JOURNAL OF BIOMEDICAL OPTICS 2(4), 364–374 (OCTOBER 1997)

RATIO-BASED DECISIONS AND THE QUANTITATIVE ANALYSIS OF cDNA MICROARRAY IMAGES

Yidong Chen,[†] Edward R. Dougherty,[‡] and Michael L. Bittner[†]
[†]National Institutes of Health, National Human Genome Research Institute, Bethesda, Maryland 20892; [‡]Texas A&M University, Texas Center for Applied Technology and Department of Electrical Engineering, College Station, Texas 77843

(Paper JBO-150 received Mar. 24, 1997; revised manuscript received June 23, 1997; accepted for publication July 8, 1997.)

ABSTRACT

Gene expression can be quantitatively analyzed by hybridizing fluor-tagged mRNA to targets on a cDNA microarray. Comparison of gene expression levels arising from cohybridized samples is achieved by taking ratios of average expression levels for individual genes. A novel method of image segmentation is provided to identify cDNA target sites and a hypothesis test and confidence interval is developed to quantify the significance of observed differences in expression ratios. In particular, the probability density of the ratio and the maximum-likelihood estimator for the distribution are derived, and an iterative procedure for signal calibration is developed. © *1997 Society of Photo-Optical Instrumentation Engineers.* [S1083-3668(97)00504-2]

Keywords cDNA; microarray; gene expression; image segmentation; Mann–Whitney target detection; ratio density, ratio confidence interval.

1 INTRODUCTION

The recent development of complementary DNA microarray technology provides a powerful analytical tool for human genetic research.[1] One of its basic applications is to quantitatively analyze fluorescence signals that represent the relative abundance of mRNA from two distinct tissue samples. cDNA microarrays are prepared by automatically printing thousands of cDNAs in an array format on glass microscope slides, which provide gene-specific hybridization targets. Two different samples (of mRNA) can be labeled with different fluors and then cohybridized onto each arrayed gene. Ratios of gene expression levels between the samples are calculated and used to detect meaningfully different expression levels between the samples for a given gene.

This paper studies ratio distributions and develops a hypothesis test and confidence interval so that expression ratios may be used for deciding significant differences in sample expressions across the gene population discernible on a microarray. Assuming sample expression levels are independent, levels are normally distributed, and there is a constant coefficient of variation for the entire gene set (a biochemical consequence of the mechanics of transcript production), we derive the probability density of the ratio, find the maximum-likelihood estimator for the distribution, and develop an iterative procedure for signal calibration. Under the aforementioned conditions, we can process a single image and identify outliers. Expression measurements are achieved by processing digitized microarray images, the key imaging development being a nonparametric statistical technique to extract cDNA sites on the slide.

2 BIOLOGICAL BACKGROUND AND cDNA MICROARRAY TECHNOLOGY

A cell relies on its protein components for a wide variety of its functions. The production of energy, the biosynthesis of all component macromolecules, the maintenance of cellular architecture, and the ability to act upon intra- and extracellular stimuli are all protein dependent. Each cell within an organism contains the information necessary to produce the entire repertoire of proteins which that organism can specify. This information is stored as genes within the organism's DNA genome. The number of human genes is estimated to be 30,000 to 100,000. Within any individual cell, only a portion of the possible gene set is present as protein. Some of the proteins present in a single cell are likely to be present in all cells because they serve functions required in every type of cell, and can be thought of as ''housekeeping'' proteins. Other proteins serve

Address all correspondence to Yidong Chen. NIH/NHGRI/LCG, Bldg. 49, Rm. 4B24, 49 Convent Drive, MSC 4470, Bethesda, MD 20892-4470. Tel: (301) 402-3150; Fax: (301) 402-3241; E-mail: yidong@nhgri.nih.gov

1083-3668/97/$10.00 © 1997 SPIE

specialized functions only required in particular cell types. For example, muscle cells contain specialized proteins that form the dense contractile fibers of a muscle. Given that a large part of a cell's specific functionality is determined by the genes it is expressing, it is logical that transcription, the first step in the process of converting the genetic information stored in an organism's genome into protein, would be highly regulated by the control network that coordinates and directs cellular activity.

Regulation is readily observed in studies that scrutinize activities evident in cells configuring themselves for a particular function (specialization into a muscle cell) or state (active multiplication or quiescence). As cells alter their status, coordinate transcription of the protein sets required for this state can be observed. As a window both on cell status and on the system controlling the cell, detailed, global knowledge of the transcriptional state could provide a broad spectrum of information useful to biologists. Knowledge of when and in what types of cell the protein product of a gene of unknown function is expressed would provide useful clues as to the likely function of that gene. Determination of gene expression patterns in normal cells could provide detailed knowledge of the way in which the control system achieves the highly coordinated activation and deactivation required for development and differentiation of a mature organism from a single fertilized egg. Comparison of gene expression patterns in normal and pathological cells could provide useful diagnostic ''fingerprints'' and help identify aberrant functions that would be reasonable targets for therapeutic intervention.

The ability to carry out studies in which the transcriptional state of a large number of genes is determined has, until recently, been severely inhibited by limitations on our ability to survey cells for the presence and abundance of a large number of gene transcripts in a single experiment. A primary limitation has been the small number of identified genes. In the case of humans, only a few thousand of the complete set (30,000 to 100,000 genes) have been physically purified and characterized to any extent. Another significant limitation has been the cumbersome nature of transcription analysis. Even a large experiment on human cells could track expression of only a dozen genes, clearly an inadequate sampling for inference about so complex a control system.

Two recent technological advances have provided the means to overcome some of these limitations to examining the patterns and relationships in gene transcription. The cloning of molecules derived from mRNA transcripts in particular tissues, followed by the application of high-throughput sequencing to the DNA ends of the members of these libraries has yielded a catalog of expressed sequence tags (ESTs).[2] These signature sequences provide unambiguous identifiers for a large cohort of genes. At present, approximately 40,000 human genes have been ''tagged'' by this route, and many have been mapped to their genomic location.[3]

In addition, the clones from which these sequences were derived provide analytical reagents that can be used in the quantitation of transcripts from biological samples. The nucleic acid polymers, DNA and RNA, are biologically synthesized in a copying reaction in which one polymer serves as a template for the synthesis of an opposing strand, which is termed its complement. Even after separation from each other, these strands can be induced to pair quite specifically with each other to form a very tight molecular complex, a process called *hybridization*. This specific binding is the basis of most analytical procedures for quantitating the presence of a particular species of nucleic acid, such as the mRNA specifying a particular protein gene product. Microarray technology is a recent hybridization-based process that allows simultaneous quantitation of many nucleic acid species.[1,4,5] This technique combines robotic placement (spotting) of small amounts of individual, pure nucleic acid species on a glass surface, hybridization to this array with multiple fluorescently labeled nucleic acids, and detection and quantitation of the resulting fluor-tagged hybrids with a scanning confocal microscope. When used to detect transcripts, a particular RNA transcript (an mRNA) is copied into DNA (a cDNA) and this copied form of the transcript is immobilized on a glass surface.

The entire complement of transcript mRNAs present in a particular cell type is extracted from cells and then a fluor-tagged cDNA representation of the extracted mRNAs is made *in vitro* by an enzymatic reaction termed *reverse transcription*. Fluor-tagged representations of mRNA from several cell types, each tagged with a fluor emitting a different color light, are hybridized to the array of cDNAs and then fluorescence at the site of each immobilized cDNA is quantitated.

The various characteristics of this analytic method make it particularly useful for directly comparing the abundance of mRNAs present in two cell types. An example of such a system is presented in Figure 1. In this experiment,[4] an array of cDNAs was hybridized with a green fluor-tagged representation of mRNAs extracted from a tumorigenic melanoma cell line (UACC-903) and a red fluor-tagged representation of mRNAs was extracted from a nontumorigenic derivative of the original cell line (UACC-903 +6). Monochrome images of the fluorescent intensity observed for each of the fluors were then combined by placing each image in the appropriate color channel of a red-green-blue (RGB) image, as shown in Figure 2 (color plate). In this composite image, one can see the differential expression of genes in the two cell lines. Intense red fluorescence at a spot indicates a high level of expression of that gene in the nontumorigenic cell line, with little expression of the

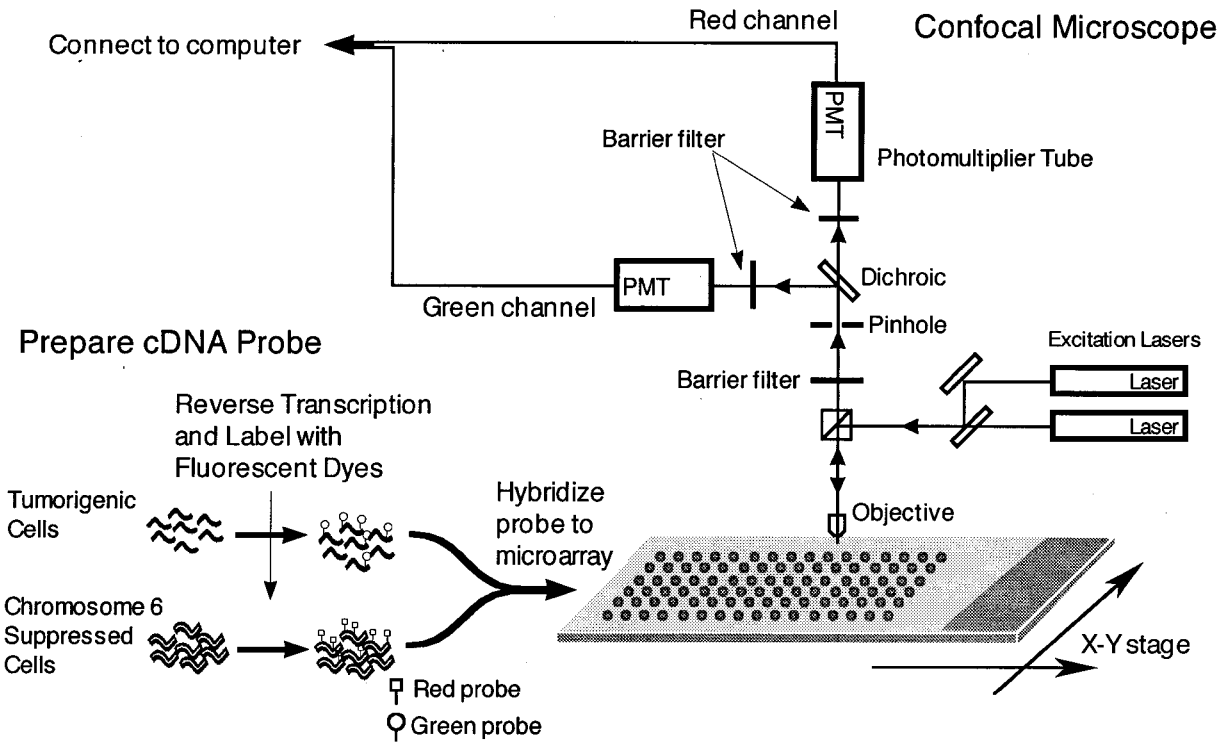

Fig. 1 Illustration of a microarray system.

same gene in the tumorigenic parent. Conversely, intense green fluorescence at a spot indicates high expression of that gene in the tumorigenic line, with little expression in the nontumorigenic daughter line. When both cell lines express a gene at similar levels, the observed array spot is yellow.

Visual inspection of such results is sufficient to find genes where there is a very large differential rate of expression. A more thorough study of the changes in expression requires the ability to discern more subtle changes in expression level and to determine whether observed differences are the result of random variation or whether they are likely to be meaningful changes.

3 IMAGE PROCESSING AND MANN–WHITNEY SEGMENTATION

Assuming that DNA products from two samples have an equal probability of hybridizing to the target, the intensity measurement is a function of the quantity of the specific DNA products available within each sample. Locally (or pixelwise), the intensity measurement is also a function of the concentration of the target segments. On the scanning side, the fluorescent light intensity also depends on the power and wavelength of the laser, the quantum efficiency of the photomultiplier tube, and the efficiency of other electronic devices. The resolution of a scanned image is largely determined by processing requirements and acquisition speed. The scanning stage imposes a calibration requirement, though it may be relaxed later. The image analysis task is to extract the average fluorescence intensity from each target site (cDNA region).

There are several fluorescent light sources for each slide: background, target, the target hybridized with sample 1 or sample 2, and (possibly) a glass surface. The average intensity within a target site is measured by the median image value on the site. This intensity serves as a measure of the total fluors emitted from the sample mRNA probes hybridized on the target site. The median is used as the average to mitigate the effect of outlying pixel values created by noise.

Some image processing is required prior to measuring intensity. Most is quite standard and need not be described here. For instance, the image needs to be segmented into target patches, but this task is straightforward since the robot positions the cDNA targets in a predetermined manner. Because the number of pixels in the target site is limited, both smoothing and sharpening filters need to be avoided.

The difficult image processing task is to identify the target site within the target patch (see Figure 3 color plate). Each target site is somewhat annular owing to how the robot finger places the cDNA on the slide and how the slide is treated; however, there is variability in this placement (within the patch) from image to image and from target to target. This variability can be so great that the target region is simply a collection of subregions within

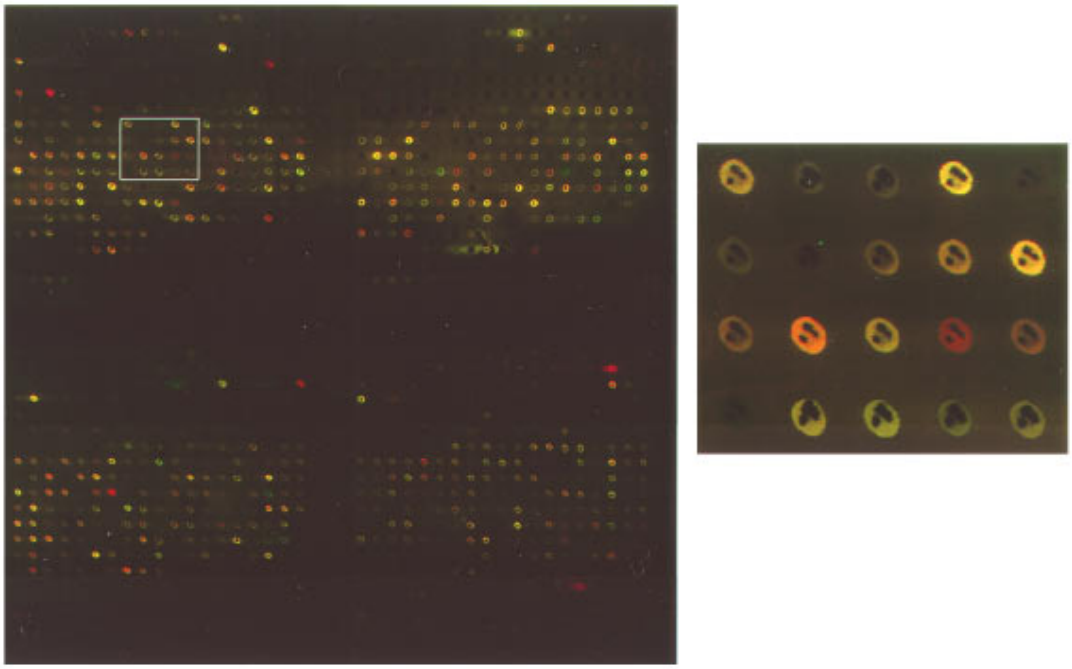

Fig. 2 cDNA microarray image.

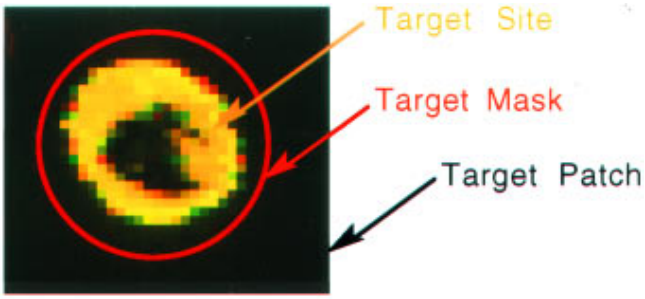

Fig. 3 Target patch, mask, and site.

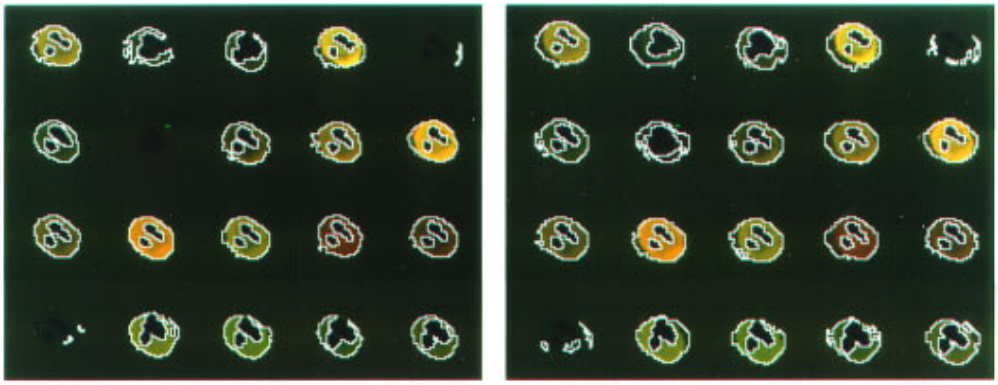

Fig. 4 Target detection results at different significant levels.

COLOR PLATE

the nominal circular target region. This instability in the target region is manifested in the irregular way the mRNA is hybridized to the target and the consequent irregular brightness pattern (created by the fluors) within the target site. It is important that mRNA intensity be measured over these fluor regions because only they correspond to probe-hybridized-to-target areas. Conventional adaptive thresholding segmentation techniques are unsatisfactory when the signal is weak because there is no marked transition between foreground and background. Standard morphological methods also fail because for weak signals there is no consistent shape information for the target area.

To overcome these difficulties, we propose a pixel selection method based on the Mann–Whitney test. There are three key points associated with this approach: (1) it associates a confidence level with every intensity measurement based on the significance level of the test and, if desired, it enables multiple readouts at different confidence levels; (2) it meets the real-time requirement of the system; and (3) it is a distribution-free test, thereby eliminating the need for normality assumptions.

We briefly describe the Mann–Whitney test as employed here. Assume that $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are independent samples arising from two random variables X and Y possessing means μ_X and μ_Y, respectively. The rank-sum statistic W, which is the sum of the ranks of all X samples in the combined ordered sequence of the X and Y samples, is used to test the null hypothesis,

$$H_0: \mu_X - \mu_Y = 0$$
$$H_1: \mu_X - \mu_Y > 0. \tag{1}$$

The Mann–Whitney criterion reveals the relation between the X and Y positions in the combined ordered sequence. Rejection of H_0 occurs when $W \geq w_{\alpha,n,m}$, the critical value corresponding to the significance level α. (See Ref. 6 for a detailed discussion on the Mann–Whitney hypothesis test and notations.)

A target site is segmented from the target patch according to the following procedure. A predefined target mask is used to identify a portion of the target patch that contains the target site. The target mask is based on the geometry of the potential target area and can be constructed from specially tagged targets or other strong targets (e.g., the target mask is obtained by finding all strong targets, aligning them together, averaging, and then thresholding). We randomly pick 8 sample pixels from the known background (outside the target mask) as $Y_1, Y_2, \ldots, Y_8$, and select the lowest 8 samples from within the target mask as $X_1, X_2, \ldots, X_8$. The rank-sum statistic W is calculated and, for a given significance level α, compared with $w_{\alpha,8,8}$. We choose 8 samples here for both foreground and background because the Mann–Whitney statistic is approximately normal when $m = n \geq 8$.[6] If the null hypothesis is not rejected, then we discard some predetermined number (perhaps only 1) of the 8 samples from the potential target region and select the lowest 8 remaining samples from the region. The Mann–Whitney test is repeated until the null hypothesis is rejected.

When H_0 is rejected, the target site is taken to be the 8 pixels causing the rejection, together with all pixels in the target mask whose values are greater than or equal to the minimum value of the eight. The resulting site is said to be a target site of significance level α. If the null hypothesis is never rejected, then it is concluded that there is no appreciable probe at the target site. Furthermore, one can require that the Mann–Whitney target site contain at a minimum some number of pixels for the target site to be considered valid and measured for fluor intensity. Figures 4(a) and 4(b) (color plate) show the detection results of target sites at $\alpha = 0.0001$ and $\alpha = 0.05$, respectively, where the detected site boundaries are superimposed on the original images. Once a target site is determined, gene expression is measured by the median of the target site minus the median of the background area (outside the target mask area).

4 PROBABILITY DENSITY FUNCTION OF RATIO PARAMETERS

We wish to use the expression ratio to determine whether gene expression differs significantly for the red and green samples. Such an approach is intuitive because equal distributions for red and green values lead to a red/green ratio close to 1, and significantly unequal distributions lead to a red/green ratio significantly different from 1. This approach is typically being applied by biologists developing microarrays.

A key purpose of this paper is to examine expression ratios. Two points need to be taken into consideration. First, even if red and green measurements are identically distributed, the mean of the ratio distribution will not be 1; second, the hypothesis test needs to be performed on expression levels from a single microarray. A salient factor in using expression ratios rather than expression differences is that gene expression levels are determined by the intrinsic properties of each gene, which means that differences in expression levels vary widely among genes, regardless of the truth of the null hypothesis; therefore it is inappropriate to pool statistics on gene expression differences across the microarray. Labeling the red and green microarray values for the genes by $R_1, R_2, \ldots, R_n$ and $G_1, G_2, \ldots, G_n$, respectively, the desired hypothesis test is

$$H_0: \mu_{R_k} = \mu_{G_k}, \quad H_1: \mu_{R_k} \neq \mu_{G_k} \tag{2}$$

using the test statistic $T_k = R_k / G_k$. This requires

finding a critical region for T_k, recognizing that the mean of T_k under the null hypothesis is not 1.

It is well known that working with ratio distributions can be problematic,[7–8] and recent research on the matter is generally confined to normality studies of the ratio distribution,[9] and numerical calculations.[10–11] However, as we now discuss, a special situation arises for gene expression that permits a more detailed statistical analysis, as well as hypothesis tests and confidence intervals based on a single microarray.

While it would be possible to gather data on the routine level of expression for each specific gene in each specific tissue, this would be a very difficult undertaking. The method currently requires substantial quantities of mRNA (and thus tissue) for each determination. Extending the studies to pathological situations would further complicate the ability to gather material for replicates, since it will initially be necessary to assume that diseases with complex molecular etiologies may have many forms, making pooling of samples from different individuals counterproductive. The most practical and informative version of an assay of this type would be achieved if information on the variance of all or most of the genes in a sample could be used to derive a statistically sound measurement of variance for each transcript. Fortunately, it appears that the biology of transcription makes such an approach possible.

A transcript's abundance at a given time is governed by the current rates of production and degradation of that transcript. As would be expected of a system faced with routine generation and destruction of these information intermediates, the processes that produce and destroy transcripts rely on common, core enzymatic machinery (polymerases and nucleases) whose specificity of activity is modulated by accessory proteins that bind to the core enzymes, the nucleic acid sites of action, or both. As might also be expected of a system that must constantly synthesize and hydrolyze tens of thousands of molecules, molecular interactions are based on very similar intermolecular affinities. Nimbleness at this scale requires that the core machinery operate without too much bias, so that no single or small class of transcripts consumes too large a share of the machinery's capacity. This type of bulk processing is thus predicted to be an approximation of a much simpler reaction, in which the level of a transcript will depend roughly on the concentration of the accessory factors driving its selection, and the variations for any particular transcript would be expected to be normally distributed and constant (as a fraction of abundance) relative to most of the other transcripts.

Such assumptions about the variances produce a special situation that can be exploited to great advantage, allowing the use of the variation data from all transcripts surveyed to be pooled to estimate the global variation of transcript synthesis and destruction. An important caveat to this hypothesis is that transcripts present at extremely high or extremely low levels could require a different method of control of synthesis/degradation and would not necessarily have variances representative of transcripts present at a common level.

Assuming there is constant coefficient of variation c for the entire gene set,

$$\sigma_{R_k}=c\mu_{R_k}, \quad \sigma_{G_k}=c\mu_{G_k}. \tag{3}$$

Under the null hypothesis H_0, $\mu_{R_k}=\mu_{G_k}$. Letting μ_k denote the common value, the condition of Eq. (3) becomes $\sigma_{R_k}=\sigma_{G_k}=c\mu_k$. From the experimental protocol, we assume that R_k and G_k are independent, identically distributed normal random variables.

If X and Y are continuous random variables, $T=X/Y$, and X and Y possess the joint probability density function $f_{X,Y}(x,y)$, then the probability distribution function for T is

$$\begin{aligned} F_T(t) &= P(X\leqslant tY, Y>0)+P(X\geqslant tY, Y<0) \\ &= \int_0^\infty \left[\int_{-\infty}^{ty} f_{X,Y}(x,y)\,dx\right] dy \\ &\quad + \int_{-\infty}^0 \left[\int_{ty}^{\infty} f_{X,Y}(x,y)\,dx\right] dy. \end{aligned} \tag{4}$$

For independent X and Y, differentiation yields the probability density function for T as

$$\begin{aligned} f_T(t) &= \int_0^\infty y f_{X,Y}(ty,y)\,dy - \int_{-\infty}^0 y f_{X,Y}(ty,y)\,dy \\ &= \int_0^\infty y f_X(ty) f_Y(y)\,dy - \int_{-\infty}^0 y f_X(ty) f_Y(y)\,dy, \end{aligned} \tag{5}$$

where the second equality follows from independence.

We apply Eq. (5) under the normality, independence, and constant-coefficient-of-variation conditions. Since microarray intensity measurements are positive, densities for both red and green values are assumed to be 0 for negative arguments. The error created by the simultaneous normality and positive-value assumptions is negligible because measurement intensities are sufficiently positive to render the portions of the left tail of the ratio distribution falling to the left of the y axis negligible. Letting $T_k=R_k/G_k$,

$$\begin{aligned} f_{T_k}(t) &= \int_0^\infty g f_{R_k}(tg) f_{G_k}(g)\,dg - \int_{-\infty}^0 g f_{R_k}(tg) f_{G_k}(g)\,dg \\ &= \int_0^\infty \frac{1}{\sigma_{R_k}\sqrt{2\pi}} \exp[-(tg-\mu_{R_k})^2/2\sigma_{R_k}^2] \end{aligned}$$

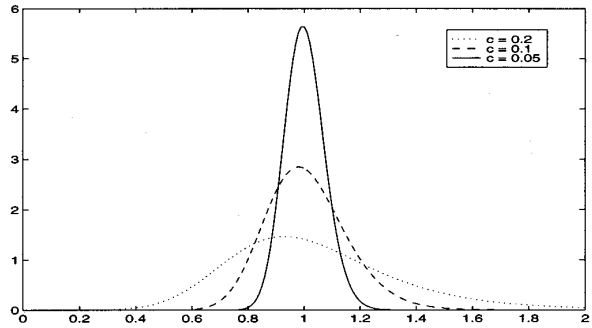

Fig. 5 Ratio density functions for $c=0.05$, 0.1, and 0.2.

$$\times\frac{1}{\sigma_{G_k}\sqrt{2\pi}}\exp[-(g-\mu_{G_k})^2/2\sigma_{G_k}^2]g\,dg$$

$$=\frac{1}{2\pi c^2}\int_0^\infty \exp[-(tu-1)^2/2c^2]$$

$$\times\exp[-(u-1)^2/2c^2]u\,du, \qquad (6)$$

where the second equality follows from the positive-value assumption and the third from Eq. (3) and the substitution $g/\mu_k=u$. Note that the density for T_k is independent of k. This property is not merely a consequence of Eq. (3), but depends on normality.

The integration of Eq. (6) yields a solution that is given by the standard error equation. Note that the second exponential in the integrand is similar to the normal density function with $\mu=1$ and $\sigma=c$. When c is small (less than 0.3), the second exponential is close to 0 for $u<0$. Therefore, by extending the integration to $-\infty$, we have the approximation

$$f_{T_k}(t)\approx\frac{1}{2\pi c^2}\int_{-\infty}^\infty \exp[-(tu-1)^2/2c^2]$$

$$\times\exp[-(u-1)^2/2c^2]u\,du$$

$$=\frac{(1+t)\sqrt{1+t^2}}{c(1+t^2)^2\sqrt{2\pi}}\exp[-(t-1)^2/2c^2(1+t^2)]. \qquad (7)$$

The approximation error of Eq. (7) can be numerically evaluated. For example, given $c=0.3$, at $t=1.0$, the approximation error between Eqs. (6) and (7) is 4.8×10^{-8}, and at $t=3.0$, the error is 1.2×10^{-8}. Figure 5 depicts the probability density function given in Eq. (7) for $c=0.05$, 0.1, and 0.2. The density function of Eq. (7) is an asymmetric function and its peak is close to 1 under the null hypothesis. Since Eqs. (6) and (7) are not functions of k, we denote the density function by $f_T(t;c)$ with parameter c.

5 CONFIDENCE INTERVALS AND MAXIMUM-LIKELIHOOD ESTIMATION

Confidence intervals can be obtained via Eq. (7). Table 1 lists the upper (right) limit and lower (left) limit of 95% confidence intervals for different c values, as well as the mean and standard deviation of the corresponding distributions. As functions of c, the mean and standard deviations of the confidence interval limits can be approximated by polynomial functions

$$y=a_3c^3+a_2c^2+a_1c+a_0. \qquad (8)$$

Table 2 gives the appropriate polynomial coefficients for the upper limit, lower limit, mean, and standard deviation. Figure 6 provides curves for 95, 90, 85, and 80% confidence levels. Most results obtained here have been verified by Monte Carlo simulation. Referring back to the hypothesis test of Eq. (2), for each k, the acceptance region for the test statistic T_k is the confidence interval for the appropriate value of c and the confidence level.

Typically, c needs to be estimated from the data. Using the density of Eq. (7), we can obtain a maximum-likelihood estimator for c. The likelihood function is

$$L(c)=\prod_{i=1}^n \frac{(1+t_i)\sqrt{1+t_i^2}}{c(1+t_i^2)^2\sqrt{2\pi}}\exp[-(t_i-1)^2/2c^2(1+t_i^2)], \qquad (9)$$

where $t_1,t_2,\ldots,t_n$ are ratio samples taken from a single collection of expression values, for example, all ratios from the housekeeping genes in a microarray. The maximum-likelihood criterion requires that $d[\log L(c)]/dc=0$. Hence, the estimator for c is

$$\hat{c}=\left[\frac{1}{n}\sum_{i=1}^n \frac{(t_i-1)^2}{(1+t_i^2)}\right]^{1/2}. \qquad (10)$$

6 UNCALIBRATED SIGNALS

The null hypothesis of equal means is appropriate for calibrated signal acquisition, but in practice this may not be the case. Therefore we consider the uncalibrated situation in which the means of the red and green signals are related by a constant amplification (or reduction) gain factor m, $\mu_{R_k}=m\mu_{G_k}$. If $m\geq1$, then the red signal is stronger than the green. We can follow the same derivation as in the cali-

Table 1 Lower and upper limits at 95% confidence level, and other statistics of ratio density.

Input dist. C.V. (c)	Output distribution parameters					
	L. limit	U. limit	Mean (μ)	Dev. (σ)	C.V. (σ/μ)	Peak t_{max}
0.01	0.972	1.026	1.000	0.014	0.014	1.000
0.02	0.945	1.052	1.000	0.028	0.028	0.999
0.03	0.919	1.080	1.001	0.042	0.042	0.998
0.04	0.894	1.108	1.002	0.056	0.057	0.997
0.05	0.869	1.137	1.003	0.071	0.071	0.995
0.06	0.845	1.167	1.004	0.085	0.085	0.993
0.07	0.822	1.198	1.005	0.100	0.100	0.990
0.08	0.798	1.230	1.007	0.115	0.114	0.987
0.09	0.776	1.263	1.008	0.130	0.129	0.984
0.10	0.754	1.297	1.010	0.145	0.144	0.980
0.11	0.732	1.332	1.013	0.161	0.159	0.976
0.12	0.710	1.369	1.015	0.177	0.174	0.972
0.13	0.689	1.407	1.018	0.193	0.190	0.967
0.14	0.669	1.447	1.021	0.210	0.206	0.962
0.15	0.648	1.488	1.024	0.227	0.222	0.957
0.16	0.628	1.531	1.028	0.245	0.239	0.951
0.17	0.609	1.576	1.032	0.264	0.256	0.946
0.18	0.589	1.623	1.036	0.283	0.274	0.939
0.19	0.570	1.672	1.041	0.304	0.293	0.933
0.20	0.551	1.724	1.046	0.326	0.312	0.926
0.21	0.532	1.778	1.052	0.349	0.332	0.919
0.22	0.514	1.835	1.057	0.372	0.353	0.912
0.23	0.495	1.895	1.064	0.397	0.374	0.905
0.24	0.477	1.958	1.070	0.423	0.396	0.897
0.25	0.459	2.026	1.077	0.450	0.418	0.890
0.26	0.441	2.098	1.084	0.477	0.440	0.882
0.27	0.424	2.174	1.091	0.504	0.460	0.874
0.28	0.407	2.257	1.098	0.531	0.484	0.866
0.29	0.390	2.346	1.105	0.558	0.506	0.858
0.30	0.373	2.442	1.111	0.585	0.527	0.850

brated case except that now the ratio density has two parameters, c and m. This results in the recursive relation

$$f_T(t;c,m)=\frac{1}{m}f_T(t/m;c,1), \tag{11}$$

where $f_T(\bullet;c,1)$ is given by Eq. (7). Figure 7 shows cases for $m=0.5$, 1, and 2 (when $c=0.1$). For $m=0.5$, we expect R_k/G_k to be about 0.5, which is what Figure 7 indicates.

In the uncalibrated setting, estimators are required for both c and m; however, a closed-form solution as in the calibrated case is precluded by reliance on the recursion of Eq. (11). We proceed iteratively to obtain estimators. Note from Table 1 that the means for different c values are very close to 1 when $m=1$. Intuitively, when two signals are approximately the same, the mode of the ratio density will be around 1. Therefore, a usual calibration practice is to move the ratio histogram mode to 1 when the red and green signals are not calibrated. This calibration procedure is not strictly correct because the peak of the ratio density changes with parameter c. To account for this effect, we first assume the population mean μ_0 to be 1 and let the first approximation m_1 of the calibration parameter be the sample mean. The sample data are then calibrated by m_1. After that, Eq. (10) is used to estimate the first approximation $\hat{c}_1$ of c. Estimation proceeds by iteratively repeating the procedure. The following algorithm results:

1. Initialize mean estimate $\hat{\mu}_0$ of the ratio density of Eq. (7) to be 1 (equivalent to assuming $c_0=0$).
2. Calibrate ratio samples so that the input red and green signals are approximately equal by taking the estimator of m, say $\hat{m}_i$, to be the sample mean divided by the previous mean estimator,
$$\hat{m}_i=\frac{1}{\hat{\mu}_{i-1}}\left(\frac{1}{n}\sum_{j=1}^{n}t_j\right). \tag{12}$$
The calibration factor is taken to be $1/\hat{m}_i$. The normalized ratio data set is
$$(t_1',t_2',\ldots,t_n')=(t_1/\hat{m}_i,t_2/\hat{m}_i,\ldots,t_n/\hat{m}_i). \tag{13}$$
3. Use the maximum-likelihood estimator of Eq. (10) to calculate $\hat{c}_i$ by evaluating the estimator with the newly calibrated ratio data (t_i', $i=1,2,\ldots,n$).
4. Estimate the mean $\hat{\mu}_i$ of the ratio distribution given the new $\hat{c}_i$ by using the polynomial regression given in Table 2 ($\mu=0.364c^3+1.279c^2-0.0427c+1.001$).
5. Repeat steps 2 through 4 until a satisfactory result is obtained. Since the ratio mean μ is close to 1 for even relatively large values of c, five iterations are usually sufficient.
6. Upper and lower confidence limits (θ_1,θ_2) can be obtained using Tables 1 or 2, and then converting them to the desired interval ($\theta_1\cdot\hat{m}$, $\theta_2\cdot\hat{m}$).

Table 2 Parameters of fitting polynomial functions.

Conf. level		a_3	a_2	a_1	a_0	Goodness of fit (R^2)
	Lower limit	−2.805	2.911	−2.706	0.979	0.999994
95%	Upper limit	28.644	−2.830	3.082	0.989	0.99993
	Lower limit	−5.002	4.462	−3.496	0.9968	0.99998
99%	Upper limit	78.349	−15.161	4.810	0.9648	0.99998
	Mean (μ)	0.364	1.279	−0.0427	1.001	0.9997
	SD	6.259	0.190	1.341	0.00225	0.9998

To verify the accuracy of the iterative method under the H_0 condition ($\mu_{R_k}=\mu_{G_k}$), we performed the following simulation assuming 100 red and 100 green intensity data points. For $k=1,2,...,100$, the kth red signal's (representing the kth gene expression level in sample 1) mean intensity μ_{R_k} is drawn from a uniform random process with a range from 100 to 30,000 (simulating a 16-bit integer range). For a given m and c value, along with the normality for both red and green signals, we generate a single datum for both the kth red and green signals, thereby obtaining a sample of the red/green ratio for each k. Simulations were done for m from 0.3 to 3 with a step of 0.1, and for c from 0.01 to 0.3 with a step of 0.01, each simulation involving the full iterative procedure. The entire simulation was repeated 30 times for each value of m and c. Average estimation errors for c and m are under 1% where the error for c is defined as $|(\hat{c}-c)/c|$ and similarly for m.

7 EXPERIMENTAL RESULTS

Consider the superimposed microarray image from UACC-903 (red channel) and UACC-903 (+6) (green channel) shown in Figure 2. The full array contains 1368 clone segments. A total of 88 ratio samples whose ratios are believed to be about 1 (whose gene expression levels are assumed unchanged in both cell lines), such as housekeeping genes, are listed in Web site http://www.nhgri.nih.gov/DIR/LCG/ARRAY/expn.html. Since acquisition does not ensure perfect calibration, the iterative procedure is used. The result is as follows:

m	1.1316
c	0.1727 (or 17.27%)
99% confidence interval:	(0.566, 1.977)

The step-by-step iterative estimation is shown in Table 3. The 99.0% confidence interval for $c=0.1727$ and $m=1.1316$ is (0.56617, 1.97684).

Based on this interval, 92 ratio samples are found to be significant. Of these, 70 were found to be significant using the inappropriately narrow confidence interval of Ref. 4 (see Web site http://www.nhgri.nih.gov/DIR/LCG/ARRAY/expn.html). Table 4 lists the ones missed by the confidence limits of Ref. 4. Some of the newly found significant changes are biologically in-

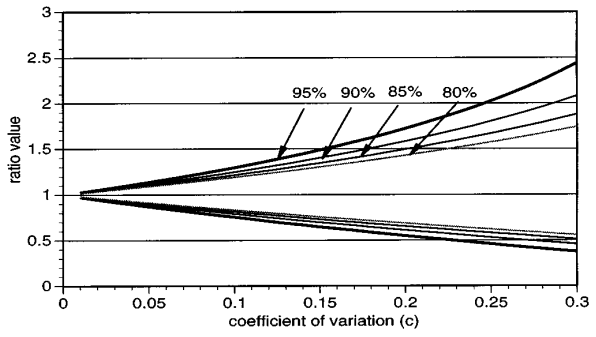

Fig. 6 Limits for different confidence levels.

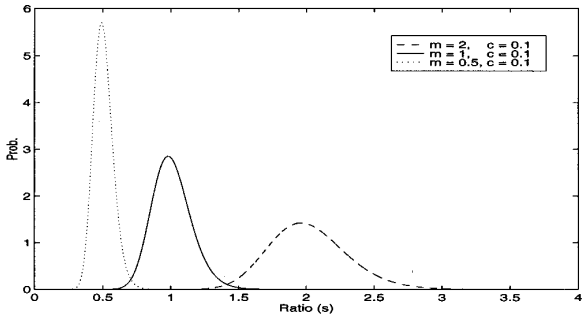

Fig. 7 Ratio density functions for $m=0.5$, 1, and 2 when $c=0.1$.

Table 3 Step-by-step illustration of the iterative estimation.

Step i	Sample scaling factor	c_i (Eq. 10)	μ_i (Table 2)	m_i (Eq. 12)
Initial	—	—	$\mu_0 = 1.0$	$m_0 = 1.1697$
1	$1/m_0$	0.1741	1.03425	1.1420
2	$1/m_1$	0.1728	1.03370	1.1315
3	$1/m_2$	0.1727	1.03365	1.1316
4	$1/m_3$	0.1727	Stop!	—

teresting and further bolster general impressions resulting from the original cohort of genes showing significant changes.

Two further examples of the tendency of the chromosome 6 suppressed line toward increased expression of genes associated with differentiation are the myeloid leukemia cell differentiation protein (mcl1) and the cell adhesion regulator protein (CAR/CMAR). Increased expression of the mcl1 gene has been found to be a very early indicator of induced differentiation in cancer cells.[12,13] Increased expression of the CAR gene has been correlated with reduced spontaneous metastatic potential in the HT-29 (human adenocarcinoma) cell line,[14] presumably due to a greater repertoire of integrins, with increased adherence of the cells to the extracellular matrix. In addition to the tendency toward expression of genes associated with differentiation, changes are observed that suggest that the suppressed cells are more capable of modulating oncogene activity. In addition to the strong increase in p21 expression previously seen, a significant increase in the expression of the *ras* suppressor Rsu-1 is observed. Rsu-1 has been shown to be a potent inhibitor of Jun kinase activation.[15]

8 CONCLUSION

Ratios are used to quantify gene expression distinctions on a cDNA microarray arising from different samples. Under the mathematical conditions assumed for average mRNA expression intensities, the ratio distribution has been derived, maximum-likelihood estimation characterized, and calibration achieved via an iterative algorithm. Empirically, a careful mathematical analysis of calibration and confidence limits has revealed significant gene expression ratios that were missed with a less precise analysis.

Table 4 Additional genes showing different expression levels.[a]

Gene name	R/G Ratio
Pre-mRNA splicing factor SRp7	2.33
Casein kinase I delta	2.33
MAC25	2.32
Endothelin-1 (EDN1)	2.30
B12 protein	2.25
RSU-1/RSP-1	2.25
Id1	2.24
Similar to induced myeloid leukemia cell differentiation protein	2.22
Male-enhanced antigen mRNA (Mea)	2.20
PP15 (placental protein 15)	2.20
Vascular endothelial GF	2.18
Calphobindin II	2.18
Similar to mouse transplantation antigen p35B	2.15
22 kDa smooth muscle protein (SM22)	2.15
Alternative guanine nucleotide-binding regulatory protein (G)	2.13
Nuclear autoantigen GS2NA	2.13
Cadherin-associated protein-related (cap-r)	2.13
Mitochondrial phosphate carrier protein	2.12
Alpha NAC	2.10
Thymopoietin beta	2.08
B-lymphocyte serine/threonine protein kinase	2.07
Platelet alpha SNAP	2.06
Lamin B2 (LAMB2)	2.06
CMAR	2.06
Inosine-5′-monophosphate dehydrogenase (IMP)	2.02
I-Rel	1.99
DNA-binding protein (CROC-1A)	1.99
PolyA binding protein	1.98
bcr (break point cluster gene)	0.57
Mitotic feedback control protein Madp2 homolog	0.55
Protein-tyrosine phosphatase	0.55
Human poly(ADP-ribose) synthetase	0.53

[a] The named genes, shown in decreasing ratio order, are additional genes found to have different expression levels in a chromosome 6–suppressed melanoma cell line than in the tumorigenic parent (99% confidence level). The original findings were reported in Ref. 4.

REFERENCES

1. M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, ''Quantitative monitoring of gene expression patterns with a complementary DNA microarray,'' *Science* **270**(5235), 467–470, (1995).
2. M. S. Boguski and G. D. Schuler, ''Establishing a human transcript map,'' *Nature Genetics* **10**(4), 369–371 (1995).
3. G. D. Schuler, M. S. Boguski, et al., ''A gene map of the human genome,'' *Science* **274**(5287), 540–546 (1996).
4. J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent, ''Use of a cDNA microarray to analyse gene expression patterns in human cancer,'' *Nature Genetics* **14**(4), 457–460 (1996).
5. M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis, ''Parallel human genome analysis: microarray-based expression monitoring of 1000 genes,'' *Proc. Nat. Acad. Sci. U.S.A.* **93**(20), 10614–10619 (1996).
6. E. R. Dougherty, *Probability and Statistics for the Engineering, Computing, and Physical Sciences*, Prentice-Hall, Englewood Cliffs, NJ (1990).
7. D. Feldman and M. Fox, *Probability, the Mathematics of Uncertainty*, Marcel Dekker, New York (1991).
8. R. R. Sokal and F. J. Rohlf, *Biometry: the Principles and Practice of Statistics in Biological Research*, 3rd ed., W. H. Freeman, New York (1995).
9. S. Shanmugalingam, ''On the analysis of the ratio of two correlated normal variables,'' *The Statistician* **31**(3), 251–258 (1982).
10. H. Schneeberger and K. Fleischer, ''The distribution of the ratio of two variables,'' *J. Statist. Comput. Simulation* **47**, 227–240 (1993).
11. P. J. Korhonen and S. C. Narula, ''The probability distribution of the ratio of the absolute values of two normal variables,'' *J. Statist. Comp. Simulation* **33**, 173–182 (1989).
12. A. Umezawa, T. Maruyama, J. Inazawa, S. Imai, T. Takano, and J. Hata, ''Induction of mcl1/EAT, Bcl-2 related gene, by retinoic acid or heat shock in the human embryonal carcinoma cells, NCR-G3,'' *Cell Struct. Funct.* **21**(2), 132–150 (1996).
13. T. Yang, H. L. Buchan, K. J. Townsend, and R. W. Craig, ''MCL-1, a member of the BLC-2 family, is induced rapidly in response to signals for cell differentiation or death, but not to signals for cell proliferation,'' *J. Cell Physiol.* **166**(3), 523–536 (1996).
14. H. Yamamoto, F. Itoh, Y. Hinoda, and K. Imai, ''Inverse association of cell adhesion regulator messenger RNA expression with metastasis in human colorectal cancer,'' *Cancer Res.* **56**(15), 3605–3609 (1996).
15. L. Masuelli and M. L. Cutler, ''Increased expression of the Ras suppressor Rsu-1 enhances Erk-2 activation and inhibits Jun kinase activation,'' *Molec. Cell Biol.* **16**(10), 5466–5476 (1996).