Journal of Cellular Biochemistry 80:192–202 (2000)

Analyzing High-Density Oligonucleotide Gene Expression Array Data

Eric E. Schadt,[1,2]* Cheng Li,[3] Cheng Su,[4] and Wing H. Wong[3]

[1]Department of Biomathematics, University of California, Los Angeles, California 90095
[2]Departments of Bioinformatics and Biomathematics, Roche Bioscience, Palo Alto, California 94304
[3]Department of Statistics, University of California, Los Angeles, California 90095
[4]Department of Biomathematics, Roche Bioscience, Palo Alto, California 94304

Abstract We have developed methods and identified problems associated with the analysis of data generated by high-density, oligonuceotide gene expression arrays. Our methods are aimed at accounting for many of the sources of variation that make it difficult, at times, to realize consistent results. We present here descriptions of some of these methods and how they impact the analysis of oligonucleotide gene expression array data. We will discuss the process of recognizing the "spots" (or features) on the Affymetrix GeneChip® probe arrays, correcting for background and intensity gradients in the resulting images, scaling/normalizing an array to allow array-to-array comparisons, monitoring probe performance with respect to hybridization efficiency, and assessing whether a gene is present or differentially expressed. Examples from the analyses of gene expression validation data are presented to contrast the different methods applied to these types of data. J. Cell. Biochem. 80:192–202, 2000. © 2000 Wiley-Liss, Inc.

The use of microarray technologies to monitor gene expression in model organisms, cell lines, and human tissues has become an important part of biological research over the last several years [Wodicka et al., 1997; Der et al., 1998; Alon et al., 1999]. Teasing apart biochemical pathways, identifying genes responsible for a particular phenotype, and assessing the effect of a drug compound on the expression levels of any number of genes have all benefited from expression array technology. While many of these early successes clearly demonstrate the importance of this technology, the experiments have centered on profiling simple model organisms or laboratory cell lines. Gene expression experiments performed at Roche Bioscience (RBS) have exhibited a more complicated variation structure (with respect to the expression intensities) when profiling more complex samples, such as human and murine tissue samples. The more complicated variation structure is most probably due to the genetic and environmental heterogeneity of these more complex samples.

Given this more complicated variation structure, we found it useful to enhance the methods used to analyze the GeneChip probe array data to account for as much of the technology variation possible. While we have found that the methods used to analyze expression data provided by the GeneChip software often yield high-quality results, the false positive and false negative gene presence/differential expression call rates, we realized in a portion of replicate human and murine GeneChip expression experiments, could be improved through the development of our own methods to analyze expression array data. Given the scarcity of many tissue samples and the small size of many of the mouse tissue samples, the number of hybridizations that can be done for any given experiment is often less than optimal. The small number of array hybridizations for many of our experiments (e.g., having only one or two samples per time point, which makes it difficult, if not impossible, to estimate the within-time-point biological variation), while useful when looking at one or two genes, is problematic when looking at thousands of genes simultaneously, and really demands developing more sophisticated algorithms to further re-

Research was supported in part by a grant to UCLA from the Roche Frontiers Program.

*Correspondence to: Eric Schadt, Roche Bioscience, 1401 Hillview Avenue, Palo Alto, CA 93404. E-mail: eric.schadt@roche.com

Received 29 July 1999; Accepted 13 October 1999

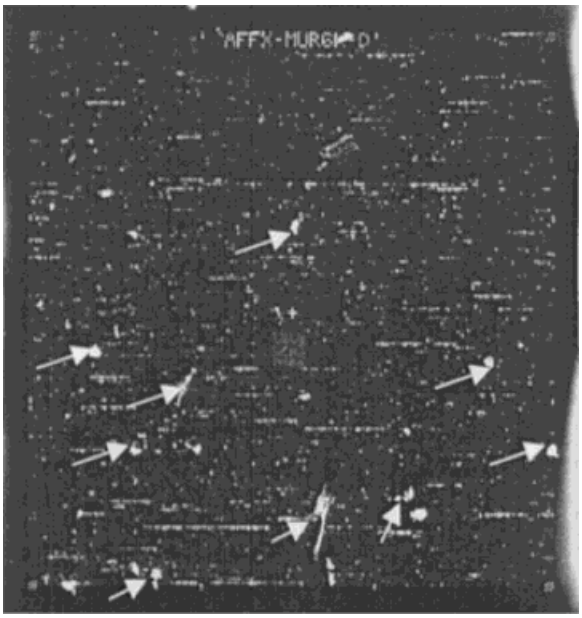

Fig. 1. A contaminated D array from the Murine 6500 Affymetrix GeneChip® set. Several particles are highlighted by arrows and are thought to be torn pieces of the chip cartridge septum, potentially resulting from repeatedly pipetting the target into the array.

duce the signal variation within and between arrays. Furthermore, a portion of the arrays we have analyzed had noticeable signal anomalies, which included intensity gradients (bright edges and fluorescing streaks), glue smears (broad fluorescing strokes resulting from the chip packaging process), and dark spots (regions where the signal is artificially low). Figure 1 illustrates some of these problems. We have found that the normalization and background correction methods currently available to analyze probe array data can be enhanced to better account for such problems, and that many of the underlying assumptions on which these methods depend do not hold in a significant percentage of the experiments we have analyzed. Finally, we have found it useful to supplement the gene detection and differential expression detection methods employed by the GeneChip software with our own methods, to make these results easier to interpret at the biological level and to provide a more quantitative measure of significance on whether a gene is present or differentially expressed.

We will discuss in more general terms some of the methods and tools we have developed to facilitate the analysis of GeneChip data; methods aimed at reducing variation at a variety of sources, variation that serves only to obscure the very biological variation we are actually interested in detecting. We will begin with a brief overview of the oligonucleotide expression array technology developed by Affymetrix, and then proceed to describe each of the low-level analysis methods we have found useful in analyzing gene expression array data.

OVERVIEW OF THE OLIGONUCLEOTIDE EXPRESSION ARRAY TECHNOLOGY

There are several publications discussing the fundamentals of the oligonuceotide expression array technology [see, e.g., Lockhart, 1996, or the supplement to Nature Genetics, Volume 21, January, 1999]. However, for our purposes in this article, it will be useful to review some of the elements of the probe array analysis provided by the GeneChip® software. As described by Lockhart et al. [1996], genes are represented on a probe array by some number of sequences (typically 20) of a particular length (typically 25 nucleotides) that uniquely identify the genes and, ostensibly, have relatively uniform hybridization characteristics, with respect to the experimental protocol used in these experiments. Each oligonucleotide, or probe, is synthesized in a small region (the length and width of the features are either 50 μm for the low-density arrays or 24 μm for the high-density arrays), which can contain anywhere from 10^6 to 10^7 copies of a given probe. Designed to correspond to the perfect match (PM) oligonucleotide pulled from a gene sequence (or EST), is a mismatch (MM) oligonucleotide in which, typically, the center base position of the oligo has been mutated; the MM probes give some estimate of the random hybridization and cross hybridization signals, although, as we can see in Figure 2, there is a nonlinear functional relationship between the paired PM and MM probe intensities.

Ostensibly, this functional relationship stems from the hybridization kinetics of the different probe sequences and from nonspecific RNA hybridizations. Figure 2 illustrates a hypothetical tiling pattern of probes pulled from a gene sequence, the length of the probes, and how each PM probe is paired with a corresponding MM probe, and the intensity differential between PM and MM features when a gene is present in a sample (i.e., high intensity for the designed perfect match probes, low intensity for the corresponding mismatch

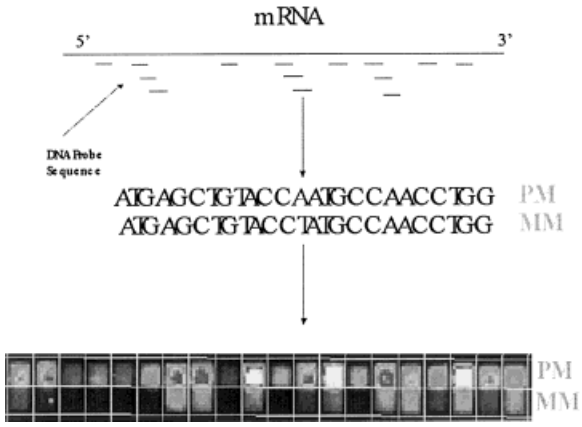

Fig. 2. Hypothetical arrangement of oligonucleotides selected to interrogate a single gene transcript (top). The perfect match (PM) and mismatch (MM) probes designed to correspond to a gene are synthesized in adjacent features (middle of figure). The intensity plot represents the sort of hybridization intensities we see for genes that are present at a moderately high abundance (bottom). Note the functional dependency of the MM intensity on the PM intensity; further note that, as expected, this functional dependency is not linear with respect to PM intensity.

probes). RNA samples are prepared according to the protocol defined by Lockhart [1996], and then the labeled RNA sample is hybridized to the corresponding probes on the array. The array then goes through an automated staining/washing process using the Affymetrix fluidics station, and upon completion of this process, the array is scanned using the Affymetrix confocal laser scanner. The scanner generates an image of the array by exciting each feature with its laser, detecting the resulting photon emissions from the fluorescently labeled RNA that has hybridized to the probes in the feature, and converting the detected photon emissions into a 16-bit intensity value. The images generated by the scanner are then ready for analysis. We can determine whether a gene is present and the quantity at which it is present by examining various statistics formed from the PM/MM feature intensities. Most of the statistics used are based on PM/MM differences (e.g., the average difference intensity and the positive fraction and positive/negative fraction statistics described later) and the PM/MM ratio (e.g., the average log-ratio). When a gene transcript is actually present, one would expect the PM intensities to be significantly greater than the MM intensities, which would be reflected in the PM/MM differences, ratios and associated statistics.

The GeneChip software supplied by Affymetrix to process array images from the scanner performs all of the fundamental operations necessary to analyze an array, including (1) image segmentation, (2) background correction, (3) scaling/normalizing arrays for array-to-array comparisons, (4) calculation of statistics to indicate whether a gene transcript is present, and (5) calculation of statistics to indicate whether a gene transcript is differentially expressed. As will be detailed by Schadt et al. [1999], we have developed and implemented our own algorithms for each of the operations listed above. We will discuss many of these methods in a less technical manner throughout the remainder of this article. For a more detailed description of these methods and for a more exhaustive comparison of these methods with currently available ones, refer to Schadt et al. [1999].

Computing Reliable Feature Intensities

Image Segmentation. In analyzing gene expression array data generated by the Affymetrix GeneChip® technology, perhaps the simplest operation to perform is that of segmenting the image. The GeneChip software employs a dynamic gridding algorithm to segment the image and then uses a percentile algorithm to compute the feature intensities once the feature boundaries have been determined [Lockhart et al., 1996]. We describe, in Schadt et al. [1999], our own image-processing algorithm. Employing our own image segmentation algorithm allowed us to directly analyze the distribution of pixel intensities for a given feature, devise new algorithms to compute feature intensities, and directly estimate the blurring effects that can affect probe intensity calculations. The image files generated by the Affymetrix GeneChip® scanner are 16-bit, binary image files, with header information prepended as described in the GATC specification [GATC Consortium, 1998]. Except for anomalous examples (e.g., when the laser in the Affymetrix GeneChip® scanner is not properly aligned or when the image is extremely bright), we have found it straightforward to compute robust feature intensity estimates for a probe array. Aligning the basic grid to an image to determine the feature locations is greatly simplified because the arrays contain alignment features at each corner of the image (these features can be seen in Fig. 1), which, when

used in conjunction with the known feature sizes, can be used to compute the locations of each feature.

Once the basic grid has been aligned, we allow the grid to "deform" at each feature location when computing the intensity of the signal at that location. Toward this end, we have implemented an adaptive pixel selection algorithm. At current scanner resolutions and feature sizes, a feature generally consists of 64 pixels. For each feature defined in the basic grid, we first compute its coefficient of variation (CV), which is a function of the pixel intensities for that feature. Then we remove a pixel row or column from the feature in order to attain the greatest reduction in CV, if this reduction is judged to be statistically significant. This process continues until we can no longer achieve a significant reduction in the CV or until the modified feature size has been reduced to a predefined threshold (typically 16 pixels). After removing outliers, we compute the modified feature mean and standard deviation using the selected pixels. Data will be presented in Schadt et al. [1999] to demonstrate the significant reduction in variation between replicate samples that is achieved in using our adaptive pixel selection algorithm, when compared to the percentile algorithm that is currently most frequently used.

Background/Gradient Correction. Computing the raw feature intensities represents only the beginning in obtaining high-quality intensity measurements for each feature. Background noise correction is instrumental for determining intensities that accurately reflect the amount of RNA present for each gene on an array. The GeneChip software corrects for background variation by segmenting an image into 16 (by default) squares that cover the entire image. For each block, the lower 2% (by default) of the feature intensities for that block are averaged, and this average is subtracted from each feature in the block. One assumption implicit in this method is that feature-to-feature background variation is not significant. We have found this to be true in many cases, but then there are many other cases in which this assumption breaks down. Figure 3 illustrates the usefulness of computing a background intensity value for each feature by analyzing the neighbors of that feature. The image shown in Figure 3 represents the lower-right portion of an array covered by four of the

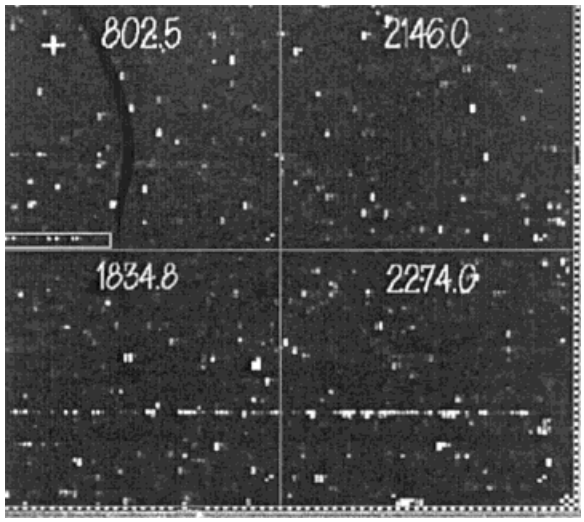

Fig. 3. The lower right portion of a C array from the low-density Murine 6500 Affymetrix GeneChip® set. See the background/gradient correction section for details on this image.

background correcting blocks described above. The intensity values listed in these blocks were computed using the Affymetrix algorithm described above. Because of the dark scar in the upper left block, the background intensity for this block is severely underestimated since the lower 2% of the feature intensities in this block are represented by features in the area devoid of signal. By examining the features surrounding a given feature and taking the median of some number of these surrounding features (we eliminate the brightest features from consideration since the brightest features are not generally informative in estimating the background intensity), the median background for the features in the rectangular region of the upper-left block was estimated to be 2,034, compared to the intensity estimate 803 computed using the GeneChip algorithm. For the gene in question, this resulted in a fourfold increase in the average log-ratio for the gene, which, in general, has a significant impact on assessing whether a gene is present. This localized method of background correction also goes far in reducing intensity gradients/strokes across an image. As one would expect, background correction rarely affects the PM/MM differences.

Artifact Detection. Any number of contaminants can cause large numbers of adjacent features to fluoresce brightly, thus obscuring the true hybridization intensities for these fea-

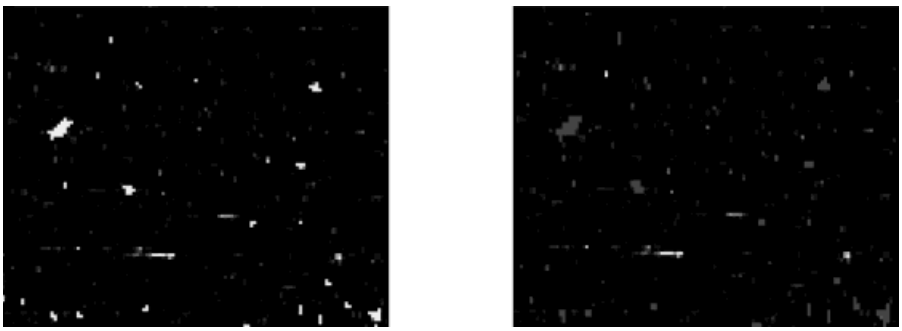

Fig. 4. The feature-level image on the left represents a portion of the pixel-level image shown in, displayed using our prototype gene expression software prior to masking. We currently mask at the feature level because masking at the pixel level can have undesirable effects on feature intensity calculations. Several spot artifacts are visible in this image (bright yellow spots of irregular shapes and sizes). The image on the right has been masked using the semiautomated masking technique described in the text; the masked spots are highlighted in blue.

tures (see Fig. 1). Note that some of the artifacts in Figure 1 cover from 40 to 100+ features (the feature sizes in this image are 50 μm; in the currently available high-density arrays, the feature sizes are 24 μm, which leads to bigger problems when array debris is present, since the features have one-fourth the area). Because this array debris can cover so many features, the summary statistics computed for the corresponding genes can be greatly affected, since the true hybridization signal will be obscured and a false signal will be given in its place. To extract meaningful information from these cases, users should mask the problem areas to prevent the obstructed features from being used in the analysis. While the GeneChip software has a tool that allows users to manually mask out the problem features, this process is tedious and can take several hours for a single array; because this process is so time consuming, scientists are usually loathe to mask out problem regions on an array. We have developed a semiautomated way to mask problem regions on an array that reduces the time needed to mask these problem regions. We have developed prototype software that allows users to click on an image problem region using the mouse, which invokes a function that recursively connects neighboring features with similar intensities. If the ratio of adjacent features (starting with the feature highlighted by the mouse click) is within a particular interval (we currently use the interval [0.70, 1.43]), then those features are automatically masked; the corresponding PM (MM) feature is masked when a given feature is masked. Figure 4 demonstrates this process. We are currently working on methods to completely automate the artifact detection procedure.

Allowing for Array-to-Array Comparisons: Scaling and Normalization

Scaling and normalization present one of the greatest challenges in getting the most from GeneChip data. Unlike the microarray technology in which multiple samples are competitively hybridized to an array, each GeneChip probe array has only a single sample hybridized to it. Therefore, to compare two or more arrays, the arrays must be brought into the same scale, or one array must be normalized against another. The GeneChip software currently assumes intensity differences between two or more arrays are linearly related with a zero y-intercept. This allows us to define a very simple and robust normalization factor:

$$\hat{\beta} = \frac{\sum_{i=1}^{N} (PM_i - MM_i)_{Chip_2}}{\sum_{i=1}^{N} (PM_i - MM_i)_{Chip_1}},$$

where N is the number of features on an array. $\hat{\beta}$ is an unbiased estimate of the slope of a weighted, linear least squares regression. Therefore, to normalize $chip_1$ against $chip_2$, we set

$$(PM_i - MM_i)^{new}_{Chip_1} = \beta(PM_i - MM_i)^{old}_{Chip_1}.$$

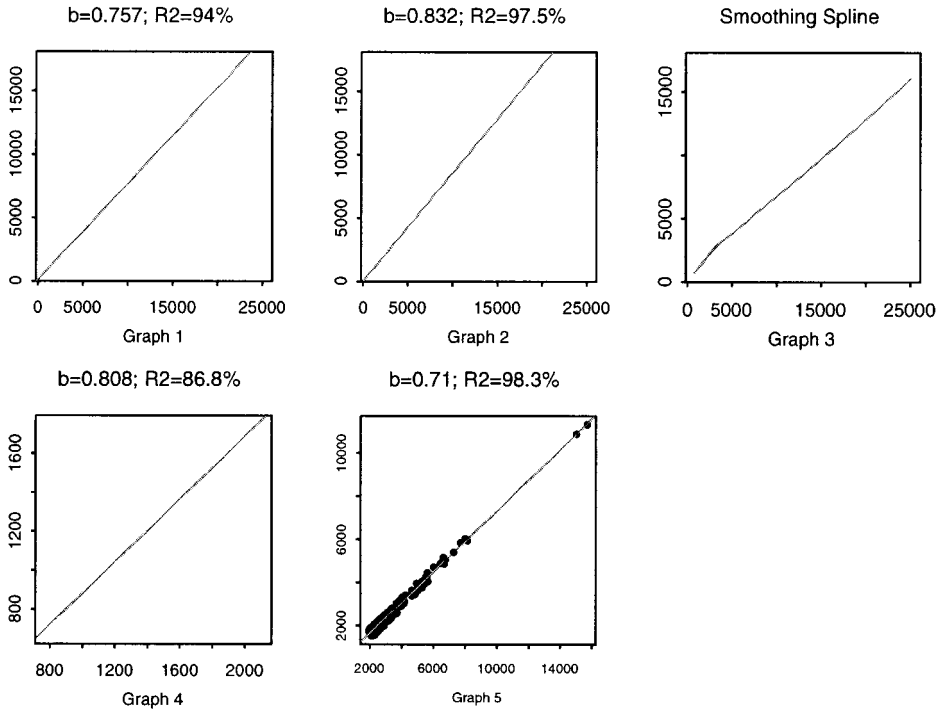

Fig. 5. Five graph representations of four normalization methods. The x-axis and y-axis for each graph represent the intensities of two arrays, where the intensities on the y-axis are to be normalized to the intensities on the x-axis. Graph 1 normalizes the array on the y-axis against the array on the x-axis using an ordinary least-squares regression, without assuming a zero y-intercept; the slope of this line (which represents the normalization factor) is given by $b = 0.757$, with a statistically significant positive y-intercept. Graph 2 uses the Affymetrix method of scaling (ordinary least-squares regression with a zero y-intercept). The third graph uses the smoothing spline normalization method; note the slight shift in slope between intensities 2,000 and 3,000. The behavior of the spline between intensities 0 and 2,000 is given by graph 4, which is an ordinary least squares regression on intensities less than 2,000; the behavior of the spline for intensities $> 2{,}000$ is given by graph 5. Note the difference in slope between graph 4 ($b = 0.818$) and graph 5 ($b = 0.716$) the slope in graph 5 is 87% the slope of graph 4. While this change-point is subtle (it represents the typical nonlinear behavior seen between arrays), note that a signal intensity of 5,000 on the y-axis array gets taken to an intensity value of 4,200 using the GeneChip® normalization method, and to an intensity value of 3,580 using the smoothing spline normalization method.

for all $i = 1..N$. Here the superscripts are used to distinguish the intensity values in chip 1 before and after the normalization. By examining many arrays, we found this linear method of normalization to be adequate in many cases, but in many other cases, the linear relationship simply does not hold (see Fig. 5). We have found that the distribution of the low-intensity signals behave differently than the distribution of the high-intensity signals; graphs 3, 4, and 5 in Figure 5 illustrates this point. The low-intensity/high-intensity distribution differences will often yield β estimates in which the estimate for the low-intensity signals is 10–50% less than or greater than the estimate for the high-intensity signals. Again, this can have a significant impact on reliably detecting differentially expressed genes.

To account for these differences in behavior across the dynamic range of an array, we apply change-point detection techniques to determine at which intensity point the slope of the linear regression line changes (i.e., where to define the low-intensity/high-intensity signal boundary) as we sweep through the dynamic range of the array. This results in dividing arrays into two intensity blocks, where a linear regression can be performed in each block, so that arrays are normalized one block at a time.

There are several problems in applying this sort of change-point analysis to obtain a normalization curve, including the fact that the normalization curve is piece-wise linear, that is, at the change-point, the normalization curve is not analytic. To eliminate this problem, we currently employ a smoothing spline technique

[Hastie and Tibshirani, 1997], which is capable of picking up the slope changes across the dynamic range of an array and, simultaneously, keeping the curve smooth at the various change-points. This work will be more fully described in Schadt et al. [1999], but Figure 5 demonstrates the nonlinearity in the expression signals that can arise.

Finally, it is worth noting that the median coefficient of variation for probe intensities over a set of replicate experiments is typically lowest when smoothing spline normalization is employed (compared to no normalization and linear normalization methods). For a set of six replicate, high-density, Affymetrix Human 6800 GeneChip® arrays generated at RBS for validation purposes, the median coefficient of variation (CV) for the probe intensities across the six replicates was only 7.5%, after normalizing five of the six arrays to the sixth using the smoothing spline method. The corresponding median CVs for the raw probe intensities (no normalization) and the probe intensity differences normalized using the GeneChip software were 32.3% and 8.9%, respectively.

Interestingly enough, the median CV for the raw probe intensities was 36.4% after normalizing the probe intensity differences using the GeneChip software. This suggests that normalizing on probe intensity differences does nothing to reduce the variation in the raw probe intensities. We are currently investigating whether it is best to normalize on the probe intensity differences or on the raw probe intensities.

ASSESSING GENE PRESENCE AND DIFFERENTIAL EXPRESSION SIGNIFICANCE

For the gene expression experiments to be useful, one must be able to assess whether genes are present or differentially expressed. The methods Affymetrix employs to determine whether a gene is present are similar to the methods used to determine whether a gene is differentially expressed. Therefore, only the gene presence calls will be discussed here. As described by Lockhart et al. [1996], the methods Affymetrix can employ to determine if a gene is present or absent depend on a variety of statistics:

1. *PositiveFraction* = the number of *PM*/*MM* differences that are significantly positive, where significance is determined by an empirically determined threshold constant.
2. Average Log−Ratio $= \dfrac{10(\sum_{i=1}^{N} \log(PM_i/MM_i))}{N}$, where N is the number of probe pairs for the gene and PM_i, and MM_i indicate the perfect match and corresponding mismatch feature intensities, respectively, for feature i.
3. $\dfrac{Positive}{Negative}$ = the number of *PM*/*MM* differences that are significantly positive divided by the number that are significantly negative. As in 1, the significances of the *PM*/*MM* differences are determined by an empirically determined threshold constant.

These statistics and the associated empirically determined parameters are used to estimate the parameters of a decision tree, which is then used to classify genes as present, marginally present, or absent. The GeneChip software associates significances with each classification, but biologically, they are difficult to interpret. Furthermore, scientists use these classifications as a means of filtering genes (e.g., a scientist may simply exclude all genes from consideration that were not classified present). This can turn out to be a rather unfortunate mistake as it is often the case that genes are at the threshold of being detected by the decision tree, and so, filtering on the categorical calls for these genes can result in missing many potentially informative genes. Also, the default parameters of the decision tree are estimated and set using data generated internally at Affymetrix. These parameter estimates are not automatically updated (the parameters can be changed manually by users) as users generate significant amounts of data, and so, the parameter estimates generated by Affymetrix will not typically coincide with estimates that would obtain if all historical data were taken into account.

To give users of this technology a more meaningful way to filter data, we propose testing the presence of a gene based on two simple null hypotheses:

$$PM_i \overset{D}{=} MM_i \text{ and } \frac{PM_i}{MM_i} \overset{D}{=} \frac{MM_i}{PM_i} \text{ for } i = 1..N,$$

TABLE I.

Gene	S1 ADI	S1 CALL	S1 PV	S2 ADI	S2 CALL	S2 PV	S3 ADI	S3 CALL	S3 PV	S4 ADI	S4 CALL	S4 PV
W	192.61	A	0.0033	160.75	P	8.77E-05	219.52	P	0.0018	174.32	P	0.0007
X	137.10	P	0.0925	203.21	A	0.1021	145.22	A	0.6661	137.46	A	0.4925
Y	145.42	P	0.0026	131.57	P	0.0017	180.42	P	0.0068	162.29	P	0.0028
Z	−6.38	A	0.4776	−22.91	A	0.7213	−3.69	A	0.4925	36.28	A	0.7773

[a]Four genes are represented in this table across 4 replicate samples (S1, S2, S3, and S4) hybridized to the Affymetrix Human 6800 High-Density GeneChip™. Each sample has three columns associated with it: 1) ADI gives the average difference intensity for the gene of interest, 2) CALL gives the GeneChip™ gene presence call for the gene of interest (A for absent, P for present), and 3) PV is the p-value for the average difference, randomization test described in the text. Gene W is called A in one of four of the replicate samples, while the p-value is significant for all four samples. Gene X is called A in three of four of the replicate samples while none of the p-values are significant at the 0.05 significance level. Genes Y and Z are consistently called P and A, respectively, with consistently significant and non-significant p-values.

where N, PM_i, and MM_i are as defined above, and the = indicates the intensities are equal in distribution. Because normality assumptions for probe intensities often do not hold and because the probe pairs are not necessarily independent [Alon et al., 1999], we have developed a randomization test, which will be described in more detail in Schadt et al. [1999], in which the following statistics are computed to empirically estimate the distributions of the PM/MM differences and the PM/MM ratios:

$$S_k = \sum_{i=1}^{N} (-1)^{I_i}(PM_i - MM_i),$$

and

$$R_k = \sum_{i=1}^{N} \left(\frac{PM_i}{MM_i}\right)^{(-1)^{I_i}},$$

where I_i is a random indicator function, N is the number of probe pairs for the gene, and $k = 1..M$, where M is the number of permutations considered. Each of the M sums for S_k and R_k are then stored in sorted lists, S and R, respectively. Then the sums:

$$S_0 = \sum_{i=1}^{N} (PM_i - MM_i),$$

and

$$R_0 = \sum_{i=1}^{N} \left(\frac{PM_i}{MM_i}\right),$$

are computed and these values are compared against the sorted lists. A quantitative measurement of significance is formed by determining where S_0 and R_0 would be inserted in the respective sorted lists and then dividing the index value at this insert position by the number of elements in the list. It can be shown that this measurement of significance for the PM/MM differences is asymptotic to the P-value given by the t statistic, when the underlying assumptions for this distribution are met [Cox and Hinkley, 1979]. The resulting P-values can then be used to augment the calls made by the GeneChip software and to quantitatively assess the significance of the gene calls.

Table 1 illustrates the usefulness of the randomization test P-values described above, as well as the potential danger in filtering genes based on the GeneChip software categorical calls. The associated P-values allow a more quantitative assessment of the presence or absence of a gene. For example, gene W in this table demonstrates that the A call in sample 1 has a significant P-value, which is consistent with the other 3 P calls for the other samples. The consistency of the other calls and the fact that these samples are replicates would indicate that the GeneChip software has given a false negative call for this sample, while the P-value accurately reflects that the gene is present. The same sort of randomization test can be applied in determining whether a gene is differentially expressed. For differential expression, the difference of the PM/MM differences or the log of the ratio of the PM/MM ratios are the statistics we have found most

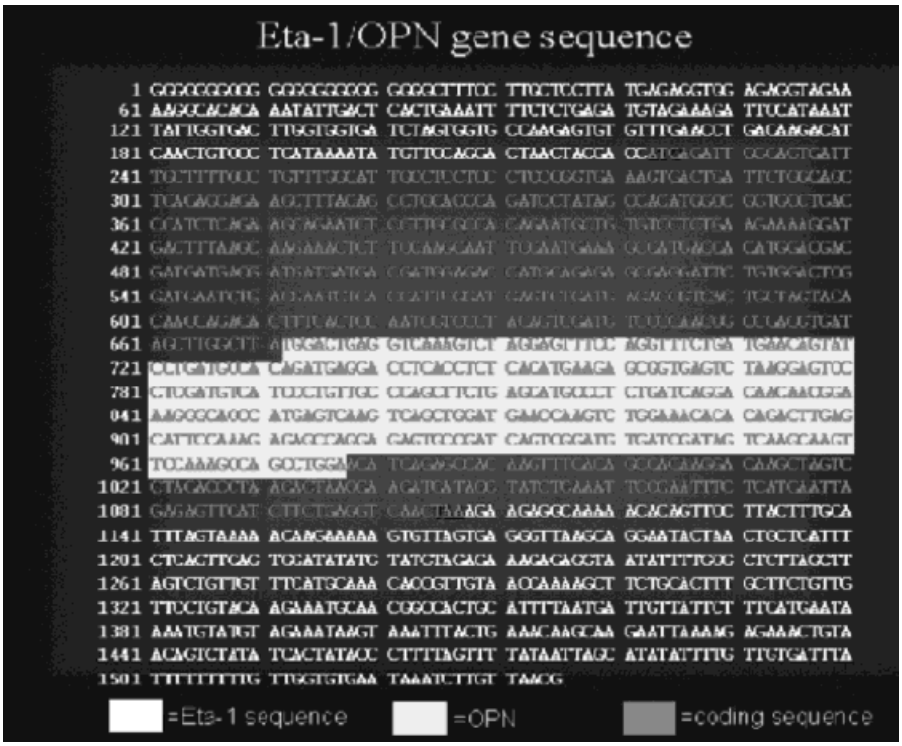

Fig. 6. Probes from Eta-1 (a full-length gene) and OPN (an EST within Eta-1) were independently selected from the corresponding sequences illustrated here. Note that GenBank was updated after the design of the corresponding probe array, to reflect that OPN is actually part of the Eta-1 gene. The white indicates the full-length gene sequence for the Eta-1 gene, the red indicates the coding sequence for the Eta-1 gene, and the yellow indicates the sequence for the OPN EST.

useful in reliably detecting differential expression. The permutation test does not address the probe dependency problem, but as probe sequence data becomes available, we will incorporate the dependency structure into these tests.

ASSESSING PROBE PERFORMANCE

When a gene transcript is actually present in a target sample hybridized to a probe array, the intensities of the individual probes corresponding to the transcript can vary greatly, and these intensity fluctuations are a function of the hybridization kinetics and nonspecific RNA hybridizations. Without tracking the performance of these probes, false gene presence or absence calls, or false differential expression calls will result. As an example, Figure 6 illustrates a gene, Eta-1, and an EST, OPN, from which probe sets were formed; a set of 20 probes was pulled from the 3′ UTR of the Eta-1 gene, and an additional 20 probes were pulled from the OPN EST, which, at the time the Affymetrix Generic Murine 6500 array was designed, was not known to belong to the translated region of the Eta-1 gene. Figure 7 shows the hybridization results of the Eta-1 gene and the OPN EST in untreated and treated samples. Neither probe set detected the presence of the corresponding sequence in the untreated case (both array regions shown for the untreated sample are from the same array, as are the regions shown for the treated sample), however, in the treated case, the Eta-1 gene was detected as present, but the OPN EST was not detected as present. Such conflicting results for gene transcripts represented on a single array multiple times is not uncommon, and data from our experiments indicate that when a gene transcript is present, PM probes from a given probe set for the gene are more likely to light up, when compared to PM probes from a more 5′ probe set for the same gene. This bias demands that we weight the 3′ probes more heavily with respect to their ability to indicate gene presence, while placing less weight on probes that are more 5′. The presence of the Eta-1 gene in the treated sample was confirmed by kinetic PCR (data not shown).

As indicated by Alon et al. [1999], the probe sequences in neighboring features can contain common sequence. Our understanding of the Affymetrix probe selection process is that probe sequences are picked from a given gene sequence after eliminating palindromic subsequences, Alu subsequences, and other subse-

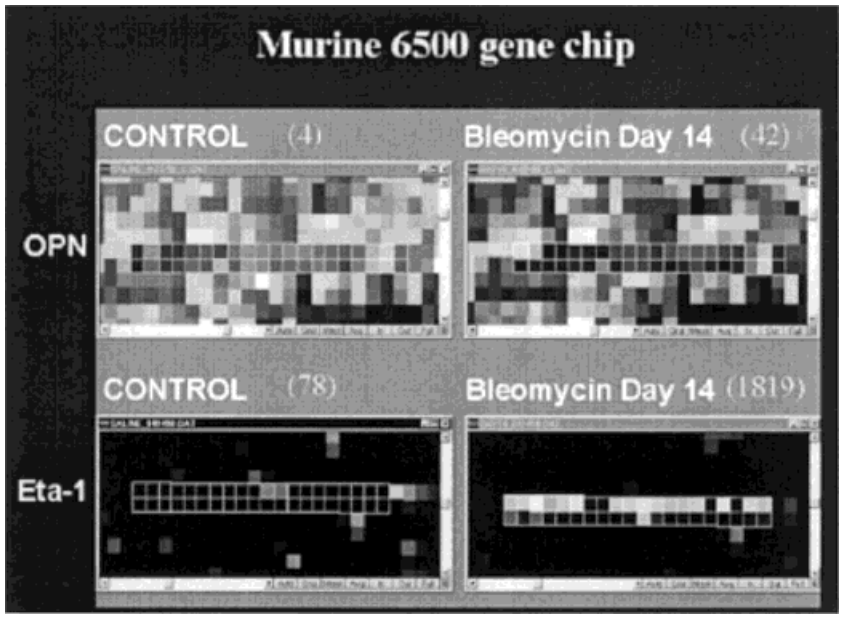

Fig. 7. The OPN and Eta-1 probe sets on the Affymetrix Murine 6500 GeneChip® array. The 20 probe pairs representing the OPN and Eta-1 sequences are outlined in white. The top row (OPN) indicates no detectable expression of the OPN EST in the untreated (control) and treated (bleomycin) samples. The bottom row (Eta-1) indicates no detectable expression of Eta-1 in the untreated sample but definite expression in the treated sample. The untreated data shown for OPN and Eta-1 are from the same array hybridization, as are the treated data. As shown in Figure 6, OPN and Eta-1 represent the same gene. The probes for Eta-1 were pulled from the 3′ UTR of the gene, while the OPN probes were pulled from the coding region of the gene (the area highlighted in yellow in Fig. 6).

quences that are homologous to other gene sequences, among a host of other exclusion criteria aimed at making the hybridization characteristics of a probe relatively uniform. Under such circumstances, it is not always possible to pick probes that are do not contain common sequence or that have hybridization characteristics that are similar to the "ideal" probe. Therefore, it is imperative that the common sequence structure and hybridization characteristics be accounted for in computing average intensity statistics and in determining whether a gene transcript is present or differentially expressed.

Current protocols do not call for high temperature hybridizations, and so, variations in the melting temperature can have a significant impact on the PM/MM differences, which, in turn, can greatly impact gene presence or differential expression detection. If the melting temperature is too high, relative to the experimental protocols defined by Affymetrix (these protocols call for carrying out the hybridization for 16 h at 45°C in a rotisserie oven set at 60 RPM), then the PM as well as the corresponding MM features will bind the RNA tightly, thus giving small PM/MM differences, which could severely bias gene presence and differential expression calls. Other serious variations we have seen with respect to hybridization efficiencies include arrays that have a brighter median probe intensity (over the entire array) than a baseline array, but which have PM/MM differences that are less than the differences of the baseline array. This would seem to indicate that the intensity signals for the PM features of the "bright" array are saturated, and that the intensity signals for the corresponding MM features are boosted, giving lower PM/MM differences.

If a significant number of PM features are reaching this saturation point, the data on the array are completely unreliable. In fact, this can lead to very misleading results, since when the arrays are compared, if normalization takes place on the PM/MM differences, genes that are actually up-regulated in the "bright" array may be called down-regulated. Similarly, genes that are actually present in the "bright" array may be called absent. We currently examine PM/MM differences as well as PM/MM sums, since information lost in the PM/MM difference may be partially recovered using both statistics.

The probe sequences contain a plethora of information that could be used to enhance current gene expression and differential expression detection algorithms. For instance, knowledge of the probe sequences would allow the hybridization efficiency (and hence, the quality of the probe) to be assessed based on the probe's position in the gene sequence, on its GC content, on its GC trend, or on any other attribute of the probe sequence that would be

highly predictive of its ability to efficiently hybridize (this would include things like alternative splicing, SNPs, etc.) Current systems do not provide for any sort of quality score for the probes on an array (i.e., past arrays are not used to assess performance of probes in future experiments). We strongly believe such a scoring system for the probes would greatly improve gene detection algorithms and the ability to quantify levels of a gene transcript when a gene transcript is actually present.

CONCLUSION

The various methods discussed in this article go further than currently available methods in accounting for many of the sources of variation that can obscure the very biological variation the technology aims to detect. Clearly, there is much further to go in analyzing gene expression array data. Sophisticated background/gradient correction methods, scaling/normalization methods, methods to assess a probe's hybridization efficiency, and methods to detect gene presence or differential expression, only begin to approach the goal of extracting as much information as possible from these data. Power analysis, introducing orthogonal biological factors to increase the information content of these data, and building on the current clustering techniques used to explore gene expression data [Eisen, 1998; Tamayo, 1999], will all be necessary to get the most from this technology. Furthermore, as expression libraries become widely available, we will be able to estimate the variation structures of genes in a variety of tissue and across many different organisms, which will result in more informative differential expression analyses. Making better use of the currently available sequence information will enhance probe selection algorithms, thus increasing the efficiency of the probes and better accounting for the many complicated biological phenomena (e.g., alternative splice sites) that are currently not well understood. Of course, these issues represent only the computational aspects of this technology and do not even begin to address the sort of informatics infrastructure necessary to intelligently store and mine this type of expression data, which, at RBS, has already hit the terabyte scale.

The ability to simultaneously monitor tens of thousands of genes across a series of experiments is truly a great technology breakthrough for the biomedical and life sciences. However, these types of high-throughput assays generate data that require sophisticated computational and statistical techniques for their analysis. By spending the time to develop such methods up front, we believe the better quality data that result, will make it possible to detect many of the more subtle gene expression patterns and gene interactions that give rise to all of the complexities of living systems.

ACKNOWLEDGMENTS

We thank Renu Heller, Gary Peltz, John Allard, Fengrong Zuo, Andrew Grupe, and Dee Aud of Roche Bioscience for providing us with the gene expression array data.

REFERENCES

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96:6745–6750.

Cox DR, Hinkley DV. 1979. Theoretical statistics. London: Chapman and Hall.

Der SD, Zhou A, Williams BRG, Silverman RH. 1998. Identification of genes differentially regulated by interferon α, β, or γ using oligonucleotide arrays. Proc Natl Acad Sci USA 95:15623–15628.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868.

Eisenberg DS, Crothers DM. 1979. Physical chemistry: with applications to the life sciences. Menlo Park, CA: Benjamin/Cummings Publishing Company.

GATC Consortium. 1998. GATC Specifications: Software Specifications. Available at http://www.gatconsortium.org/.

Hastie TJ, Tibshirani RJ. 1997. Generalized additive models. New York: Chapman and Hall.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallow MV, Chee MS, et al. 1996. Expression monitoring by hyrbridization to high-density oligonucleotide arrays. Nature Biotechnol 14:1675–1680.

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 96:2907–2912.

Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ. 1997. Genome-wide expression monitoring in *Saccharomyces cerevisiae.* Nature Biotechnol 5:1359–1366.