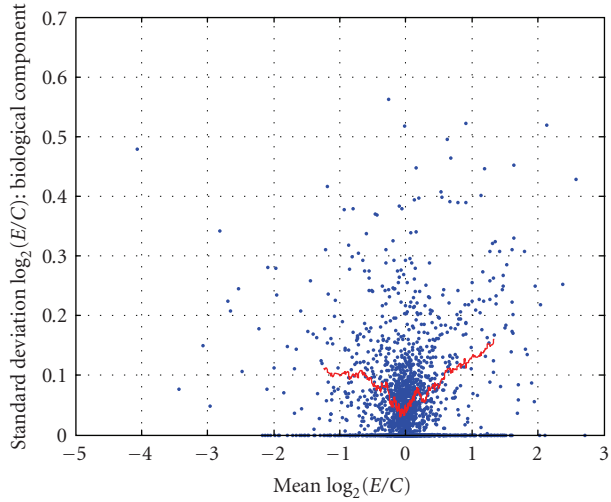

FIGURE 11: $\hat{\sigma}_b^2$ versus $\hat{G}_g$.

The estimates for σ_w^2 and σ_b^2 were obtained using methods for unbalanced data described in [17, page 72]. The second argument (0.063) in the definition for $\hat{\sigma}_{\hat{G}_g^*(\text{com})}$ is the variance of $\hat{G}_g^*$ obtained by assuming $\tilde{\sigma}_d(0)$, $\tilde{\sigma}_b(0)$, and $\tilde{\sigma}_\varepsilon(0)$.

From the simulation, we found approximate percentiles of the distribution of S_g^*. For example, the 0.000025 (0.999975) percentile was found to be about -3.95 (3.95). Thus, with a type-1 error of $\alpha = 0.00005$, an individual gene is declared as being significantly expressed if $|S_g^*| > 3.95$. Of these 271 genes in question, 90 were deemed to be significantly expressed (43 positive and 47 negative).

Overall, across all 2408 genes considered (the 2137 genes represented on 4 slides plus the 271 genes represented on 3 slides), 719 genes were deemed to be significantly up- or downregulated. Tables 3 and 4 list the 15 genes that were the most upregulated and the 15 genes that were the most downregulated. The supplementary tables list all genes that were significantly up- or downregulated (See Tables 1 and 2 in the Supplementary Material available online at doi:10.1155/2009/950171).

Figure 14 provides the cumulative distributions of $|\hat{G}_g|$ for both the selected and unselected genes. Based on the floor for $\hat{\sigma}_{\hat{G}_g}(\text{com})$ ($\tilde{\sigma}_0 = 0.058$ or $\tilde{\sigma}_0 = 0.063$) and the selected threshold of 4.2 (or 3.95), the minimum level of $|\hat{G}_g|$, such that gene is declared significant, is $4.2 \cdot 0.058 \approx 0.24 \log_2$ relative expression units. About 1400 of the 2408 genes are associated with values of $|\hat{G}_g|$ less than $0.24 \log_2$ relative expression units. Almost three quarters of the remaining genes (719 out of 993) were deemed to have been significantly expressed relative to the control. About 70% of the 719 significant genes exhibited less than a 2-fold change in intensity. About 35% of the significant genes exhibited less than a 1-fold change in intensity. Thus, we are able to identify large numbers of genes for which the treatment causes a small, but significantly different level in expression when compared to the control.

3.5. Biological Interpretations. One interesting biological outcome of these results is the extent to which changes in the phosphorus regulatory system seem to affect the gene

Table 3: Statistically significant genes with highest level of upregulation. $\hat{G}_g$: estimated relative expression level, S_g: test statistic.

Gene ID	$\hat{G}_g$	S_g	Gene description
SYNW1555	2.72	13.47	Hypothetical
SYNW2478	2.58	7.42	Conserved hypothetical protein
SYNW2480	2.37	11.28	ABC transporter, ATP binding component, possibly zinc transport
SYNW0524	2.13	6.10	Conserved hypothetical protein
SYNW0424	2.13	5.46	Possible HMGL-like family protein
SYNW2481	2.10	9.36	Putative zinc transport system substrate-binding protein
SYNW1305	2.03	11.34	Hypothetical
SYNW0947	2.03	12.86	Two-component response regulator, phosphate
SYNW1463	2.02	32.03	Hypothetical
SYNW2479	1.96	9.04	ABC transporter component, possibly Zn transport
SYNW1654	1.95	13.34	Conserved hypothetical protein
SYNW2486	1.91	15.10	Putative cyanate ABC transporter
SYNW0454	1.84	12.43	Possible glycosyltransferase
SYNW1947	1.81	14.14	Conserved hypothetical protein
SYNW0456	1.79	7.00	Possible glycosyltransferase

Table 4: Significant genes with highest level of downregulation. $\hat{G}_g$: estimated relative expression level, S_g: test statistic.

Gene ID	$\hat{G}_g$	S_g	Gene description
SYNW2508	−4.07	−10.24	Molecular chaperone DnaK2, heat shock protein hsp70-2
SYNW0514	−3.44	−21.37	GroEL chaperonin
SYNW1503	−3.06	−17.05	Endopeptidase Clp ATP-binding chain B
SYNW1797	−2.96	−47.82	Putative iron ABC transporter, substrate binding protein
SYNW0513	−2.94	−41.756	GroES chaperonin
SYNW1278	−2.90	−19.209	Heat shock protein HtpG
SYNW2391	−2.81	−8.68	Putative alkaline phosphatase
SYNW1018	−2.69	−11.50	ABC transporter, substrate binding protein, phosphate
SYNW1798	−2.65	−11.14	Putative iron ABC transporter
SYNW1511	−2.58	−25.108	Conserved hypothetical
SYNW0938	−2.54	−12.69	Endopeptidase Clp ATP-binding chain C
SYNW2390	−2.48	−22.48	Putative alkaline phosphatase/5′ nucleotidase
SYNW0835	−2.22	−15.82	Probable oxidoreductase
SYNW1842	−2.17	−10.44	Apocytochrome f
SYNW0670	−2.14	−7.950	Conserved hypothetical protein

expression of multiple genes beyond those strictly involved in phosphorus acquisition. This may be due to the many uses of phosphorus in the cell. It may also be due to the relatively small number of two-component regulatory systems in open ocean cyanobacteria, for example, only 5 histidine kinase sensors and 9 response regulators [11] and the possibility of substantial cross-talk among these systems. Inactivating one response regulator may affect this regulatory cross-talk. One unknown is whether the inactivation of SYNW0947 caused polar effects on nearby genes, especially the downstream *phoR* (SYNW0948) although this would still be part of changing the phosphorus regulatory system.

In addition, this statistical approach should allow for a much more robust identification of operons especially if gene expression in genes later in an operon are attenuated. The microarray results presented here suggest that several clusters of genes are potentially operons. For example, SYNW1016 and SYNW1017 were both significantly downregulated (see the supplementary tables). These are genes that are next to two other genes known to be involved in phosphate metabolism (SYNW1018 and SYNW1019). In addition a set of genes (SYNW0465-SYNW0470) were all highly upregulated and thus are a potential operon involved in phosphate metabolism. Interestingly, a third region probably comprising several operons (SYNW2477-2491) was also upregulated. These predictions merit further experimentation such as gene knockouts. As can be seen in Supplementary Figure 1 no spatial clustering of genes is apparent, suggesting that the operons detected are being found purely as a consequence of their place in regulatory networks affected by phosphate limitation.

We utilized the pathway analysis package DAVID (http://david.abcc.ncifcrf.gov/home.jsp) to examine the extent to which pathways, potentially involving multiple

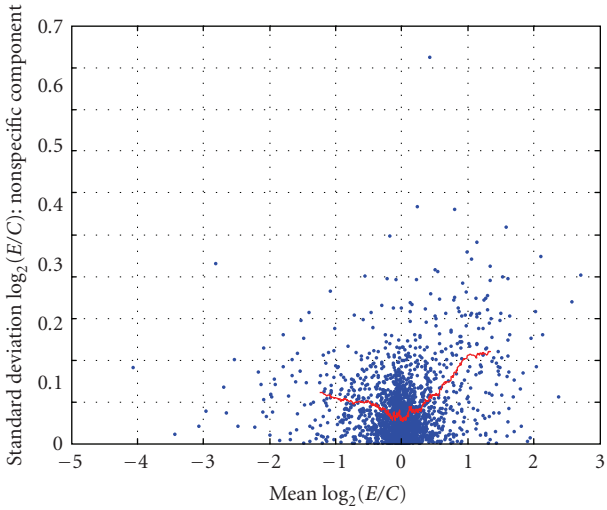

FIGURE 12: $\hat{\sigma}_\varepsilon^2$ versus $\hat{G}_g$.

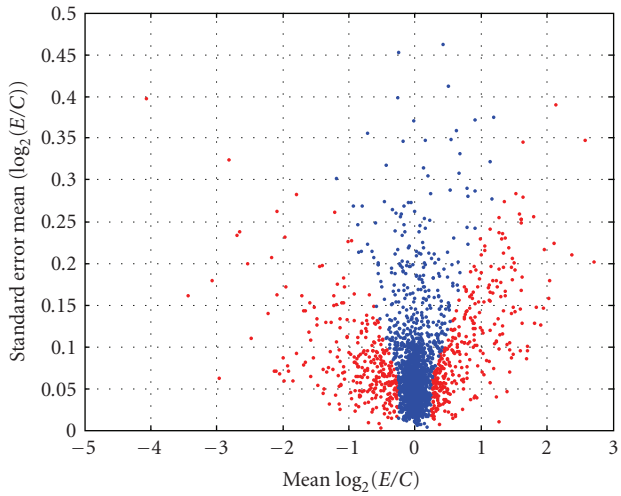

FIGURE 13: Significantly upregulated or downregulated genes (red): $|S_g| > 4.2$.

operons, are altered in the SYNW0947 mutant. The up- and downregulated genes from our analysis as well as using a simple 2-fold change statistic were mapped to KEGG pathways (see Supplementary Tables 1 and 2 where genes with simple 2-fold changes are shown in bold). We mapped 67 upregulated (of 360) genes to KEGG pathways while only 20 (of 100) genes were mapped using a 2-fold change. Our results demonstrated a much more convincing upregulation of the photosynthetic antenna proteins (9 genes) compared to the simpler analysis (5 genes). In addition new pathways involving mannose metabolism (SYNW0422, SYNW0423, SYNW0919) and other sugars were convincing upregulated in our analysis but were not seen with a 2-fold change statistic. We mapped 154 downregulated (of 337) genes to KEGG pathways compared to 40 (of 83) genes with a 2-fold change. Interestingly, we were able to map a larger fraction of downregulated genes to KEGG pathways. Again we saw a much more convincing downregulation of specific

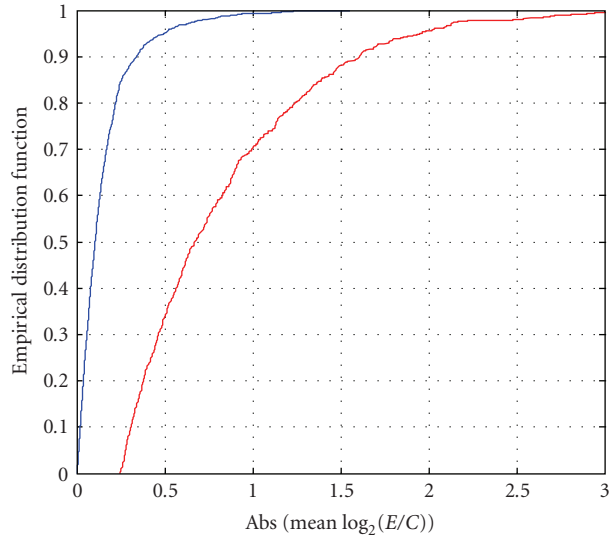

FIGURE 14: Cumulative distributions of $|\hat{G}_g|$ for selected (unselected) genes.

pathways. We found 28 ribosomal genes downregulated compared to 8 using a 2-fold statistic. Since these genes are likely to be coregulated, our results are biologically coherent. Similarly 15 photosynthesis genes were downregulated compared to 5 in the simpler analysis. Interestingly, the cells are downregulating core phycobilisome antenna proteins while upregulating rod proteins. This suggests that they are making fewer but larger light harvesting antenna complexes.

4. Conclusions

We have used a replicated dye-swap experiment with multiple spots per gene per array as a platform for comparing a regulatory mutant of *Synechococcus* sp. WH8102 with the wild type under defined growth conditions in artificial seawater. Our process for analyzing the experimental data includes utilizing simple graphical displays. These displays were used to assess spot quality, spatial variability within an array, array-to-array reproducibility, as well as other effects due to special causes (e.g., well plate). Quantitative analysis was based on the median expression level (within an array) of each gene. Following array normalization, a variance components analysis was used to partition the observed variability in expression level across replicate arrays. The level of variability introduced by dye swapping was found to be relatively small and independent of the apparent expression level. The variation in gene expression across biological replicates was found to be more significant and was found to be dependent on the apparent expression level. As only one strain was utilized, the biological significance of the data cannot be extended beyond the wild type strain used, but the statistical method developed with this model will allow greater sensitivity than was previously possible. The assessment of whether a particular gene is upregulated or downregulated was based on a test statistic that excludes genes that would otherwise be identified solely on the basis of a chance abnormally low level of variation across arrays.

The null distribution of the test statistic was computed making a number of assumptions and by carefully constructing a simulation that mimicked the experiment and the observed sources of variation. A relatively large proportion of the genes were identified as being significantly upregulated or downregulated by the treatment, albeit with relatively small changes in the levels of expression. The ability to detect these small changes in the levels of expression (as small as about $0.25 \log_2$ units) is a direct consequence of the replication within the array.

Acknowledgments

This work was funded largely by the US Department of Energy's Genomes to Life program (http://www.doegenomestolife.org/) under the project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling." This work was partly supported by a US Department of Energy Grant, DOE DE-FG03-O1ER63148 to BP, BB, and IP. The authors would also like to thank Rob Herman and Lori Crumbliss for technical assistance.

References

[1] M. K. Kerr and G. A. Churchill, "Statistical design and the analysis of gene expression microarray data," *Genetical Research*, vol. 77, no. 2, pp. 123–128, 2001.

[2] M. K. Kerr, C. A. Afshari, L. Bennett, et al., "Statistical analysis of a gene expression microarray experiment with replication," *Statistica Sinica*, vol. 12, no. 1, pp. 203–217, 2002.

[3] M.-L. T. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar, "Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 18, pp. 9834–9839, 2000.

[4] J. J. Chen, R. R. Delongchamp, C.-A. Tsai, et al., "Analysis of variance components in gene expression data," *Bioinformatics*, vol. 20, no. 9, pp. 1436–1446, 2004.

[5] G. Balázsi, K. A. Kay, A.-L. Barabási, and Z. N. Oltvai, "Spurious spatial periodicity of co-expression in microarray data due to printing design," *Nucleic Acids Research*, vol. 31, no. 15, pp. 4425–4433, 2003.

[6] C. J. Schaupp, G. Jiang, T. G. Myers, and M. A. Wilson, "Active mixing during hybridization improves the accuracy and reproducibility of microarray results," *BioTechniques*, vol. 38, no. 1, pp. 117–119, 2005.

[7] Z. Su, F. Mao, P. Dam, et al., "Computational inference and experimental validation of the nitrogen assimilation regulatory network in cyanobacterium *Synechococcus* sp. WH 8102," *Nucleic Acids Research*, vol. 34, no. 3, pp. 1050–1065, 2006.

[8] D. J. Scanlan and W. H. Wilson, "Application of molecular techniques to addressing the role of P as a key effector in marine ecosystems," *Hydrobiologia*, vol. 401, pp. 149–175, 1999.

[9] J. B. Stock, A. J. Ninfa, and A. M. Stock, "Protein phosphorylation and regulation of adaptive responses in bacteria," *Microbiological Reviews*, vol. 53, no. 4, pp. 450–490, 1989.

[10] T. A. Hirani, I. Suzuki, N. Murata, H. Hayashi, and J. J. Eaton-Rye, "Characterization of a two-component signal transduction system involved in the induction of alkaline phosphatase under phosphate-limiting conditions in *Synechocystis* sp. PCC 6803," *Plant Molecular Biology*, vol. 45, no. 2, pp. 133–144, 2001.

[11] B. Palenik, B. Brahamsha, F. W. Larimer, et al., "The genome of a motile marine *Synechococcus*," *Nature*, vol. 424, no. 6952, pp. 1037–1042, 2003.

[12] B. Brahamsha, "A genetic manipulation system for oceanic cyanobacteria of the genus *Synechococcus*," *Applied and Environmental Microbiology*, vol. 62, no. 5, pp. 1747–1751, 1996.

[13] D. Amaratunga and J. Cabrera, *Exploration and Analysis of DNA Microarray and Protein Array Data*, John Wiley & Sons, New York, NY, USA, 2004.

[14] S. N. Peterson, C. K. Sung, R. Cline, et al., "Identification of competence pheromone responsive genes in *Streptococcus pneumoniae* by use of DNA microarrays," *Molecular Microbiology*, vol. 51, no. 4, pp. 1051–1070, 2004.

[15] A. I. Saeed, V. Sharov, J. White, et al., "TM4: a free, open-source system for microarray data management and analysis," *BioTechniques*, vol. 34, no. 2, pp. 374–378, 2003.

[16] R. D. Wolfinger, G. Gibson, E. D. Wolfinger, et al., "Assessing gene significance from cDNA microarray expression data via mixed models," *Journal of Computational Biology*, vol. 8, no. 6, pp. 625–637, 2001.

[17] S. R. Searle, G. Casella, and C. E. McCulloch, *Variance Components*, John Wiley & Sons, New York, NY, USA, 1992.

[18] X. Cui and G. A. Churchill, "Statistical tests for differential expression in cDNA microarray experiments," *Genome Biology*, vol. 4, article 210, no. 4, pp. 1–10, 2003.

[19] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[20] S. Draghici, *Data Analysis Tools for DNA Microarrays*, Chapman & Hall/CRC, Boca Raton, Fla, USA, 2003.