MARCEL DEKKER, INC. • 270 MADISON AVENUE • NEW YORK, NY 10016

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
Vol. 13, No. 1, pp. 57–74, 2003

Normalization Methods for Analysis of Microarray Gene-Expression Data

Yi-Ju Chen,[1] Ralph Kodell,[1] Frank Sistare,[3] Karol L. Thompson,[3] Suzanne Morris,[2] and James J. Chen[1,*]

[1]Division of Biometry and Risk Assessment and [2]Division of Genetic and
Reproductive Toxicology, National Center for Toxicological Research,
Food and Drug Administration, Jefferson, Arkansas, USA
[3]Division of Applied Pharmacology Research, Center for Drug Evaluation
and Research, Food and Drug Administration, Laurel, Maryland, USA

ABSTRACT

This paper investigates subset normalization to adjust for location biases (e.g., splotches) combined with global normalization for intensity biases (e.g., saturation). A data set from a toxicogenomic experiment using the same control and the same treated sample hybridized to six different microarrays is used to contrast the different normalization methods. Simple *t*-tests were used to compare two samples for dye effects and for treatment effects. The numbers of genes that reproducibly showed significant *p*-values for the unnormalized data and normalized data from different methods were evaluated for assessment of different normalization methods. The one-sample *t*-statistic of the ratio of red to green samples was used to test for dye effects using only control data. For treatment effects, in addition to the one-sample *t*-test of the ratio of the treated to control samples, the two-sample *t*-test for testing the difference between treated and control samples was also used to compare the two approaches. The method that combines a subset approach (median or *lowess* fit) for location adjustment with a global lowess fit for intensity adjustment appears to perform well.

[*]Correspondence: Dr. James J. Chen, NCTR/FDA/HFT-20, Jefferson, AR 72079; Fax: (870) 543-7662; E-mail: jchen@nctr.fda.gov.

DOI: 10.1081/BIP-120017726 1054-3406 (Print); 1520-5711 (Online)

www.dekker.com

Key Words: Consistency; Dye-swap experiment; Global and subset normalizations; Median and lowess adjustments; One-sample and two-sample t-tests.

1. INTRODUCTION

The development of cDNA and oligonucleotide microarray technology provides exciting tools for studying the expression levels of thousands of distinct genes simultaneously. A common cDNA microarray experiment contains two cDNA samples differently labeled using fluorescent dyes. In the experiment, samples of DNA clones with known sequence content are spotted and immobilized onto a glass slide or other substrate. Next, pools of mRNA from the cell populations under study are purified, reversed-transcribed into cDNA, and labeled with one of two fluorescent dyes, “red” or “green.” Two pools of differentially labeled cDNA are combined and applied to a microarray. Labeled cDNA in the pool hybridizes to the spots containing complementary sequences on the array. After this hybridization, the amount of individual hybridization of each of the two samples to each spot is quantified by scanning with laser for each dye on the array. The intensity of the red and green signals measured corresponds to the levels of gene expression for the two samples. Procedures must be applied to the set of red and green signals to most accurately represent relative gene-expression ratios.

It has been recognized that there are many sources of systematic variation in assigning expression levels to the measured fluorescence intensities. Variation factors include efficiency of dye incorporation, variable dye self-quenching, regional hybridization and quenching artifacts, differential spot quality, variable experimental conditions in the labeling process, scanner settings, image capture options, etc. Various normalization methods have been proposed (e.g., Efron et al., 2000; Kerr et al., 2000; 2001; Schuchhardt et al., 2000; Spellman et al., 1998; Wolfinger et al., 2001; Yang et al., 2001) to reduce some of the variability. Normalization strategies depend on the experimental design and data collection process. A simple approach is to scale the intensities based on a small subset of genes, called housekeeping genes, that are believed to have constant expression across a variety of conditions. However, it is very hard to identify a set of housekeeping genes that consistently do not change significantly under all treatment conditions. Also, the housekeeping genes may not be representative of other genes of interest (Yang et al., 2001). Furthermore, in practice, too often the ratios of proposed housekeeping genes within an array are not consistent across the array.

In a given experiment, it is assumed that only a small number of genes is expected to be differentially expressed and that if there is a large number of genes that are differentially expressed, there will be an equal number of genes that are down-regulated or up-regulated. The remaining genes are expected to have constant expression and so can be used for normalization (Delongchamp et al., 2002). Since the list of genes that are differentially expressed is not known, the commonly applied global normalization method uses all genes on the array for adjustment. Global normalization mathematically assumes a constant proportionality factor across all genes on the array. In many experiments, the proportionality appears to be different in different locations and/or at different intensity levels (Chen et al., 1997; Delongchamp et al., 2002; Herzel et al., 2001). In this paper, we propose subset normalization to account for systematic biases which may vary locally, and also to investigate intensity normalization.

Since the ratio of the fluorescence intensity for each spot measures the relative abundance of the corresponding gene under two different experimental conditions, a popular concept is to use the ratio to provide an adjustment given by Chen et al. (1997). Yang et al. (2001) considered normalization methods in terms of the ratio of fluorescence intensities within each array. They used the lowess fit to adjust for intensity and location dependency biases. Kerr and Churchill (2001) recommended that an analysis should use all the information in the data and not reduce to ratios. They proposed an analysis of variance (ANOVA) model for individual red and green intensities. The ANOVA model simultaneously adjusts for the dye, within- and among-array effects globally. The ANOVA model uses the mean to estimate normalization factors. Delongchamp et al. (2002) recommended the median estimate since the median is more robust against the highly over- or underexpressed genes. Also, the median is preferred over the mode since it is straightforward to calculate. Each approach has its advantages under certain experimental conditions or assumptions. Depending on the experimental design and the method of normalization, either one-sample or two-sample t-statistics of the adjusted intensities are used to test for treatment effects.

It is difficult to evaluate differential gene expression simply based on comparisons of the treated and control sample fluorescent intensities. It is well known that when two identical RNA samples are reverse-transcribed, labeled with different dyes, and hybridized on the same slide, the green intensities tend to be brighter than the red intensities. Also, the magnitude of the difference may depend on overall intensity. However, two control samples labeled with different color dyes can be used to evaluate a normalization procedure and six replicate hybridizations of the same control and same treated samples can be used to test consistency. This paper presents several adjustment methods using the median and lowess fit sequence estimates to account for location biases and lowess fits to account for intensity biases. We use a microarray data set from a toxicogenomic project conducted at the Center for Drug Evaluation and Research (CDER), FDA, to illustrate different normalization methods. This experiment provided data for both treatment-control and control-control comparisons in a dye-swap design.

This paper is organized as follows. In Section 2, we give a brief description of the toxicogenomic data and present the issues of normalization for the analysis of microarray data. In Section 3, we describe different normalization methods and two approaches to modeling the microarray data of two-dye intensity measurements. In Section 4, we present the results of applying different normalization methods to the toxicogenomic data. The numbers of genes identified and the consistency of the specific genes identified to be differentially expressed are evaluated for the two approaches using one-sample and two-sample t-tests. In Section 5, we discuss several issues regarding normalization for the analysis of microarray data.

2. ANALYSIS OF THE DATA BEFORE NORMALIZATION

2.1. Description of the Experiment

The data set is from a study of gene-expression levels of kidney samples from rats dosed with cisplatin, a known kidney toxin. Details of the study are given in Thompson et al. (2002). Control and treated samples from this study were hybridized on six replicate experiments (arrays A1–A6). A separate array (array A0) was hybridized with control samples labeled separately with Cy5 (red dye) and one with Cy3 (green dye). Each replicate sample was labeled independently (one treated and one control). There were 12 separate labeling reactions. On the arrays A1–A3, the control samples were assigned to the green dye and treated samples were assigned to the red dye. The dye assignments to the control and treated samples were reversed on the arrays A4–A6. The array is a 700 gene cDNA rat chip from Phase-1 Molecular Toxicology (Santa Fe, NM). In each array there are 4×4 grids of 14×14 spots. Grids 9–12 are replicates of grids 1–4, and grids 13–16 are replicates of grids 5–8. On each grid, genes were spotted in duplicate, so each gene has four replicate values on each array. In addition, sequences of four genes from plant and one from bacteria were also spotted on the array, each with four replications, to monitor for nonspecific background binding of labeled cDNA. There are also numerous “blank” and “empty” spots with buffer (blank) or nothing (empty) spotted in that region, respectively. The control data consisted of one array, A0, with both red and green measurements, three green measurements from arrays A1–A3, and three red measurements from arrays A4–A6. Arrays A1 to A6 were all labeled and hybridized on one date; A0 was labeled and hybridized on a separate date.

Figure 1a shows the density plots for array A0. The dark and light lines represent the red and green fluorescent readings, respectively. The solid lines represent the genes from the upper grids (G1–G8) and the dotted lines represent the lower grids (G9–G16). The upper grids will be referred to as replicate 1 and the lower grids as replicate 2. Figure 1b is the plot of the log-intensity ratio (M) vs. the mean log-intensity (A) for the two samples, where $M = \log_2(R/G)$ and $A = (1/2)\log_2(RG)$. If no dye normalization is needed, the points should center around zero on the vertical axis. Note that the points “ $+$ ” are the plant genes and bacterial gene and the green points are blank and empty spots. The plots show that there are systemic location effect differences (Fig. 1a) in dye intensity across the individual DNA probes (Fig. 1a and b).

2.2. Tests for Dye Effects

The control samples of two different dye labelings on the array A0 can be used to test for dye effects. For comparison purposes, we used the average of the two duplicate spots in the same grid in this analysis. Each gene thus had two measurements. That is, we treated the observations from upper grids and lower grids as two replicates on the array. We did not treat each of the four spots as independent since the variations between the duplicates are too small (more details are given below). Let r_{il}^{c} denote the log-ratio (in base 2) of the red to green intensities for the l^{th} replicate in the i^{th} gene, $i = 1, \ldots, 707$ and $l = 1, 2$.

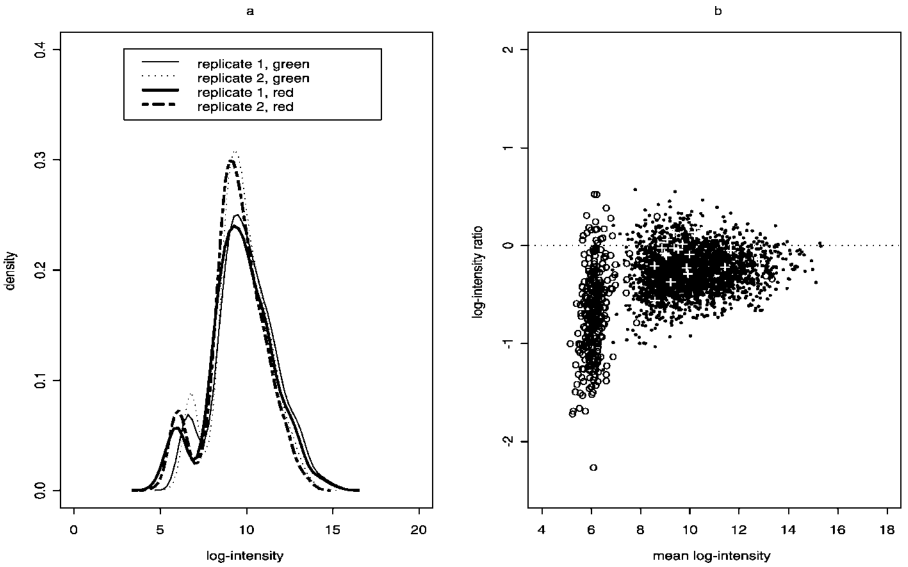

Figure 1. Data display for control vs. control data (A0).

The one-sample t-statistic for comparing the two color measurements is

$$l_i = \frac{\bar{r}_i^c}{S_i^c},$$

where $\bar{r}_i^c$ is the mean for the i th gene and s_i^c is the standard error of the $\bar{r}_i^c$. The number of genes with p-values less than or equal to 0.05 for array A0 is 138, of which the blank is significant (the expected number is 35 if the dyes are the same). It shows a highly significant dye effect.

Dye effects can also be tested from the ratio of the relative intensities of A1–A6 to A0. Six ratios of the red to green intensities were formed as follows. The red intensity on A1–A3 was divided by the A0 green intensity, and the A0 red intensity was also divided by each green intensity on A4–A6. For this data set, the number of genes with p-values less than or equal to 0.05 is 15. These results do not show any evidence of a dye effect, which is contrary to the conclusion shown from the array A0. The discrepancy can be due to a large array to array variation, or it may be because the intensities measured from different arrays are scaled differently. We applied different normalization methods to investigate these discrepancies.

In this design, there are four intensity measurements for each gene (more than four for the blank and empty spots) on an array. In this analysis and the analysis of treatment and control comparisons, the average of the four measurements is used for the comparison. However, the normalization is performed on the individual spot intensities for each array. The standard deviation of the duplicate spots on the same grid is generally only about two-thirds of the standard deviation of replicate spots (upper and lower grids). Moreover, the standard deviations among arrays are generally greater than the standard deviations within arrays (details are not shown). A nested error model incorporating variations between the duplicates within a replicate and the variations between the two replicates within an array is worth a future exploration.

2.3. Test for Treatment Effects

The main purpose of the experiment is to identify the differentially expressed genes between the control and treated samples. The one-sample *t*-test of the ratio of treated to control samples can be used for the comparison. For each gene, there are six ratios. A total of 707 comparisons (genes) is made. We set the significance level to be $\alpha = 0.001$ to take into account both false positive and false negative issues. The number of genes with *p*-values less than or equal to 0.001 is 30. Alternatively, treatment effects can be tested by comparing the six treated samples to the six control samples using the two-sample *t*-test. The number of genes having a significant *p*-value is 39. We also performed a similar analysis using $\alpha = 0.000035(.25/707)$ based on the Bonferroni adjustment. The number of genes with significant *p*-values are 16 and 21 for the one-sample and two-sample *t*-tests, respectively. We will compare the one-sample vs. two-sample *t*-tests approaches in the analysis of microarray data. The two-sample *t*-test for dye effects using the control samples is not considered since this is not an effect of interest—in addition to there being a small sample size in each dye group.

3. NORMALIZATION

In this dye-swap experiment where the dye assignments for arrays A1–A3 and A4–A6 were reversed, we paired (A1,A4), (A2,A5), and (A3,A6) for an evaluation of a dye-swap self-normalization. Figure 2 displays the plots of $(\mathrm{M} - \mathrm{M}')/2$ vs. $(\mathrm{A} + \mathrm{A}')/2$, where M and A correspond to arrays A1–A3 and M' and A' to arrays A4–A6. The red points should be centering around the dotted line corresponding to log-ratios of zero. Figure 2 shows slight bias among these points. In addition, the plots show negative biases, in particular, for the higher intensity points [large values of $(\mathrm{A} + \mathrm{A}')/2$]. However, Fig. 2 does show an improvement as compared to the control array (Fig. 1b). Different normalization methods considered location/intensity adjustments for the dye and treatment effect are presented below, respectively, by comparison between control and treated samples.

3.1. Global Normalization of Intensity Ratios

Fluorescence intensity data are obtained after image analysis using some method for local background correction. Let y_{ij} denote the base-2 logarithm of the ratio of the intensity measurements for two RNA samples for the j^{th} gene on the i^{th} array, $i = 1, \ldots, a$, and $j = 1, \ldots, g$. Consider the additive model

$$y_{ij} = m + A_i + e_{ij}$$

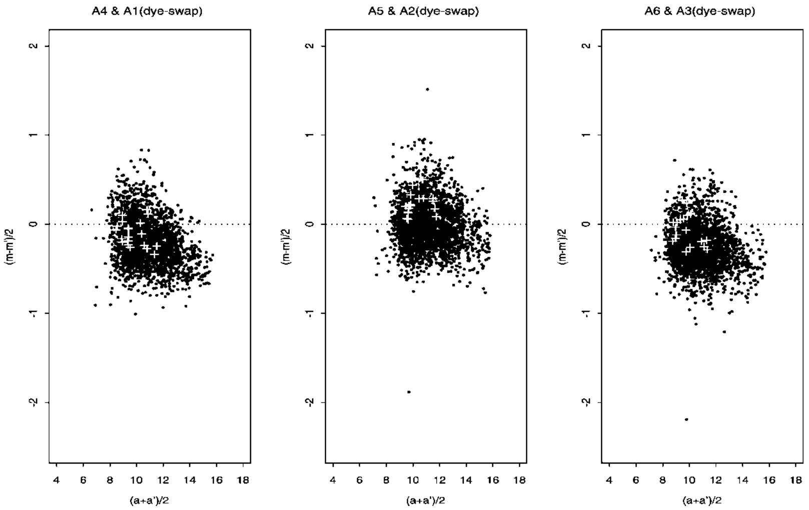

Figure 2. Graph for dye-swap before normalization.

where m is an overall effect, A_i represents the effect of the array i, and e_{ij} is the stochastic error. The additive model (with respect to the log-transformed data) refers to an additive error term which comes from the background effect and within-array variations. The array effects A_i account for overall variation in fluorescent signal from array to array. Let $\hat{m}$ and $\hat{A}_i$ be estimates of m and A_i, respectively. The residual r_{ij} from the fitted model is

$$r_{ij} = y_{ij} - \hat{m} - \hat{A}_i$$

These residuals represent the normalized values for y_{ij}, $i = 1, \ldots, a$, and $j = 1, \ldots, g$. This approach has been given by Wolfinger et al. (2001) and Draghici et al. (2001).

In the linear (additive) model given above, the parameters are conventionally estimated in terms of cell means under the normal model. An alternative is to use median estimates. However, both mean and median estimates assume constant proportionality factors in the adjustment. This approach is known as global normalization. More flexible estimates can be obtained by a nonlinear smoothing algorithm. In this paper, we adopt the Yang et al. (2001) approach using the lowess (MathSoft Inc., 1995) fit to adjust for array effects A_i.

3.2. Subset Normalization of Intensity Ratios

The approach described above considers all genes on an array simultaneously in the adjustment. A subset normalization approach is a refinement of the global normalization to account for location-dependence biases. The subset normalization consists of two steps: a) partition all genes on an array into disjoint subsets; b) perform an adjustment for each subset using either the central estimate (mean or median) or the lowess fit. Genes on an array, in principle, can be partitioned with respect to either location or intensity.

The partition of an array for location normalization can be based on contiguous rows and columns whose spotting pattern (e.g., 14×14 or 8×8) is produced by a single pin. The partition for intensity normalization can be based on the magnitudes of the intensity levels. However, the optimal number of partition subsets for intensity normalization is difficult to determine. Also, the division of the partition intervals can be either based on an equal interval length or on an equal number of genes in a partition subset. Therefore, we considered the global lowess fit using all genes on an array for the intensity adjustment.

For a given array, let s denote the number of disjoint subsets (partitions) in the array, which is based on the spotting-pattern matrix generated by a single pin with size n (e.g., $n = 14 \times 14$). Denote by $L_{i,l}$ the l^{th} subset (location) on the array i, $l = 1, \ldots, s$. The model for a subset normalization is given as

$$y_{ij,l} = m + L_{i,l} + I_i + e_{ij,l}$$

where $L_{i,l}$ represents the effect of the location l on the array i and I_i represents the effect of the intensity on the array i. The $L_{i,l}$ can be estimated either using the central estimates (mean or median) or the lowess fit. The estimate of $L_{i,l}$ can be expressed as

$$\hat{L}_{i,l} = b_i(f_l) - \hat{m}$$

For the lowess adjustment method, $b_i(f_l)$, represents the results for the subset l on the i^{th} array and f_l denotes the lowess fit sequence (with n smoothed values). For the central estimate, the f_l is replaced by the median value of the subset. Similarly, for the intensity adjustment, the estimate of I_i can be expressed as

$$\hat{I}_i = c_i(f) - \hat{m}$$

where f is a lowess fit sequence with all $(n \times s)$ values for the i^{th} array. The residuals are the normalized values of the subset normalization given as

$$r_{ij,l} = y_{ij,l} - \hat{m} - \hat{L}_{i,l} - \hat{I}_i$$

Note that a global median adjustment is applied in the final step to complete the normalization.

Our lowess smoothing method for the location estimates $b_i(f_l)$ is the same as that of Yang et al. (2001). For each subset l, the lowess function $f_l = \mathrm{f}(A)$ is fit to the $L(i, l)$ vs. $A(i, l)$ plot (A-dependent adjustment) with the 20% fraction for smoothing. For the intensity adjustment, we fit the lowess function in two ways. In addition to the A-dependent adjustment, we also fitted the lowess function $c_i(f)$ to the I_i vs. R_i plot, where R_i is the rank of spot intensity level (red and green total) on the array i. This method will be referred to as the R-dependent adjustment. The R-dependent adjustment is sometimes more preferable due to existence of a saturation-like phenomenon since the intensities of similar magnitude are locally adjusted in a neighborhood. However, the A-dependent method performs the adjustment in a pair fashion based on the total of the red and green intensities.

3.3. Subset Normalization of Individual Intensities

The additive model given above considers the ratio of intensities to model the difference of two RNA samples. Kerr et al. (2000) proposed an ANOVA model for the individual intensity measurements. Let y'_{ijkt} denote the base-2 logarithm of the intensity for the j^{th} gene on the array i in the t^{th} treatment and k^{th} dye, $i = 1, \ldots, a$, $j = 1, \ldots, g$, $k = 1, 2$, and $t = 1, 2$. Two models are considered:

$$\textbf{M1} \quad y'_{ijkt} = m + A_i + G_j + D_k + T_t + e_{ijkt}$$

and

$$\textbf{M2} \quad y'_{ijk} = m + A_i + G_j + D_k + (AD)_{ik} + e_{ijk}$$

In the M2 model, the treatment effects (T_t) are indirectly accounted for by the Array × Dye interactions where $(AD)_{ik} = T_t + (DT)_{kt} + (ADT)_{ikt}$. These two models are the modification of the Kerr's global ANOVA model. The residuals of the fitted model correspond to the Treatment × Gene interactions as the effect of interest in the context of Kerr et al. global ANOVA model (Kerr et al., 2000).

The subset normalization methods described above can be applied to individual intensities to adjust for location biases. For example, the subset normalization for the model M1 can be expressed as

$$y'_{ijkt,l} = m + L_{ikt,l} + I_{ikt} + G_j + D_k + T_t + e_{ijkt,l}$$

The residuals of the fitted model are computed in the same manner. The residuals represent the normalized intensities. Again, either one-sample or two-sample t-tests can be used to test for treatment effects from the normalized intensities. To our knowledge, the one-sample t-test has not been previously considered in this context.

4. RESULTS

The normalization methods described in the previous section were applied to comparing dye effects for two control data sets and treatment effects for the control and treated samples using the ratios of the two samples. We considered three global normalization methods, median, and A-dependent and R-dependent lowess methods. For subset normalization, we partitioned each array into 16 subsets from the 16 grids. We considered the median and A-dependent lowess normalization for location adjustment in combination with the A-dependent and R-dependent lowess methods for intensity adjustment. In addition, we also considered two methods of a subset normalization without an intensity adjustment (median and A-adjustment only). Table 1 contains the number of genes having p-values less than or equal to α from the one-sample t-test comparing dye effects ($\alpha = .05$) and treatment effects ($\alpha = .001$ and $\alpha = .00035$) for the data before normalization and for different normalized data. Note that we used the nominal level $\alpha = .05$ in testing for dye effects because of small effect sizes between the two colors.

Table 1. Number of genes with p-value less than or equal to α from the one-sample t-test for the log-ratios (base-2) of two samples (red/green for dye effects and treated/control for treatment effects) using different normalization methods.

	Dye ($\alpha = .05$)		Treatment ($A1$–$A6$)	
Methods	$A0$	($A1$–$A6$)/$A0$	$\alpha = 0.001$	$\alpha = 0.00035$
Before normalization	138[a]	15	30	16
Global normalization				
Central median	68[a]	13	138[b]	96
R-dependence	68[a]	12	138[b]	99
A-dependence	60	16	138[b]	106[b]
Subset normalization				
Median only	70[a,b]	17	152[b]	104
MR	71[a,b]	18	152[b]	107
MA	70[b]	18	158[b]	112
A-dependence only	73	21	132	88
AR	58	19	131	86
AA	72[a]	14	123	80

[a] Blank spots are significant.
[b] One bacterial gene is significant.

4.1. Dye Effects

For the dye effect, the number of significant p-values before normalization is 138 for the data from the array A0 (Column 2). All normalization methods give very similar results between 58 to 73. The results show evidence of dye effects; however, these numbers are about half of the number obtained before normalization. For the analysis of the data from (A1–A6)/A0, the number of significant p-values before the normalization is 15 (Column 3). For the normalized data, the numbers range from 12 to 21. These numbers are substantially smaller than the number obtained from the analysis of the array A0. Thus, it appears that there is a large array to array variation. The difference between the numbers of significance in the two analyses before the normalization is 123 (138–15); this number is much larger than the differences from the normalized data, from 19 to 58. Figure 3a displays the density curves for the data of the unnormalized log-intensity ratios and the normalized (AR) log-intensity ratios for the array A0. The normalized data show tighter range and more symmetric at zero than the unnormalized data. Figure 3b is the plot of M vs. A for the data after normalization. This plot, as compared to Fig. 1b, shows that the normalized data behave better than unnormalized data, where the points center around zero on the vertical axis.

4.2. Treatment Effects

For the treatment and control comparisons, the number of genes with p-values less than or equal to $\alpha = 0.001$ before normalization is 30 (Column 4) and the number for

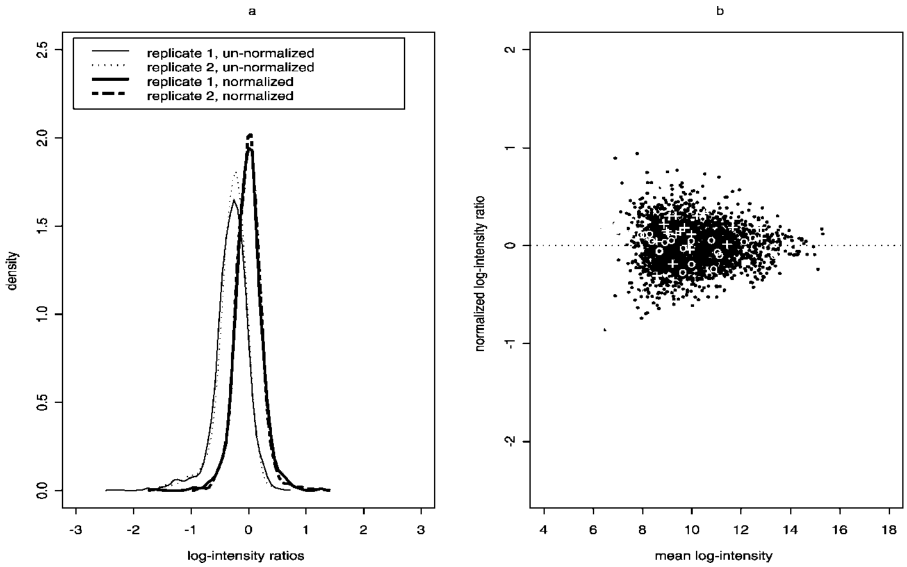

Figure 3. Graph for control vs. control data (A0) after subset normalization (AR).

$\alpha = 0.00035$ is 16 (Column 5). For the normalized data, all normalization methods generally give similar numbers of significant p-values. All methods identify at least fourfold of the number of the genes that are identified before normalization. The three A-dependent subset location normalization methods give the smallest numbers of significant p-values (132, 131, and 123) compared to the other six methods. However, of the genes that are identified from the other six methods, one is a bacterial gene. The subset location normalization without intensity adjustment method (i.e., median or A-dependence only) is presented for illustrative purposes, since a location adjustment alone is generally inadequate. The two methods, the A-dependent location with R-dependent intensity (AR) method and the A-dependent location with A-dependent intensity (AA) method, appear to perform well. The AR and AA methods identify 25 and 24, respectively, out of 30 genes that are identified before normalization. For $\alpha = 0.00035$, the numbers identified from the AR and AA methods are 86 and 80, respectively, of which 14 and 13 are out of the 16 genes that are identified before normalization. Figure 4 displays the plots of $(\mathrm{M} - \mathrm{M}')/2$ vs. $(\mathrm{M} + \mathrm{M}')/2$ for the normalized data from the AR method. The plots for the AA method are similar (not shown). As can be seen, the AR method performed well on these data.

For the analysis of treatment and control comparisons with the models M1 and M2, we considered four global normalization methods—mean, median, R- and A-dependent methods—and four subset normalization methods—median, and A-dependent normalization for location adjustment in combination with the A-dependent and R-dependent lowess methods for intensity adjustment. The mean normalization method is similar to the ANOVA approach proposed by Kerr et al. (2000), except they used a bootstrap confidence interval to test for a difference. The results are summarized in Table 2.

The M1 model shows fewer significant p-values than the M2 model in both one-sample and two-sample t-tests. The difference between the M1 and M2 models is that

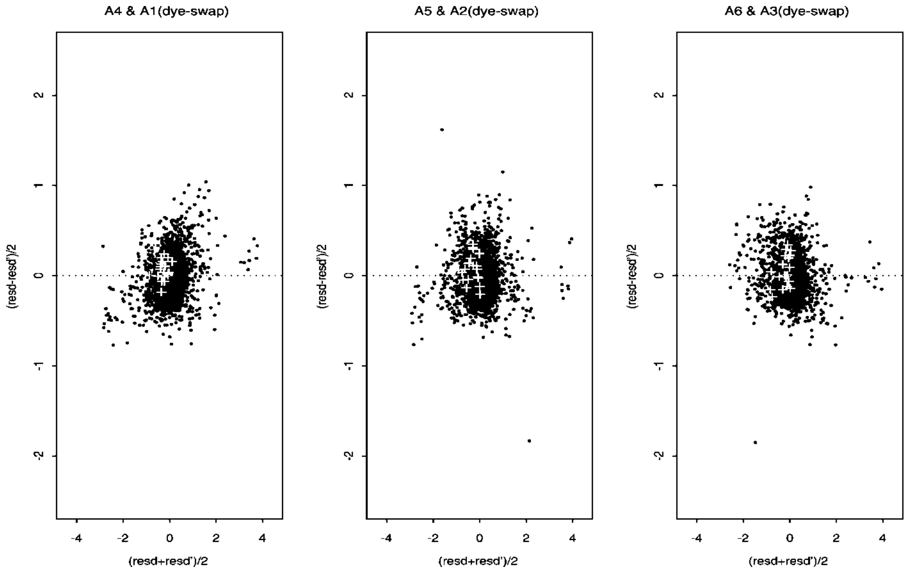

Figure 4. Graph for dye-swap after subset normalization (AR).

Table 2. Number of genes with p-value less than or equal to α from the one-sample and two-sample t-tests for different normalization methods using normalization models M1 and M2 of individual intensities.

	One-sample t-test				Two-sample t-test			
	M1		M2		M1		M2	
Methods	α_1[a]	α_2	α_1	α_2	α_1	α_2	α_1	α_2
Before normalization	30	16			39	21		
Global normalization								
Mean	27	14	143[b]	97	38	30	65	51
Median	30	14	141[b]	102[b]	40	27	60	43
R-dependence	34[c]	18	118[b]	86[b]	43	32	64	45
A-dependence	30	14	140[b]	107[b]	121[c]	101[c]	313[b]	282[b]
Subset normalization								
MR	34[c]	18	123[b]	83[b]	50	39	76	57
MA	30	14	149[b]	104	83[c]	63[c]	146[b]	121[b]
AR	34[c]	18	125[b]	78	141[b,c]	106[c]	264[b]	228[b]
AA	30	14	121	77	41	27	48	33

[a] $\alpha_1 = 0.001$; $\alpha_2 = 0.00035$.

[b] One bacterial gene is significant.

[c] Blank spots are significant.

the normalized data from the M2 model account for variations due to the dye × array interaction. For this data set, there is highly significant dye × array interaction. Therefore, M2 is a more appropriate model than M1. For the one-sample *t*-test approach, the numbers of genes identified among the eight methods are between 118 to 149 for $\alpha = 0.001$ and between 77 to 107 for $\alpha = 0.00035$. The two subset normalization methods AA and AR appear to perform consistently. The number of significant *p*-values for the two methods are 125 and 121 for $\alpha = 0.001$ and 78 and 77 for $\alpha = 0.00035$. For the two-sample *t*-test approach, the results fluctuate aberrantly. Mainly, the four A-dependent methods behave badly because the method adjusts the two measurements (red and green intensities) of the same genes in the same array in a dependent manner, which may violate the independence assumption of two-sample t-tests. For the remaining four methods (global mean, median and *R*-dependence, and subset MR) the subset normalization MR method has the most number of significant *p*-values, 76 for $\alpha = 0.001$. Of the 76 genes, 33 are among the 39 genes (about 85%) that are identified before normalization. The results for $\alpha = 0.00035$ are similar.

With regard to the two approaches to applying the one-sample test shown in Tables 1 and 2, the two approaches are very similar. Overall, the AA and AR subset normalization methods perform more consistently than other methods in both approaches. For example, for $\alpha = 0.001$ the numbers of significant *p*-values identified by the AA method are 123 and 121 (Tables 1 and 2), among which there are 108 genes in common.

4.3. Consistency

We conducted a further analysis to evaluate the two subset normalization AR and AA methods that appear to perform the best among the methods considered. We used the average of the two duplicate spots in the same grid in this analysis, as was done on array A0 in the analysis of dye effects. Each slide will have two measurements. Pair data from two arrays with a dye swap were analyzed to assess consistency. There are nine possible pairings: A1A4, A1A5, A1A6, A2A4, A2A5, A2A6, A3A4, A3A5, and A3A6. Three pairs from the six arrays A1–A6 were analyzed at a time, e.g., A1A4, A2A5, and A3A6, to evaluate the consistency in identifying (the same) differentially expressed genes for two pairs and three pairs. In the analysis, the number of significant *p*-values was computed with the one sample *t*-test on the ratio of treated to control samples for each pair and for all three pairs at $\alpha = 0.01$,.001, and.00035. Table 3 contains three analyses among nine possible combinations before normalization and AA and AR methods. The numbers shown on the last row are from Table 1 for $\alpha = .001$ and .00035. The numbers 124, 254, and 264 were obtained using the same analysis with $\alpha = .01$. Table 4 shows the distribution of genes among the nine array pairs that are significant on each pair.

For individual pairs, the results show serious fluctuation from the unnormalized data, whereas the adjusted intensities behave more uniformly. For the consistency across array pairs, the normalized data gives much more consistent results. For example, for the analysis with $\alpha = 0.01$, the number of genes with significant *p*-values for the three single-hybridization pairs, A1A4, A2A5, and A3A6, before the normalization are 295, 153, and 72, respectively, but there are only 21 in common. For the AR method, the numbers are 144, 138, and 135 with 55 in common. Although the results are not very satisfactory,

Table 3. Number of genes with p-value less than or equal to α from one-sample t-test for data of pair-chips using means of duplicates.

	Before normalization			AA method			AR method		
Data	α_0[a]	α_1	α_2	α_0	α_1	α_2	α_0	α_1	α_2
a: A1A4	295[b](32%)	74(20%)	34(15%)	149[b,c](85%)	32(72%)	7(57%)	144(89%)	27(85%)	14(67%)
b: A2A5	153[b](51%)	46(24%)	21(29%)	136[b](90%)	34(82%)	14(71%)	138[b](88%)	32(84%)	17(76%)
c: A3A6	72(67%)	15(60%)	6(50%)	137(85%)	39(74%)	17(71%)	135(90%)	37(86%)	14(86%)
a + b	73(78%)	9(56%)	5(40%)	76(100%)	9(100%)	1(100%)	79(99%)	8(100%)	3(100%)
a + c	33(88%)	6(67%)	0	77(99%)	9(100%)	1(100%)	71(100%)	8(88%)	2(100%)
b + c	47(79%)	7(86%)	2(100%)	70(99%)	12(100%)	2(100%)	79(100%)	10(100%)	1(100%)
a + b + c	21(100%)	2(100%)	0	52(100%)	6(100%)	0	55(100%)	4(100%)	1(100%)
d: A1A5	68(59%)	13(54%)	7(29%)	144[c](89%)	36[c](83%)	16(81%)	144[b](89%)	38(82%)	16(81%)
e: A2A6	86(50%)	24(38%)	9(22%)	124(87%)	30(77%)	13(69%)	132(89%)	25(84%)	11(91%)
f: A3A4	232(39%)	69(16%)	36(17%)	154(83%)	47(79%)	22(77%)	154(87%)	48(85%)	18(78%)
d + e	16(94%)	2(100%)	0	65(100%)	8(100%)	2(100%)	74(100%)	5(100%)	2(100%)
d + f	28(96%)	3(100%)	1(100%)	78(99%)	10(100%)	2(100%)	85(98%)	9(100%)	3(100%)
e + f	47(79%)	8(63%)	2(0%)	74(100%)	13(100%)	6(83%)	80(99%)	11(100%)	3(100%)
d + e + f	14(100%)	1(100%)	0	51(100%)	5(100%)	1(100%)	60(100%)	4(100%)	1(100%)
i: A1A6	24(79%)	1(0%)	1(0%)	135(85%)	33(76%)	21(76%)	142(87%)	30(73%)	18(72%)
j: A2A4	167(42%)	47(26%)	27(33%)	128[b](88%)	34(94%)	12(75%)	133(89%)	33(88%)	13(92%)
k: A3A5	171[b](57%)	46(35%)	23(22%)	164(84%)	51(75%)	23(74%)	159(89%)	50(74%)	24(71%)
i + j	10(100%)	0	0	71(100%)	8(100%)	4(100%)	81(98%)	8(100%)	3(100%)
i + k	16(100%)	1(0%)	1(0%)	82(100%)	8(100%)	1(100%)	84(99%)	8(100%)	2(100%)
j + k	75(76%)	8(75%)	5(80%)	86(99%)	20(100%)	2(100%)	87(99%)	15(100%)	2(100%)
i + j + k	10(100%)	0	0	58(100%)	6(100%)	1(100%)	63(100%)	6(100%)	2(100%)
3 replicate pairs	124	30	16	254	123	80	264	131	86

The values % refers to the percentage of genes represented in both identified by the three replicate pairs and by each single pair (and their combinations).

[a] $\alpha_0 = 0.01$; $\alpha_1 = 0.001$; $\alpha_2 = 0.00035$.

[b] One bacterial gene is significant.

[c] Blank spots are significant.

Table 4. Distribution of genes on each array pair that are significant.

	Unnormalized			AA method			AR method		
#	α_0[a]	α_1	α_2	α_0	α_1	α_2	α_0	α_1	α_2
0	210	513	600	369	561	621	383	563	629
1	180	112	68	85	62	47	77	64	34
2	135	48	27	71	37	23	62	41	22
3	61	18	6	41	14	6	38	12	11
4	43	9	2	25	13	1	31	12	2
5	37	2	2	24	5	2	16	5	0
6	23	2	1	25	4	1	26	4	2
7	9	1	0	14	2	2	18	1	0
8	4	1	0	15	2	0	15	2	0
9	5	0	0	37	3	0	40	2	1

[a] $\alpha_0 = 0.01$; $\alpha_1 = 0.001$; $\alpha_2 = 0.00035$.

the normalization does help in improving the consistency. Furthermore, the number of genes identified from a single hybridization pair data set is much less than that from the three replicate pairs (i.e., six hybridization replicates, A1–A6) used in this study. With three replicate pairs, the number increases from about 140 to 264 (bottom of Table 3) for $\alpha = .01$. We further examined those genes that were identified by the three replicate pairs and by each single pair (and their intersections). We evaluate consistency as the proportion of genes identified by both methods. For example, the number of genes with significant p-values were 144 for A1A4, of which 128 (89%) were among the 264 identified from the three replicate pairs. Moreover, for the intersection of three pairs the consistency with the 264 is 100%. Table 4 shows the frequency of genes by number of array pairs that were significant among the nine possible array pairs. Again, the AR and AA methods appear to have good sensitivity and help in providing consistent results, although we do not know which genes are truly significant without extensive further testing using independent single-gene hybridization measurements.

5. DISCUSSION

DNA microarray experiments incorporate novel technologies that have been used increasingly in biological and medical research. There are inherent biases in microarray data generated from a typical DNA array experiment. Researchers agree that data need to be normalized before a proper statistical analysis can be performed. Normalization has become an integral part of data analysis. However, different normalization methods might lead to different conclusions (Tables 1 and 2). This paper does not seek an optimal normalization method. It should not be expected to have a method to work for all data sets. The purpose is to investigate various commonly proposed normalization methods and to identify a general approach for use in normalization. To do so with actual experimental data is challenging because it is not known which genes are truly differentially expressed between treated and control samples. However, a good normalization method is expected to result in detection of more genes than an analysis based on unnormalized data, without incurring known false positives (e.g., plant and bacterial genes). Based on these criteria, we show that the global adjustment appears inadequate and a subset location normalization with a global (array) lowess fit for intensity adjustment performs well. Also, for the subset location normalization, the lowess fit performs better than median.

In the subset normalization, the location adjustment does not include the *R*-dependent fit, instead the median estimate is presented. In general these two methods are similar. The median adjustment is the limiting case of the *R*-dependent lowess fit. Furthermore, either the *R*-dependent or *A*-dependent fit is ordered according to the magnitude (*R* or *A*) of the intensity in the location normalization. It is possible that there are location biases within a partition. A remedy is to reduce the partition size. (In this case, the median and *R*-dependent fit should be very close.) Alternatively, it may be feasible to have a lowess fit according to an ordered location in the subset. For intensity normalization, on the other hand, since both *R*-dependent and *A*-dependent fits are based on the magnitude of intensity, it is not necessary to divide into several subsets. In addition, the partition of subsets for intensity normalization is somewhat arbitrary.

This paper describes the data from a two-dye system cDNA experiments. With modifications, the method can be applied to data from different types of arrays, for example, arrays with a single signal measurement such as the filter array. The *A*-dependent lowess fit can not be applied, but the method can be modified by replacing *A*, the mean log-intensity for the two samples, with the mean log-intensity for all arrays.

Kerr et al. (2000) proposed a global model for both normalization and testing. We break the analysis into two steps: 1) normalization, and 2) testing. Breaking the analysis into two steps can increase the flexibility of the use of nonconventional techniques for normalization. The main difference is that the Kerr et al. (2000) model makes assumptions about the distributions of the error terms on the unnormalized measurements, while the two-step approach makes the assumptions on the normalized measurements. We used the *t*-test to determine the number of genes that are differentially expressed between two samples. Kerr et al. (2000) proposed using the bootstrap method to compute the distribution of the *t*-statistic as an alternative method to identify differentially expressed genes.

Since more than 700 comparisons were made, simple use of a significance test without adjustment for the multiple comparison artifacts could lead to a large chance of false positive findings. We do not attempt to compute the tailed probability of the maximum *t*-statistic or the *p*-value from all possible permutations to control the family-wise error rate. Using a different significance level will result in different numbers of significant *p*-values. The conclusion that the subset location normalization combined with a global lowess fit for intensity is a good method should remain valid. Furthermore, for this data set the use of the one-sample *t*-test is recommendable rather than using the two-sample *t*-test, whether normalizing on the basis of ratios of intensities or individual intensities (see results in Tables 1 and 2). Finally, a preliminary analysis of the differentially expressed genes that are identified is presented in Thompson et al. (2002). One surprising finding is that when more conservative significance levels from $\alpha = 0.01$ to $\alpha = 0.00035$ are applied, the consistency of significant gene findings is generally somewhat reduced. The number of statistically significant genes is markedly reduced as would be expected.

Another noteworthy finding is that for any single hybridization pair, using statistics alone as a guide, the inconsistency ranges from 8% to 33% (consistency ranges from 67% to 92%, the last column of Table 3). When any four microarray hybridizations are combined, the apparent inconsistency drops to 0–2% (consistency ranges from 98% to 100%), except for one pair (the row a + c).

ACKNOWLEDGMENTS

We thank Drs. Robert Delongchamp, Cruz Velasco, James Fuscoe, and Sue-Jane Wang for comments on the manuscript. We are indebted to them for pointing out some important issues. The views presented in this paper are those of the authors and do not necessarily represent those of the U.S. Food and Drug Administration.

REFERENCES

Chen, Y., Dougherty, E. R., Bittner, M. L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* 2(4):364–374.

Draghici, S., Kuklin, A., Hoff, B., Shams, S. (2001). Experimental design, analysis of variance and slide quality assessment in gene expression array. *Drug Discov. Dev.* 4(3):332–337.

Delongchamp, R. R., Velasco, C., Evans, R., Harris, A., Casciano, D. (2002). *Adjusting cDNA Array for Nuisance Effects*. Technical Report. Jefferson, AR: National Center for Toxicological Research.

Efron, B., Tibshirani, R., Goss, V., Chu, G. (2000). *Microarrays and Their Use in a Comparative Experiment*. Preprint 37B/213. Stanford University.

Herzel, H., Beule, D., Kielbasa, S., Korbel, J., Sers, C., Malik, A., Eickhoff, H., Lehrach, H., Schuchhardt, J. (2001). Extracting information from cDNA arrays. *Chaos* 11(1):98–107.

Kerr, M. K., Churchill, G. A. (2001). Experimental design for gene expression microarrays. *Bio-Stat.* 2:183–201.

Kerr, M. K., Martin, M., Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *J. Comput. Biol.* 7(6):819–838.

Kerr, M. K., Afshari, C. A., Bennett, L., Bushel, P., Martinez, J., Walker, N. J., Churchill, G. A. (2001). Statistical analysis of a gene expression microarray experiment with replication. *Stat. Sinica* 7(6):819–838.

MathSoft Inc. *S-PLUS Guide to Statistical and Mathematical Analysis, version 3.3*. Seattle, WA, 1995.

Schuchhardt, S., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., Herzel, H. (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Res.* 28(10):e47.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9(12):3273–3297.

Thompson, K. L., Mirsky, M. L., Kadyszewski, E., Sistare, F. D. (2002). Concordance of degree of renal injury with gene expression in individual animals treated with the nephro-toxicant cisplatin. *The Toxicol.* 66(1):297.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol*.

Yang, Y. W., Dudoit, S., Luu, P., Speed, T. P. (2001). Normalization of cDNA microarray data. Bittner, M. L., Chen, Y., Dorsel, A. N., Dougherty, E. R., eds. *Microarrays: Optical Technologies and Informatics*. Proceedings of SPIE Vol. 4266., pp. 141–152.