Objective method of comparing DNA microarray image analysis systems

Edward L. Korn[1], Jens K. Habermann[2], Madhvi B. Upender[2], Thomas Ried[2], and Lisa M. McShane[1]

BioTechniques 36:960-967 (June 2004)

Many image analysis systems are available for processing the images produced by laser scanning of DNA microarrays. The image processing system takes pixel-level intensity data and converts it to a set of gene-level expression or copy number summaries that will be used in further analyses. Image analysis systems currently in use differ with regard to the specific algorithms they implement, ease of use, and cost. Thus, it would be desirable to have an objective means of comparing systems. Here we describe a systematic method of comparing image processing results produced by different image analysis systems using a series of replicate microarray experiments. We demonstrate the method with a comparison of cDNA microarray data generated by the UCSF Spot and the GenePix® image processing systems.

INTRODUCTION

Dual-label DNA microarrays allow the assessment of the relative expression or relative copy number of thousands of genes when comparing biological specimens. Briefly, two samples to be compared are labeled with two different fluorescent dyes (e.g., Cy™3 and Cy5), and the samples are mixed and allowed to cohybridize to an array spotted with nucleic acids. The hybridized slide is scanned, and fluorescence signal intensities in two channels (one channel per labeled sample) are recorded at each very small region (pixel) on the array. These pixel-level intensities are then processed into values that represent gene expression or copy number ratios. There are many steps involved in this processing in the context of cDNA gene expression microarrays (1–4). The question we address here is how one can objectively compare the quality of the ratio measurements that are produced by different image processing systems. It is useful to make such comparisons because it has been recognized that different image processing methods can make comparisons between investigations difficult (5). Note that this is a different problem than (*i*) assessing the effects of different experimental conditions on gene expression or (*ii*) assessing the different components of the variability of gene expression results. For (*i*), different arrays would be treated under different conditions, with the results being compared (6,7). For (*ii*), the results from a set of arrays treated identically would be partitioned into different sources of variability (8–10). For the problem considered here, a set of arrays treated identically are evaluated with different image processing methods, and the corresponding results are compared.

In theory, one could compare image analysis systems by comparing the gene summaries to the results of a gold standard assay applied to each gene. However, comparisons of estimated expression levels from cDNA microarrays and gold standards such as Northern blot analysis or quantitative PCR are difficult (3). In addition, it would be impractical to apply here due to cost, time, and specimen availability considerations. Thousands of genes may need to be examined to accurately assess how the image processing systems handle a variety of artifacts on images.

The approach we propose requires the availability of replicate microarray experiments: each array is analyzed by the image processing methods one wants to compare. By comparing the results both between methods within arrays and within methods between replicate arrays, and by assessing the observed variations relative to the variations between genes, we can objectively compare image processing methods. This type of comparison allows one to choose between methods based on accuracy and time-efficiency considerations. We demonstrate the approach with an empirical comparison of two popular software packages for image analysis, the UCSF Spot [University of California, San Francisco (UCSF), San Francisco, CA, USA; http://jainlab.ucsf.edu/] and GenePix® (Axon Instruments, Union City, CA, USA) using two different modes (GenePix-automatic or GenePix-manual). This paper is not intended to be a comprehensive review of these two software packages or as a comparison of the many current methods of image processing. Instead, it describes a methodology that can be used to compare image processing methods with a small number of existing arrays.

MATERIALS AND METHODS

Samples, Microarrays, and Experimental Design

The samples and microarrays used in this study were part of a larger study to assess the effect of an intervention on gene expression. For the purposes of the current study, four experiments were performed using a colorectal cancer cell line, DLD1, and customized cDNA microarrays obtained from the National Cancer Institute's Advanced Technology Center (http://nciarray.nci.nih.gov). The arrays have 9984 spots, 9128 of which correspond to genes or expressed sequence tags (ESTs). A common reference design was used in which each experimental sample was compared to universal human reference RNA (www.stratagene.com/faq/_answer/UniversalRNA.htm). In experiment I, RNA

[1]Biometric Research Branch and [2]Genetics Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

was extracted from the parental cell line three times, and each extraction was processed into a labeled sample that was cohybridized with the reference sample to a different array. Experiments II–IV differed from experiment I in that the cell culture used in each case was the parental cell line, to which three different interventions were performed. For experiments II, III, and IV, two hybridizations each were performed from a single RNA extraction. Thus, the replication for experiment I was between extractions, and the replication for each of experiments II, III, and IV was between hybridizations within an extraction (see http://www.riedlab.nci.nih.gov for details of the extraction and hybridization protocols). The three image processing approaches (GenePix-automatic, GenePix-manual, and UCSF Spot) were applied to each set of microarrays from each experiment.

Image Processing

Key components of image processing algorithms for microarray data include gridding, segmentation, and foreground and background signal intensity summary calculation. Gridding refers to the localization of rectangular patches or probe cells that contain the spots. Because the location of the cells is determined by the printing of the array, gridding can usually be done with little or no user interaction (11). Segmentation is the process of identifying the set of pixels within a probe cell that lie in the spot, or foreground, as opposed to lying in the background portion of the cell. GenePix assumes that the spot is a circular disc in the cell with the center and size of the circle being determined automatically (GenePix-automatic) or with the additional (time-consuming) interaction of the user (GenePix-manual). UCSF Spot uses a histogram algorithm in which, subject to some geometric constraints, the highest intensity pixels are assumed to be in the foreground and the lowest intensity pixels are in the background (12). To summarize foreground and background pixels, frequently one calculates the mean or median of the pixels within each of the foregrounds and backgrounds. We used as a summary measure of channel-specific intensity the mean foreground pixel intensity minus the median background pixel intensity. For our experiments, the user interventions required by GenePix-manual were performed by M.B. Upender (National Institutes of Health, Bethesda, MD, USA).

We calculated a ratio for each spot by dividing the background-corrected test channel summary intensity by the background-corrected reference channel summary intensity with the following minor modification. We flagged as "unreliable" spots for which the test and reference channel summary intensities were both less than 100. In our experience, low intensity spots are less reliable than higher intensity spots. The value of 100 for

a cut-off was based on various diagnostic plots; in particular, as a lower bound for which the M (difference in channel intensities) versus A (average of channel intensities) plots appeared horizontal for A values larger than this cut-off (http://www.BioTechniques.com/June2004/KornSupplementary.html; see Supplementary Figure S1). If the intensity was less than 100 in only one channel, then the intensity summary was set to 100 in that channel, and the ratio was calculated using the thresholded value. (Alternative methods of ratio calculation exist, such as forming ratios within each pixel followed by summarization over pixels, but we do not consider such methods here.) All ratios were transformed to the log base 2 scale for further analysis. Also, to simplify subsequent discussions, we will treat distinct spots as corresponding to distinct genes, even though some investigators may use arrays that are designed with replicate spots for some genes.

In addition to our rule for flagging spots as unreliable because of low intensity in both channels, the image analysis programs also flag certain spots as unusable. The GenePix system also allows the user to manually flag additional spots as unusable based on their image. For example, dust specs, fibers, scratches, or other contaminants or defects on a slide may be visible upon meticulous inspection of the image but may not be recognized automatically as artifacts by the image processing algorithms. The GenePix-manual flagging utility allows a user to flag such spots for exclusion from analyses.

The log ratios were normalized for each array by subtracting the median of the log ratios from the nonflagged spots on the array. These median-normalized log expression ratios were used for all analyses. Intensity-based normalization was also considered but deemed to be unnecessary (using a cut-off of 100) upon inspection of diagnostic plots (see Supplementary Figure S1). Using intensity-based normalization for the full range of intensity values would lead to unreliable normalization correction in the region of low intensity.

Version 4.0.17 of GenePix and version 2.0 of UCSF Spot were used in this study. For GenePix-automatic, we restricted the (foreground) circle size in the user option to between 33% and 200% of the nominal circle size. For UCSF Spot, the composite test/reference image was used for segmentation because this is reported to be superior over using just one of the reference or test images for segmentation (12). Additional settings used for UCSF Spot gridding and segmentation were: (*i*) no spot enhancement (because we did not use a DAPI image); (*ii*) a slow array optimization step (as recommended); (*iii*) the default of 0.5 to specify the degree of slope (real number of spot spacing units); (*iv*) the default thresholds for foreground (0.3) and background (0.1); and (*v*) the default spot size that is computed relative to the target spacing. It was not necessary to use the option of specifying target spacing and subarray spacing hints. Grid adjustments that change the coordinates of the grids were performed according to the Spot 2.0 User Manual.

Statistical Analysis Comparing Image Analysis Methods

To compare the three methods of image analysis, we estimated the reliability of each method for each experimental group using a components of variance model:

$$Y_{ij} = g_i + e_{ij} \qquad \text{[Eq. 1]}$$

where Y_{ij} is the log expression ratio for the ith spot and j^{th} replicate ($j = 1,2,3$ for group I and $j = 1,2$ for groups II–IV). The error variance component σ_e^2 associated with e_{ij} represents the reproducibility of the method. The variance component σ_g^2, associated with g_i, represents the true spot-to-spot (gene-to-gene) variability, heuristically, the "signal." The intraclass correlation, defined as $\sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$, represents the reliability of the method (13). We preferred the intraclass correlation as a measure of reproducibility over a measure such as the error variance or its square root (σ_e) alone because it guards against algorithms that produce ratio estimates all shrunk to a central value. For example, if one were to apply an image processing method that reported the value 1.0 for every ratio, the method would have perfect reproducibility, yet the ability to distinguish among the genes would be lost. To make a fair comparison, only the spots that were not flagged by any of the methods on any of the replicate arrays within a given experimental set (I–IV) were used for calculating the variance components and intraclass correlations. This permits the estimation of the reliability of the methods on unflagged spots separately from the assessment of the quality of spot-flagging decisions. Because the spot-to-spot gene variability (σ_g^2) depends on the expression pattern of genes on the array, extrapolation of the reliability from one type of experiment or array should be done with caution. In particular, spiking of some spots to yield large

Table 1. Setup for Evaluation of an Image Processing Approach for Microarrays

	Array in Experiment I		
Image Analysis Method	1	2	3
Method 1	$Y_1^{(M1)}$	$Y_2^{(M1)}$	$Y_3^{(M1)}$
Method 2	Flagged	$Y_2^{(M2)}$	$Y_3^{(M2)}$

Evaluation of image processing approach (method 1) for a spot flagged by another image processing approach (method 2).

Table 2. Estimated Variance Components of Log Expression Ratios Produced by Three Image Processing Methods for Microarrays

Method	Between-Spot [$\hat{\sigma}_g^2$]	Within-Spot (between replicates) [$\hat{\sigma}_e^2$]	ICC
Experiment I (n_a = 3, n_g = 8478)			
UCSF Spot	0.618	0.081	0.88
GenePix-automatic	0.730	0.102	0.88
GenePix-manual	0.731	0.101	0.88
Experiment II (n_a = 2, n_g = 8080)			
UCSF Spot	0.728	0.233	0.76
GenePix-automatic	0.902	0.217	0.81
GenePix-manual	0.900	0.215	0.81
Experiment III (n_a = 2, n_g = 8735)			
UCSF Spot	0.627	0.146	0.81
GenePix-automatic	0.683	0.148	0.82
GenePix-manual	0.684	0.146	0.82
Experiment IV (n_a = 2, n_g = 8679)			
UCSF Spot	0.681	0.122	0.85
GenePix-automatic	0.850	0.124	0.87
GenePix-manual	0.848	0.125	0.87

Only spots not flagged as unusable by any of the methods are included in the analysis. ICC, intraclass correlation.

Table 3. Analysis of Discrepancies Greater Than or Equal to 2× Between UCSF Spot and GenePix-Manual Image Processing Software for Microarrays

Experiment	Spots (No.)	Spots with Replicate Agreement[a] (No.)	Categorization[b]		
			UCSF Spot Incorrect	GenePix-Manual Incorrect	Not Determined
I	36	22	15	1	6
II	28	19	12	4	3
III	251	95	43	13	39
IV	46	39	25	3	11

[a]Spots are included if the agreement of the methods on the other array(s) is good; that is, within 1.25×.
[b]See text for the categorization rule.

expression values is not recommended because this will artificially inflate the intraclass correlation.

Formulas for computing the variance components and intraclass correlation are as follows. The error (within-gene, between replicate arrays) variance component is estimated by

$$\hat{\sigma}_e^2 = \sum_{i=1}^{n_g} \sum_{j=1}^{n_a} (Y_{ij} - \bar{Y}_{i.})^2 / [n_g (n_a - 1)] \qquad \text{[Eq. 2]}$$

where n_a = number of replicate arrays, n_g = number of genes, and $\bar{Y}_{i.} = \sum_{j=1}^{n_a} Y_{ij} / n_a$. (Note that $n_a = 3$ for experiment I, and $n_a = 2$ for experiments II–IV.) The between-gene variance component is estimated by

$$\hat{\sigma}_g^2 = \sum_{i=1}^{n_g} (\bar{Y}_{i.} - \bar{Y}_{..})^2 / (n_g - 1) - \hat{\sigma}_e^2 / n_a \qquad \text{[Eq. 3]}$$

where $\bar{Y}_{..} = \sum_{i=1}^{n_g} \sum_{j=1}^{n_a} Y_{ij} / (n_g n_a)$. The estimated intraclass correlation (ICC) is

$$\text{ICC} = \hat{\sigma}_g^2 / (\hat{\sigma}_g^2 + \hat{\sigma}_e^2). \qquad \text{[Eq. 4]}$$

R computer code to estimate the variance components and ICC are available at http://www.BioTechniques.com/June2004/KornSupplementary.html.

The components of the variance model specifies that the error variance (σ_e^2) is constant across spots. This may be an unrealistic assumption, as the variance could be a function of the mean level of expression (14). However, even if the constant-variance assumption is not satisfied, the estimator $\hat{\sigma}_e^2$ still estimates a meaningful quantity, the average error. Formally, if σ_i^2 equals the variance of e_{ij}, then $\hat{\sigma}_e^2$ estimates $\frac{1}{n_g} \sum_{i=1}^{n_g} \sigma_i^2$ (and $\hat{\sigma}_g^2$ still estimates σ_g^2). Therefore, since the image processing methods are being compared on the same set of spots on the same set of arrays, comparisons of intraclass correlations between the methods are meaningful even if σ_e^2 is not constant across spots. We note, however, that there is no absolute guarantee that the intraclass correlation will correctly assess the reliability of a method. For example, a method that consistently overestimates the expression ratio for genes with high expression ratios and underestimates the expression ratio for genes with low expression ratios would appear to have better reliability that it actually does.

Of additional interest are spots in which the methods gave very discrepant estimates. To examine this, we considered spots in which two methods produced ratios that differed by more than a factor of two on a spot but whose estimates corresponding to a replicate spot(s) in the same experimental group were close to each other (within 25%). In this way, we could attempt to evaluate which method was doing better by being closer to the average of the replicate values. In particular, let $Y_1^{(Mi)}$ be the log expression ratio as determined by method 1 for a particular spot on array 1. Then if $|Y_1^{(M1)} - Y_1^{(M2)}| > 1$, the methods are discrepant by a factor of 2 for this spot on array 1. For experiments II–IV, let $Y_2^{(Mi)}$ be the log expression ratio as determined by method 1 for the same spot on array 2, corresponding to the replicate array in the experiment. Then if $|Y_2^{(M1)} - Y_2^{(M2)}| < \log_2(1.25)$, we consider the methods to be in agreement on the replicate array. Letting $AVE = (Y_2^{(M1)} + Y_2^{(M2)}) / 2$, we categorize method 1 as giving the incorrect value for this spot (on array 1) if $|Y_1^{(M1)} - AVE| - |Y_1^{(M2)} - AVE| \geq \log_2(1.5)$, whereas we categorize method 2 as giving the incorrect value for this spot if on the replicate arrays $|Y_1^{(M2)} - AVE| - |Y_1^{(M1)} - AVE| \geq \log_2(1.5)$; otherwise the categorization is not determined. For experiment I, there are two replicate arrays; for example, array 2 and array 3. We consider the methods to be in agreement for the spot in question on the replicate arrays if $(|Y_2^{(M1)} - Y_2^{(M2)}| + |Y_3^{(M1)} - Y_3^{(M2)}|) / 2 < \log_2(1.25)$. The categorization of the values on array 1 proceeds as with experiments II–IV, except now $AVE = (Y_2^{(M1)} + Y_2^{(M2)} + Y_3^{(M1)} + Y_3^{(M2)}) / 4$. There is no absolute guarantee that this categorization is correct. For example, it is possible (but highly unlikely) that $|Y_1^{(M1)} - Y_1^{(M2)}| > 1$, $|Y_1^{(M1)} - AVE| - |Y_1^{(M2)} - AVE| \geq \log_2(1.5)$, but method 2 is giving the wrong results on array 1 if all of $Y_1^{(M2)}$, $Y_2^{(M1)}$, and $Y_2^{(M2)}$ are giving bad results in the same direction by about the same amount.

We also considered how well one method (e.g., method 1) was doing on an unflagged spot when another method (e.g., method 2) flagged that spot as unusable on, for example, array 1 (see Table 1). In particular, we attempted to evaluate whether the other method was giving "bad" values for this spot. This evaluation was done by using the three arrays in group I and comparing $Y_1^{(M2)}$ with $(Y_2^{(M2)} + Y_2^{(M2)}) / 2$ when $|Y_2^{(M1)} - Y_3^{(M1)}| < \log_2(1.25)$. A large value of $\text{DIF} = Y_1^{(M1)} - (Y_2^{(M1)} + Y_3^{(M1)}) / 2$ suggests that this spot should have been flagged by method 1 also. We examined the distribution of DIF and compared it to the distribution of DIF calculated on spots that were not flagged.

RESULTS

The images of the nine arrays used are in the supplementary figures. Table 2 presents the estimated variance components for the four experimental groups, restricted to those spots that were not flagged by any of the methods. The reliabilities of the three methods are high, with the GenePix methods appearing on average to be perhaps slightly more reliable than UCSF Spot. Note that this is true even though the within-spot variability $\hat{\sigma}_e^2$ (the "noise") can be smaller for UCSF Spot; in these cases, the between-spot variability (the "signal") was also smaller for UCSF Spot. Interestingly, the reliability in experimental group I did not appear worse (and was perhaps better) than the other groups, even though this experiment involved three separate RNA extractions.

Table 4. Reliability of UCSF Spot Value When GenePix-Manual Is or Is Not Flagged for that Spot in Microarray Experiment I

UCSF Spot value off by a factor of[a]	GenePix-Manual Flagged for Spot			
	Yes (n = 497)			No (n = 16,563)
	Flagged as bad (n = 30) (%)	Flagged as not found (n = 307) (%)	Unreliable[b] (n = 160) (%)	(%)
<0.5× (too small)	10.0	1.0	0.0	0.4
0.5–0.67×	3.3	2.6	2.5	3.4
0.67–0.8×	6.7	8.5	8.8	10.5
0.8–1.25×	36.7	65.1	80.6	71.8
1.25–1.5×	0.0	10.7	6.2	10.7
1.5–2.0×	23.3	6.5	1.9	2.9
>2.0× (too large)	20.0	5.5	0.0	0.3
Total	100.0	100.0	100.0	100.0

[a]Spots were included for analysis if the values on the two replicate arrays were within 1.25× of each other using UCSF Spot.
[b]After GenePix-manual gridding and segmentation, the spot had both background-adjusted channel intensities less than 100.

Comparing UCSF Spot with GenePix-manual, we note that over the nine arrays, there are 361 spots (0.4%) for which the log expression ratios differ by more than a factor of 2 (see Table 3 for a breakdown by experiment). For the 116 spots for which we were able to categorize as to which method was giving an incorrect value (as described earlier), UCSF Spot gave an incorrect value 82% of the time and GenePix-manual gave an incorrect value 18% of the time. Figure 1 displays the images of eight spots for which UCSF Spot gave an incorrect value (two from each experiment, with the largest discrepancy between the methods). Although one cannot be completely sure of why UCSF Spot is giving an incorrect value for these spots, the figures offer some clues. In Figure 1, A, C, and E, the bright speck is very green and was not considered as part of the foreground by GenePix; this can be seen from pixel-level data (data not shown). If these specks were being considered in the foreground by UCSF Spot, this would explain why the UCSF Spot values were biased toward green. Similarly, the specks in Figure 1, D and H, are red and were not considered in the foreground for GenePix. For the other spots in Figure 1, things are less clear. In Figure 1B, there are no specks, but there is a red haze from a neighboring spot, in Figure 1G, the speck is red (despite the bias being too green), and in Figure 1F, there are no specks.

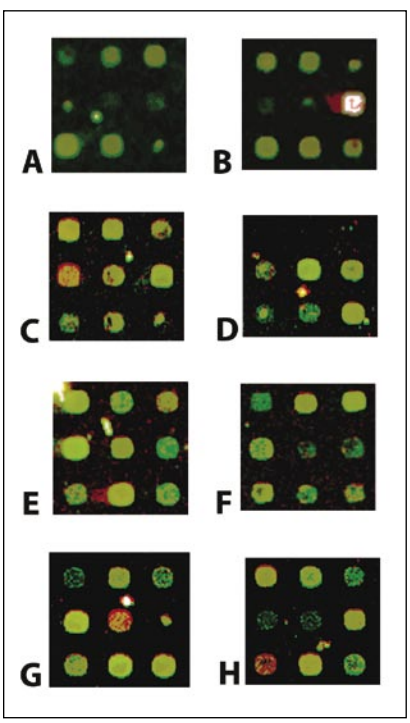

Figure 1. Microarray spots for which UCSF Spot estimates of log ratios are incorrect. An example of spots incorrectly assessed by UCSF Spot (extreme cases from Table 3). Each spot in question is in the middle of each image shown. Spots in A, C, E, and G are biased toward green by UCSF Spot. Spots in B, D, F, and H are biased toward red by UCSF Spot. The spots in the four rows correspond to the four experiments [I (A and B), II (C and D), III (E and F), and IV (G and H)].

Figure 2 displays the images of six spots for which GenePix-manual gave an incorrect value. There are two spots from each experiments II and IV, with the largest discrepancy between the methods and one spot from each of the experiments I and III. (There is only one array for experiment III because all 13 "incorrect" spots were biased toward green.) The only spot in which it is obvious what GenePix is doing wrong is Figure 2E, where the speck is red and was considered in the foreground. On the other hand, the small specks in Figure 2, D and F, should not be influencing the results. We suspect in some of these cases, GenePix is actually doing a reasonable job, but by (rare) chance, is poorly agreeing with the other arrays. For example, in Figure 2B, the speck is red but was not included in the foreground of GenePix. It is in the background of GenePix but does not influence the median background for this spot. UCSF Spot is giving a redder value for this spot, which agrees better with the other arrays for this spot.

Comparing GenePix-automatic with GenePix-manual, there were 11 spots over the arrays for which the log expression ratios differed by more than a factor of two, of which 10 spots had replicate agreement on the other arrays. Using our categorization, for 9 of these 10 spots, GenePix-automatic was giving an incorrect value, and for 1 spot, the categorization was indeterminate.

Table 4 displays the distribution of our estimate (DIF) of how far off UCSF Spot is when GenePix-manual had flagged the spot as unusable (only using experiment I data). The spot can be flagged as bad by the user or "not found" (code = -50) by the program after the user-guided manual gridding and segmentation. We have also included spots as flagged when we consider them unreliable (i.e., when their background-adjusted

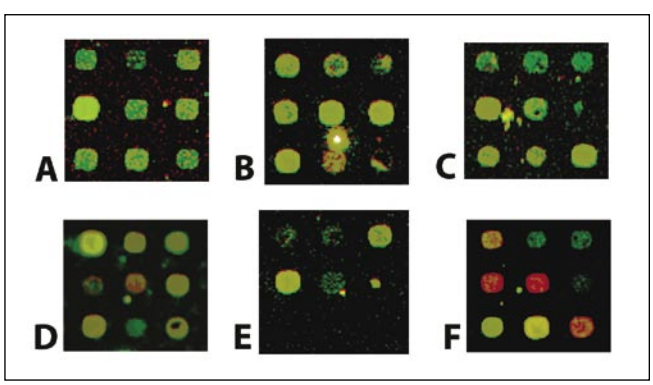

Figure 2. Microarray spots for which GenePix estimates of log ratios are incorrect. An example of spots incorrectly assessed by GenePix-manual (extreme cases from Table 3). Each spot in question is in the middle of the image shown. Spots in A, B, and C are biased toward green by GenePix-manual. Spots in D, E, and F are biased toward red by GenePix-manual. The spots correspond to the four experiments: I (D), II (C and F), III (A), and IV (B and E).

Table 5. Reliability of Low-Intensity GenePix-Manual Values Using Two Different Methods for Assigning Values to Unreliable Spots in Microarray Experiment I

	GenePix-Manual Flagged for Spot		
	Unreliable[a] (n = 135)		No flag but moderately low intensity[d] (n = 3513)
GenePix-manual value off by a factor of[b]	Flagged spot assigned value 1.0 (%)	Flagged spot given ratio using nontruncated signals[c] (%)	(%)
<0.5× (too small)	2.2	4.4	0.6
0.5–.67×	41.5	5.2	5.2
0.67–0.8×	31.9	13.3	11.8
0.8–1.25×	24.4	58.5	67.2
1.25–1.5×	0.0	11.1	11.5
1.5–2.0×	0.0	7.4	3.3
>2.0× (too large)	0.0	0.0	0.4
Total	100.0	100.0	100.0

[a]After GenePix-manual gridding and segmentation, the spot had both background-adjusted channel intensities less than 100.
[b]Spots were included for analysis if the values on the two replicate arrays were within 1.25× of each other using GenePix-manual.
[c]Background-adjusted channel intensities were less than or equal to 0 for 5 spots and were set to 1.0 for the calculation of the ratio.
[d]Background-adjusted intensities were between 100 and 500 for both channels.

intensities are both less than 100, again after the manual gridding and segmentation). Overall, GenePix-manual flagged 0.2%, 1.8%, and 0.9% of the spots as bad, not found, or unreliable, respectively. As might be expected, when spots are flagged as bad by GenePix-manual, their UCSF Spot values tend to be off more than when they are not flagged. For example, 30% of the spots flagged as bad had their UCSF Spot values differ by more than a factor of two from the estimated correct value, while only 0.7% of the UCSF Spot values were off by this much for unflagged spots. This suggests that the manual flagging of spots as bad is worthwhile. Note that the absolute numbers of spots off by a factor of two is larger for the unflagged spots than the flagged ones (116 = 0.7% × 16,563 versus 9 = 30% × 30). Thus, flagging only picks up a small percentage of spots with discrepant values. For spots flagged as not found in Table 4, there are again a higher percentage of discrepant values than when the spots are not flagged, but the results are not as dramatic as for those spots that are flagged as

bad. For spots that are flagged as unreliable, the values are less discrepant than those that are not flagged. Thus, UCSF Spot is giving reasonable values for these spots that are not being used by the GenePix-manual. This suggests that our rule for deeming spots unreliable (<100 in both channels) may not have been optimal; we return to this point below.

The distribution of discrepancies of UCSF Spot values (GenePix-automatic values) based on whether or not the spot was flagged by GenePix-automatic (GenePix-manual) are given in the Supplementary Tables, S1 and S2, respectively. Overall, 0.04% of the spots were flagged as unreliable with UCSF Spot (<100 in both channels); all of these spots were also flagged by GenePix-manual and GenePix automatic.

To investigate whether our rule for deeming spots unreliable could be modified so as to allow for more unflagged spots by GenePix, we tried two additional methods for assigning values to such spots. The first method was to assign such spots the ratio 1.0. A rationale for this assignment is that if a spot has very low intensity in both channels, it probably does not have much differential gene expression between the test and reference samples. Also, this method is somewhat consistent with the way spots that have only one channel less than 100 are handled (the channel value is set to 100), so that setting both channels to 100 yields a ratio of 1.0. The second method we considered was to use the observed ratio of background-adjusted channel intensities no matter how small they were (nonpositive background-adjusted channel intensities were set to 1.0). To evaluate how well these two methods worked, we compared the assigned value with the average value of two replicate arrays (Table 5). For comparison, the agreement of moderately low intensity spot values was included in Table 5. Neither method produced values that agreed well with the replicates. Therefore, we do not recommend these methods without further modifications.

Of importance is the time required by the investigator to assist in image processing the arrays. UCSF Spot and GenePix-automatic each took about 10–15 min per array. With GenePix-manual, it took about 15–30 min per array to flag spots as bad, but 4–8 h per array to perform the detailed circle segmentation.

DISCUSSION

In theory, a histogram segmentation algorithm such as the one used in UCSF Spot should yield more reliable log expression ratios than an algorithm using circles such as GenePix (12). However, because the specific algorithms are quite complex and microarray image data can be noisy in many different ways, we believe there is no substitute for head-to-head comparisons. In fact, Jain et al. (12) would seem to agree with one caveat: "Ideally quantitative accuracy could be assessed by direct comparison with other methods. However, because the methods available to us require significant user-specific interaction, it is difficult to make formal comparisons meaningful." Here we have shown that it is possible to make such formal comparisons and, apparently, advisable. While Jain et al. (12) found UCSF Spot superior to algorithms such as those used in GenePix, we found both procedures to be reliable, with GenePix performing slightly better on average in our study. Our comparisons also allowed us to note that very low intensity spots were handled better by UCSF Spot. We also found that the user's effort required by GenePix-manual for gridding and segmentation was very large and not useful. The effort involved in manually flagging bad spots was not extensive, however, and was useful. We decided to use GenePix-automatic with the manual flagging of bad spots (but not manual segmentation) for our larger experiment. However, an interface for manual flagging of bad spots within UCSF Spot is under development (T. Tokuyasu, personal communication). In summary, a direct comparison of various image processing software on a limited number of arrays allowed us to choose an efficient and accurate method for experimentation.

REFERENCES

1. **Hegde, P., R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J.E. Hughes, E. Snesrud, et al.** 2000. A concise guide to cDNA microarray analysis. BioTechniques *29:548-562*.
2. **Yang, Y.H., M.J. Buckley, and T.P. Speed.** 2001. Analysis of cDNA microarray images. Brief. Bioinform. *2*:341-349.
3. **Yang, Y.H., M.J. Buckley, S. Dudoit, and T.P. Speed.** 2002. Comparison of methods for image analysis on cDNA microarray data. J. Comput. Graph. Stat. *11*:108-136.
4. **Simon, R.M., E.L. Korn, L.M. McShane, M.D. Radmacher, G.W. Wright, and Y. Zhao.** 2003. Design and Analysis of DNA Microarray Investigations. Springer-Verlag, New York.
5. **Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, et al.** 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat. Genet. *29*:365-371.
6. **Wildsmith, S.E., G.E. Archer, A.J. Winkley, P.W. Lane, and P.J. Bugelski.** 2001. Maximization of signal derived from cDNA microarrays. BioTechiques *30*:202-208.
7. **Puskas, L.G., A. Zvara, L. Hackler, Jr., and P. Van Hummelen.** 2002. RNA amplification results in reproducible microarray data with slight ratio bias. BioTechniques *32*:1330-1340.
8. **Kerr, M.K., M. Martin, and G.A. Churchill.** 2000. Analysis of variance of gene expression microarray data. J. Comput. Biol. *7*:819-837.
9. **Spruill, S.E., J. Lu, S. Hardy, and B. Weir.** 2002. Assessing sources of variability in microarray gene expression data. BioTechniques *33*:916-923.
10. **Brown, J.S., D. Kuhn, R. Wisser, E. Power, and R. Schnell.** 2004. Quantification of sources of variation and accuracy of sequence discrimination in a replicated microarray experiment. BioTechniques *36*:324-332.
11. **Marzolf, B. and M.H. Johnson.** 2004. Validation of microarray image accuracy. BioTechniques *36*:304-308.
12. **Jain, A.N., T.A. Tokuyasu, A.M. Snijders, R. Segraves, D.G. Albertson, and D. Pinkel.** 2002. Fully automated quantification of microarray image data. Genome Res. *12*:325-332.
13. **Fleiss, J.L.** 1986. The Design and Analysis of Clinical Experiments. John Wiley & Sons, New York.
14. **Miller, L.D., P.M. Long, L. Wong, S. Mukherjee, L.M. McShane, and E.T. Liu.** 2002. Optimal gene expression analysis by microarrays. Cancer Cell *2*:353-361.

Received 10 March 2004; accepted 3 May 2004.

Address correspondence to:

Edward L. Korn
Biometric Research Branch, EPN-8129
National Cancer Institute
National Institutes of Health
Bethesda, MD 20892, USA
e-mail: korne@ctep.nci.nih.gov