JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 7, Number 6, 2000
Mary Ann Liebert, Inc.
Pp. 805–817

Testing for Differentially-Expressed Genes by Maximum-Likelihood Analysis of Microarray Data*

TREY IDEKER,[1,3] VESTEINN THORSSON,[1,3] ANDREW F. SIEGEL,[1,2,3] and LEROY E. HOOD[1,3]

ABSTRACT

Although two-color fluorescent DNA microarrays are now standard equipment in many molecular biology laboratories, methods for identifying differentially expressed genes in microarray data are still evolving. Here, we report a refined test for differentially expressed genes which does not rely on gene expression ratios but directly compares a series of repeated measurements of the two dye intensities for each gene. This test uses a statistical model to describe multiplicative and additive errors influencing an array experiment, where model parameters are estimated from observed intensities for all genes using the method of maximum likelihood. A generalized likelihood ratio test is performed for each gene to determine whether, under the model, these intensities are significantly different. We use this method to identify significant differences in gene expression among yeast cells growing in galactose-stimulating versus non-stimulating conditions and compare our results with current approaches for identifying differentially-expressed genes. The effect of sample size on parameter optimization is also explored, as is the use of the error model to compare the within- and between-slide intensity variation intrinsic to an array experiment.

Key words: DNA microarray, gene expression, maximum likelihood, statistical error analysis, yeast.

INTRODUCTION

DNA MICROARRAYS HAVE REVOLUTIONIZED the study of gene expression and are now a staple of biological inquiry. Using the microarray, it is possible to observe the expression level changes in tens of thousands of genes over multiple conditions, all in a single experiment. Depending on the conditions assayed, differentially expressed genes may be implicated in cancer (DeRisi *et al.*, 1996; Wang *et al.*, 1999), aging (Ly *et al.*, 2000), or a metabolic pathway of interest (DeRisi *et al.*, 1997; Roberts *et al.*, 2000).

[1]Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195.
[2]Departments of Management Science, Finance, and Statistics, University of Washington, Seattle, WA 98195.
[3]The Institute for Systems Biology, Seattle, WA 98104.
*Software is available at www.systemsbiology.org/VERAandSAM

A crucial step in the analysis of expression data is determining which genes are expressed differently between two cell populations. Usually, a gene is said to be "differentially-expressed" if its ratio of expression level in one population to expression level in a second population exceeds a certain threshold (DeRisi *et al.*, 2000; Iyer *et al.*, 1999). This threshold is set based on the observation that, in control experiments where the two cell populations are identical, few if any genes have expression ratios exceeding the threshold. However, it is common knowledge that this approach is imprecise, because the uncertainty in the expression ratio is greater for genes that are expressed at low levels than for those that are highly expressed. More sensitive methods have been employed in a few cases (Chen *et al.*, 1997; Greller and Tobin, 1999; Hilsenbeck *et al.*, 1999; Roberts *et al.*, 2000), but development of a general, formal statistical test for identifying differentially-expressed genes has remained an open problem.

To address this need, we now present an error model and an associated significance test which together provide a substantial improvement over the thresholding scheme. Using the method of maximum likelihood, model parameters are estimated from a data set consisting of gene-expression measurements obtained over repeated experiments using a standard two-color DNA microarray. Once the error structure of the model has been estimated, a generalized likelihood ratio test determines whether, for each gene, the expression levels observed with the microarray are significantly different between the two cell populations assayed. Several applications of the approach are discussed, including use of this method to identify significant differences in gene expression among yeast cells growing in galactose *vs.* raffinose. We demonstrate that, in comparison to the popular ratio-based approach, our method may occasionally implicate genes as differentially expressed even if their average expression ratio is close to one, and will not always implicate genes having extreme expression ratios. As RNA labeling and hybridization become fully automated and collection of repeated measurements becomes routine, we expect that modeling strategies such as the one proposed here will become indispensable for characterizing experimental variation involved in the array process and making decisions based on this variation.

MATERIALS AND METHODS

Experimental setup

We consider a microarray consisting of a large number of spots of DNA on glass, each containing the full open-reading-frame sequence of a gene (see Lander [1999] for a thorough review). Briefly, mRNA contained in each of two populations of cells is extracted, reverse-transcribed into cDNA, and labeled with either Cy3 or Cy5 dye. Cy3 and Cy5 preparations are combined and deposited on the microarray, where labeled molecules hybridize to the spot containing their complementary sequence. The amount of hybridization to each spot is quantified by scanning the array with a laser and observing the intensity of light emitted. Observations are made separately for the two dyes, such that two intensities x and y are observed for each spot on the microarray. This process does not behave deterministically in practice, such that multiple spots corresponding to each gene i hybridized under identical conditions will result in a distribution of intensities x_{ij} and y_{ij} ($1 \leq i \leq N$; $1 \leq j \leq M$), where N is the number of genes represented on the microarray and M is the number of spots observed for each gene.

Preprocessing of microarray data

Spot intensities are extracted from a scanned image then background-subtracted and normalized as follows. Microarray images are processed with Dapple, a software tool we have developed for array spot finding and quantitation (Buhler *et al.*, 2000). Dapple locates each spot and reports the median foreground intensity inside the spot area separately for each of the two dyes. It also provides a local background intensity estimate for each spot and dye: we smooth these estimates by spatial filtering using a 7-spot $\times$ 7-spot median filter (Lim, 1990). This smoothed background is then subtracted from the foreground of each spot to produce the background-subtracted intensities x' and y'.

In practice, x' and y' have different scales and thus are not directly comparable. This situation may occur if the total amount of labeled cDNA is greater for one dye than the other, if one dye incorporates more

efficiently or if the scanner has different sensitivities to the two dyes. Therefore, intensities are normalized to have identical medians A within each array hybridization:

$$x = \frac{Ax'}{\tilde{x}'} \quad y = \frac{Ay'}{\tilde{y}'} \quad A = \frac{1}{2}(\tilde{x}' + \tilde{y}') \tag{1}$$

where $\tilde{x}'$ denotes the median intensity of x' over all spots on a single microarray. If multiple array hybridizations are performed, normalization occurs independently for each and the resulting combined data set consists of data pairs (x_{ij}, y_{ij}) for gene i in repeat j. If three or more samples are available for a gene, these are filtered to remove outliers by Dixon's test with $\alpha = 0.1$ (Dunn and Clark, 1987), independently in x and y. We also remove extremely high intensities that are outside the dynamic range of the array scanner in either color.

Error model

Based upon extensive exploratory data analysis, we have formulated a mathematical model summarizing the influence of multiplicative and additive errors on x and y. We have consistently observed that larger intensity measurements have a proportionately larger error over repeated samples, *i.e.*, have a constant coefficient of variation $\sigma_{x'} \propto x'$ (see Fig. 1a), as would be caused by variation in spot size or labeling efficiency from gene to gene. However, the variability does not tend to zero as $x \to 0$, likely due to variation in the measured background intensity. We have also observed that, within genes, x and y are correlated and that larger intensities have a larger correlation, possibly due to errors introduced by spot-to-spot nonuniformity or during the hybridization process which affect intensity measurements for both dyes simultaneously (Fig. 1b). Finally, samples of x and y for a given gene are at least approximately normally distributed, as assessed by a normal probability plot (Dunn and Clark, 1987) (Fig. 1c).

Motivated by these observations, we postulate that the background-subtracted, median-normalized intensities observed for each gene are related to their true (or mean) intensities by the following model:

$$\begin{aligned} x_{ij} &= \mu_{x_i} + \mu_{x_i}\varepsilon_{x_{ij}} + \delta_{x_{ij}} \\ y_{ij} &= \mu_{y_i} + \mu_{y_i}\varepsilon_{y_{ij}} + \delta_{y_{ij}} \end{aligned} \tag{2}$$

where (μ_{x_i}, μ_{y_i}) is the pair of true mean intensities for gene i. For each i and j, the multiplicative errors $\varepsilon_{x_{ij}}$ and $\varepsilon_{y_{ij}}$ are drawn from a bivariate normal distribution with mean 0, standard deviations σ_{ε_x} and σ_{ε_y}, and correlation ρ_ε. The additive errors $\delta_{x_{ij}}$ and $\delta_{y_{ij}}$ are distributed analogously, with parameters σ_{δ_x}, σ_{δ_y}, and ρ_δ. Thus, multiplicative and additive errors are independent of one another but can each be highly

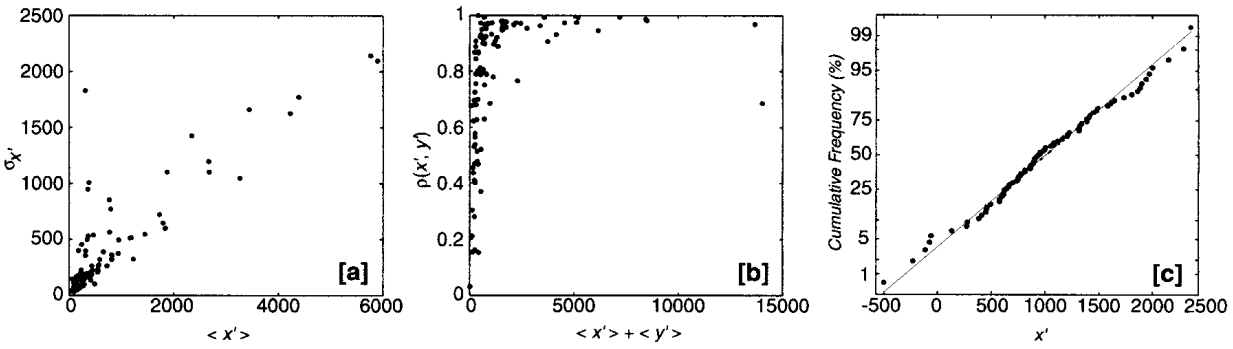

FIG. 1. Increase of standard deviation [**a**] and correlation [**b**] with absolute level of intensity x' or y'. Data were obtained over 5 separate hybridizations with identically-prepared Cy3- and Cy5-labeled cDNA mixtures to test arrays (described in Results) containing 16 replicate spots per gene over 96 genes, resulting in a total of 80 samples for each of 96 genes. [**c**] Normal probability plot for the 80 samples of x' pertaining to a single, representative gene. This plot is linear, indicating that these data are consistent with a normal distribution. The line connects the 25th and 75th percentiles of the data and represents an approximate linear fit.

correlated between x and y; in practice we have found that ρ_ε is large and that ρ_δ is small. Note that (x_{ij}, y_{ij}) may be negative if by chance foreground is less than estimated background for a spot, but that the true intensities (μ_{x_i}, μ_{y_i}) must be nonnegative.

Consequently, the samples (x_{ij}, y_{ij}) are described by a bivariate normal probability density function p with parameters μ_{x_i}, μ_{y_i}, σ_{x_i}, σ_{y_i}, and $\rho_{x_i y_i}$, where:

$$\begin{aligned} \sigma_{x_i} &= \sqrt{\mu_{x_i}^2 \sigma_{\varepsilon_x}^2 + \sigma_{\delta_x}^2} \\ \sigma_{y_i} &= \sqrt{\mu_{y_i}^2 \sigma_{\varepsilon_y}^2 + \sigma_{\delta_y}^2} \\ \rho_{x_i y_i} &= \frac{\mu_{x_i} \mu_{y_i} \rho_\varepsilon \sigma_{\varepsilon_x} \sigma_{\varepsilon_y} + \rho_\delta \sigma_{\delta_x} \sigma_{\delta_y}}{\sigma_{x_i} \sigma_{y_i}}. \end{aligned} \tag{3}$$

The model depends on six gene-independent parameters $\boldsymbol{\beta} = (\sigma_{\varepsilon_x}, \sigma_{\varepsilon_y}, \rho_\varepsilon, \sigma_{\delta_x}, \sigma_{\delta_y}, \rho_\delta)$ and a mean pair per gene, $\boldsymbol{\mu} = [(\mu_{x_1}, \mu_{y_1}), (\mu_{x_2}, \mu_{y_2}), \ldots, (\mu_{x_N}, \mu_{y_N})]$, for a total of $2N + 6$ parameters. The probability density function for gene i is $p = p(x_{ij}, y_{ij} | \boldsymbol{\beta}, \mu_{x_i}, \mu_{y_i})$.

Parameter estimation by maximum likelihood

Since $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ are generally unknown, we estimate them using maximum-likelihood estimation (MLE) (Kendall and Stuart, 1979). Likelihood functions, for gene i and over all genes, are respectively defined as:

$$\begin{aligned} L_i(\boldsymbol{\beta}, \mu_{x_i}, \mu_{y_i}) &= \prod_{j=1}^{M} p(x_{ij}, y_{ij} | \boldsymbol{\beta}, \mu_{x_i}, \mu_{y_i}) \\ L(\boldsymbol{\beta}, \boldsymbol{\mu}) &= \prod_{i=1}^{N} L_i(\boldsymbol{\beta}, \mu_{x_i}, \mu_{y_i}). \end{aligned} \tag{4}$$

The MLE parameter values maximizing L, designated $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mu}}$, are our estimates for the true parameters of the underlying statistical model. In general, these values may be found using standard optimization procedures (Press *et al.*, 1992). Because N can be large, we determine $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mu}}$ by optimizing subsets of parameters in separate stages:

(1) Choose initial values for $\boldsymbol{\mu}$.
(2) Select $\boldsymbol{\beta}$ to maximize L given current values of $\boldsymbol{\mu}$.
(3) For $i = 1, \ldots, N$: select (μ_{x_i}, μ_{y_i}) to maximize L_i, given current values of $\boldsymbol{\beta}$.
(4) Repeat (2) and (3) until $\boldsymbol{\beta}$, $\boldsymbol{\mu}$ have converged.

All stages of the optimization are performed using the procedure *fmincon* provided by Matlab (Coleman *et al.*, 1999). We have also implemented the optimization in C code, which produces comparable optimal parameters in substantially less execution time (<10 min. on a Pentium III 500 for $N = 6000$, $M = 4$, as compared with 4–5 hr for the Matlab implementation). In either case, we find that all parameters converge with 250 iterations of stages (2) and (3) and are insensitive to initial choices for $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$.

Significance testing using likelihood ratios

After the parameters have been determined for a given set of observations, it is of immediate interest to use the model to identify mean intensity pairs which are significantly unequal ($\mu_{x_i} \neq \mu_{y_i}$), representing genes that are differentially expressed between the two cell populations. For each gene i, we compute the generalized likelihood ratio test (GLRT) (Kendall and Stuart, 1979) statistic λ_i according to:

$$\lambda_i = -2 \ln \left(\frac{\max_\mu L_i(\hat{\boldsymbol{\beta}}, \mu, \mu)}{\max_{\mu_x, \mu_y} L_i(\hat{\boldsymbol{\beta}}, \mu_x, \mu_y)} \right). \tag{5}$$

Two maximizations are performed: in the numerator, the constraint $\mu_x = \mu_y = \mu$ is imposed, while in the denominator the optimization is unconstrained. When the constraint is imposed, $\hat{\boldsymbol{\beta}}$ remains a consistent estimator.

In the case that $\mu_{x_i} = \mu_{y_i}$, λ_i follows (asymptotically in M and N) a χ^2 distribution with 1 degree of freedom (DOF), whereas if $\mu_{x_i} \neq \mu_{y_i}$ the value of λ_i is expected to be larger than would be obtained from random sampling of this distribution. To select differentially expressed genes with a selection error of α (the false positive, or type-I error rate), one would first determine the critical value λ_c for which the χ^2 cumulative probability distribution is equal to $1 - \alpha$, then select the set of all genes i for which λ_i is in the critical region $\lambda_i > \lambda_c$. The particular choice of α depends on the number of genes on the array and the selection error which the individual investigator is willing to tolerate.

Laboratory protocols

Preparation of mRNA. Wild-type yeast (BY4741) or a congenic *gal80*Δ strain is inoculated in 100 ml of either galactose-inducing YPRG media (1% yeast extract, 2% peptone, 2% raffinose, 2% galactose) or noninducing YPR media (1% yeast extract, 2% peptone, 2% raffinose). Cultures are grown at 30°C to a density of 1–2 OD_{600}, and total RNA harvested by hot acidic phenol extraction (Ausubel *et al.*, 1995). Poly-A purification from total RNA is performed using Ambion Poly(A)Pure mRNA Isolation Kits (#1915).

Fabrication of DNA microarrays. A set of approximately 6,200 known and predicted gene open reading frames from the yeast *Saccharomyces cerevisiae* (Research Genetics) are amplified in separate 100 μL PCR reactions in 384-well plate format. PCR conditions are optimized depending on the length of the template, but in general are as follows: 95°C for 2 min, 35 cycles of {94°C for 30 sec, 64°C for 30 sec, 72°C for 2.5 min}, followed by 72°C for 5 min. Reactions are purified over Sephacryl S500, and the purified product is added to DMSO in a 1:1 ratio. A Molecular Dynamics Generation III microarray robotic spotter prints these products onto 25 × 75 mm glass slides (Amersham #RPK0328). Slides are spotted at 50% humidity then immediately UV cross-linked at 50 mJ of energy.

cDNA synthesis and hybridization. Two μg anchored dT25 primers and 2 μg random 9-mer primers are added to 4 μg poly-A selected mRNA and allowed to anneal at 70°C for 5 min in a 12 μL volume. After 1–2 min on ice, 4 μL 5× Superscript II buffer (Gibco), 2 μL 0.1M dTT, 1 μL dNTP mix (10 mM dATP, dTTP, dGTP, and 1 mM dCTP), 1 mM of either Cy3 or Cy5 fluorescent dye (Amersham), and 1 μL Superscript II RT are added. Reverse transcription occurs at 42°C for 2–2.5 hr in the dark. After this time, the RNA is hydrolyzed by heating at 94°C for 3 min, addition of 1 μL of 5M NaOH, and incubation at 37°C for 10 min. The pH is then adjusted by the addition of 1 μL 5M HCl and 5 μL 1M Tris (pH 6.8), and the cDNA purified through Millipore NAB plates. Dye incorporation is assessed by measuring absorbance at 550 and 650 nm, and a sample aliquot containing ~40 pmol of dye is concentrated to <5 μL. Subsequent to labeling, purification, and concentration, Cy3 and Cy5 samples are combined and suspended in 40–45 μL hybridization solution (50% formamide, 5× Denhardt's, 5× SSC, 0.1% SDS). This mixture is applied to the array slide beneath a coverslip and allowed to incubate in a sealed, humid chamber overnight (16–18 hr) at 42°C. The slide is then washed in 2× SSC/0.1% SDS for 5 min at 42°C, followed by 0.1× SSC, 0.1% SDS for 5 min at room temp and 2 additional washes in 0.1× SSC for 2 min each. The slide is rinsed briefly in d^2H_20 and immediately dried with compressed air. After hybridization and washing, array slides are scanned using a scanning laser fluorescence microscope (Molecular Dynamics Generation II Scanner).

RESULTS

Identification of genes differentially expressed in response to galactose stimulation of yeast cells

In order to explore the performance of our test for differentially-expressed genes, we compared cultures of the yeast *Saccharomyces cerevisiae* growing in the absence of galactose (YPR media) to those growing

in galactose-stimulating conditions (YPRG) using a DNA microarray of approximately 6200 nuclear yeast genes. Each gene was represented by two spots located on opposite sides of the array. We obtained a total of four (x, y) intensity pairs for each gene by performing replicate hybridizations to two of these microarrays ($N = 6200$, $M = 4$), with x and y representing intensities in YPR and YPRG respectively. In the first hybridization, RNA from the YPR condition was labeled with Cy3 dye while RNA from the YPRG condition was labeled with Cy5 dye; in the second hybridization, the reverse labeling scheme was used. Using our maximum likelihood approach, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mu}}$ were determined for these data and the λ_i statistic was computed for each gene. Values for $\hat{\boldsymbol{\beta}}$ were (0.367, 0.291, 0.862, 89.6, 339.0, 0.319).

In order to determine a reasonable choice for the critical value λ_c used to select differentially expressed genes, we performed a series of control experiments in which two cell populations were cultured separately but using otherwise identical strains and YPRG growth conditions. These two populations were compared exactly as before by obtaining a total of $M = 4$ repeat samples per gene and determining values of $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}$, and λ. In general, these control data had fewer large values of λ than did the YPR *vs.* YPRG data and followed a χ^2 distribution (as determined by a q–q plot, data not shown). However, both data sets had significantly larger values of λ than expected for a χ^2 with 1 DOF (see Materials and Methods). This result could be due to the small-sample bias of maximum likelihood methods, resulting in λ_i statistics that are not χ^2 with 1 DOF even if $\mu_{x_i} = \mu_{y_i}$ for all i.

Based on these control experiments, we selected a critical value λ_c such that less than 0.1% of genes (approximately 6 out of 6,200) had $\lambda > \lambda_c$. This value, $\lambda_c = 23.8$, was then applied to select differentially expressed genes from the YPR *vs.* YPRG data. Scatter plots of estimated μ_y *vs.* μ_x values for each gene are shown for the control experiment (Fig. 2a) and the YPR versus experiment (Fig. 2b). Red data points denote genes with $\lambda_i > \lambda_c$. The most highly significant genes (out of a total of 456 selected as significant) are shown in Table 1. These are in good agreement with previous experimental evidence (Lohr *et al.*, 1995), with the galactose-induction pathway structural genes (*GAL1*, *GAL7*, and *GAL10*) appearing as the top three most significant differentially-expressed genes.

Effect of sample size on parameter estimates

As expected, the more genes and samples per gene that are available, the more accurate are estimates of the error model parameters. To test the efficacy of parameter estimation, we used Equations (2) and (3) with fixed parameters $\boldsymbol{\beta}_{\mathbf{sim}}$ and $\boldsymbol{\mu}_{\mathbf{sim}}$ to randomly simulate data sets of several different sizes $M \times N$. $\boldsymbol{\beta}_{\mathbf{sim}}$ and $\boldsymbol{\mu}_{\mathbf{sim}}$ were set to the corresponding values optimized for the YPR *vs.* YPRG experiment; in simulations with $N < 6{,}200$, we randomly selected N of the 6,200 optimized (μ_x, μ_y) pairs. One hundred data sets were simulated for each choice of M and N, after which $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mu}}$ were estimated for each data set and the resulting distribution of $\hat{\boldsymbol{\beta}}$ characterized by parameter means $\langle \boldsymbol{\beta} \rangle$ and standard deviations $\mathbf{s}_\beta$.

Table 1. Genes Differentially Expressed between Galactose Noninducing (YPR) and Inducing (YPRG) Conditions

Gene	Cellular role	λ	μ_x	μ_y	μ_y/μ_x
GAL1	galactose metabolism	95.4	145	110644	766
GAL10	galactose metabolism	88.1	109	36656	338
GAL7	galactose metabolism	86.7	59	76849	1300
YNL194C	unknown	75.0	18533	1360	0.073
JEN1	transport	72.2	21124	889	0.042
YNL195C	unknown	72.0	7639	710	0.093
ALD6	ethanol utilization	71.5	9774	517	0.053
RHR2	glycerol metabolism	71.1	1181	22586	19
YMR318C	unknown	69.1	2457	29930	12
HSP26	diauxic shift	68.1	71988	11435	0.16

In simulations with $M = 50$, $N = 100$, parameter estimates were tightly distributed around their true values such that $\langle \boldsymbol{\beta} \rangle = \boldsymbol{\beta}_{\mathbf{sim}} \pm 4\%$ and $\mathbf{s}_\beta \leq (.09)\langle \boldsymbol{\beta} \rangle$ for all parameters $\hat{\boldsymbol{\beta}}$. In contrast, for very small data sets with $M = 4$, $N = 100$, we found that these estimates were highly variable over the 100 simulations ($\mathbf{s}_\beta \leq (.46)\langle \boldsymbol{\beta} \rangle$) and biased: $\boldsymbol{\beta}_{\mathbf{sim}}$ was under- or overestimated by 1–21% across the six parameters of $\langle \boldsymbol{\beta} \rangle$. In order to more closely model experiments performed with a yeast microarray, we also examined simulations with $M = 4$, $N = 6{,}200$. Estimates were generally biased, but this bias was smaller ($\langle \boldsymbol{\beta} \rangle = \boldsymbol{\beta}_{\mathbf{sim}} \pm 14\%$) and the variability of estimation also less ($\mathbf{s}_\beta \leq (.06)\langle \boldsymbol{\beta} \rangle$). Thus, with regard to parameter estimation, a large number of genes appears to at least partially compensate for the destabilizing effect of a small number of repeats.

We also explored the effect of sample size on significance testing in the YPR *vs.* YPRG experiment. Values for $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}$, and λ were determined using just two of the available four samples per gene by drawing one spot per gene over the two replicate hybridizations. In this case, the number of genes selected as differentially expressed was less (178 genes using $\lambda_i > 23.8$), although 85% of these genes were previously identified as significant when using four samples per gene. The genes *GAL1*, *GAL7*, and *GAL10* were also identified as significant, but were no longer among the top ten with largest λ. While these genes still had a very extreme expression ratio (μ_y/μ_x), their intensity samples were by chance more variable than those of other genes with extreme expression ratios and thus their corresponding values of λ were smaller.

Ratios of intensity are approximately equal to ratios of hybridized cDNA

Although the proposed method identifies genes having different mean intensities μ_{x_i} and μ_{y_i}, in order to conclude that these genes are differentially expressed, intensity differences (or ratios) must be at least approximately proportional to differences in RNA copy number per cell. Since it seems plausible that either low or high copy number could lead to saturation in the measured intensity, we performed a series of controlled experiments to determine whether this relationship is linear over a reasonable range of copy number. First, a mixture of *gal80*Δ cDNA was created by extracting mRNA from yeast with a complete deletion of the *GAL80* gene, labeling it with Cy3 and Cy5 dyes in separate reactions, and then combining these reactions into one tube. This mixture was hybridized to a yeast genome microarray, and the resulting image checked to ensure that intensity was not detectable above background for spots representing *GAL80* and that all spots had roughly equal Cy3 and Cy5 intensities. Next, Cy3- and Cy5-labeled DNA sequences corresponding to the *GAL80* open reading frame were added to the *gal80*Δ mixture at fixed molar ratios of Cy3:Cy5 dye. As shown in the images of Figs. 3a and 3b, array hybridizations were performed for each of eight controlled *GAL80* ratios. Data sets consisting of four (x, y) intensity measurements per gene were obtained at each controlled *GAL80* ratio by using two spots from a forward (Cy3:Cy5) labeling scheme and two spots from a reverse (Cy5:Cy3) labeling scheme. Parameters $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mu}}$ were determined separately for each data set, and the corresponding measured ratio for *GAL80* was defined as μ_y/μ_x. Fig. 3c shows a scatter plot of each measured ratio *vs.* controlled ratio. The figure suggests that saturation occurs at the lower extreme but that the system is approximately linear over a range of three orders of magnitude. Except where the controlled ratio was equal to one, all measured *GAL80* ratios had $\lambda > 23.8$ and thus were differentially expressed by our likelihood test (see Fig. 3 inset table). Since these controlled ratios were represented by independent data sets involving different estimated parameters and different samples of *GAL80* intensity, λ values do not necessarily increase monotonically as the controlled ratios deviate farther from 1.0.

At the upper end of the investigated range, *GAL80* was added at 1,000 fmol and measured at 32,436 intensity units (average over four samples). Only 14 genes on the array had higher intensities, the two largest being *TDH3* (81,255 units) and *ENO2* (55,766 units). At the lower end of the range, *GAL80* was added at 0.2 fmol and measured at 284 units: approximately 1,000 genes had lower intensities. Presumably, these are not expressed or else are beneath the range of detection. We also examined the intensities of several genes whose RNA copy number per cell has been determined experimentally (Iyer and Struhl, 1996). The gene *TRP3* has been observed at 1.9 copies per cell in YPR media, and had a corresponding intensity of 597 $\pm$ 259 (avg $\pm$ stdev) in the YPR condition of the YPR *vs.* YPRG array experiment. In contrast, *GAL1* mRNA is present at <0.1 copies per cell in YPR and was not significantly above

TABLE 2. COMPARISON OF ERROR MODEL PARAMETERS FOR 5 *Within-Slide* AND 16 *Between-Slide* DATA SETS (SEE RESULTS)

Source of variation	σ_{ε_x}	σ_{ε_y}	ρ_ε	σ_{δ_x}	σ_{δ_y}
Within-slide mean	0.35	0.306	0.981	251	374
(stderr)	(.063)	(.061)	(.0069)	(49)	(105)
Between-slides mean	0.365	0.315	0.967	422	569
(stderr)	(.0084)	(.0073)	(.0017)	(12)	(13)

background intensity on our yeast array. Thus most yeast genes (approximately 4,000 to 5,000) appear to have intensities within the linear range of the microarray system and the lower limit of detection is between 0.1 and 1.9 copies/cell.

Use of the procedure to compare and contrast parameters over different types of repeat measurements

We sought to use our error model to compare the combined variability present across an entire experiment to that introduced during array hybridization and quantitation alone. For this purpose, we constructed a test microarray having 96 genes spotted 16 times each. Ten cultures involving identical strains and YPRG conditions were grown independently in separate containers, and RNA prepared from each of these. Five of these preparations were labeled using Cy3 while the remaining five were labeled using Cy5. These mixtures were combined in Cy3–Cy5 pairs, and each of the five pairs hybridized to separate test arrays.

Two types of data sets were drawn from these experiments. In the first type of data set, repeats were drawn from the 16 replicate spots per gene on a single array (*within-slide* data, $N = 96$, $M = 16$). Parameters $\hat{\boldsymbol{\beta}}$ were estimated by maximum likelihood, independently for data sets formed using each of the five test arrays. Mean and stderr values over the estimates are shown in Table 2 (row 1). In the second type of data set, repeats were drawn from a single spot of each gene on the array over the five hybridizations to separate test arrays (*between-slide* data, $N = 96$, $M = 5$). In this case, parameters $\hat{\boldsymbol{\beta}}$ were estimated 16 times, separately for data sets formed using each of the 16 spots per gene available on the array (row 2).

FIG. 2. Scatter plots of estimated μ_y *vs.* μ_x for each gene represented on the whole-yeast-genome microarray, shown for [**a**] the control experiment YPRG *vs.* YPRG and [**b**] the YPR *vs.* YPRG comparison. Genes with $\lambda_i > 23.8$ have significantly different μ_x and μ_y and are shown in red. To show detail, axes limits are truncated to 45,000: the maximum (μ_x, μ_y) observed was $(1.8 \times 10^5, 1.4 \times 10^5)$. [**c**] The distribution of four (x, y) pairs is shown for two genes in the YPR *vs.* YPRG comparison. Samples for each gene are denoted by red or black crosses respectively, with corresponding averages $(\langle x\rangle, \langle y\rangle)$ denoted by squares and MLE-estimated means (μ_x, μ_y) denoted by filled circles. Open circles represent the estimated means under the added constraint $\mu_x = \mu_y$. Pink and gray ellipses define regions containing 95% of the error model probability distribution at these constrained means for the red- and black-colored genes, respectively. Dotted lines of constant ratio, drawn through the origin and each constrained and unconstrained (μ_x, μ_y) pair, are shown for reference. Although the genes have similar average expression ratios $\langle x\rangle/\langle y\rangle$ (2.9 *vs.* 3.5 for the red- *vs.* black-colored gene), the red-colored gene was significant by our likelihood test ($\lambda = 37.4$). The black-colored gene was not ($\lambda = 13.8$), due to its compatibility with the constrained error model.

FIG. 3. Labeled *GAL80* DNA was spiked into *gal80*Δ *vs. gal80*Δ cDNA at eight controlled Cy3:Cy5 ratios, either by fixing *GAL80*-Cy3 at 10 fmol and modulating *GAL80*-Cy5 (forward-labeling) or else fixing *GAL80*-Cy5 and modulating *GAL80*-Cy3 (reverse-labeling). Array images corresponding to each forward- and reverse-labeling experiment are shown in [**a**] and [**b**], centered on one of two spots on the array complementary to *GAL80*. Four (x, y) samples were measured for the *GAL80* gene in each controlled ratio by combining two *GAL80* spots from the forward-array in [**a**] with two *GAL80* spots from the reverse-array in [**b**]. The scatter plot [**c**] compares each controlled ratio to measured ratio (y/x) for the forward-array (red dots) or reverse-array (green dots). The ratio of estimated means μ_y/μ_x is denoted by an open circle. The inset table shows values of λ for the *GAL80* gene in each of the eight controlled ratios.

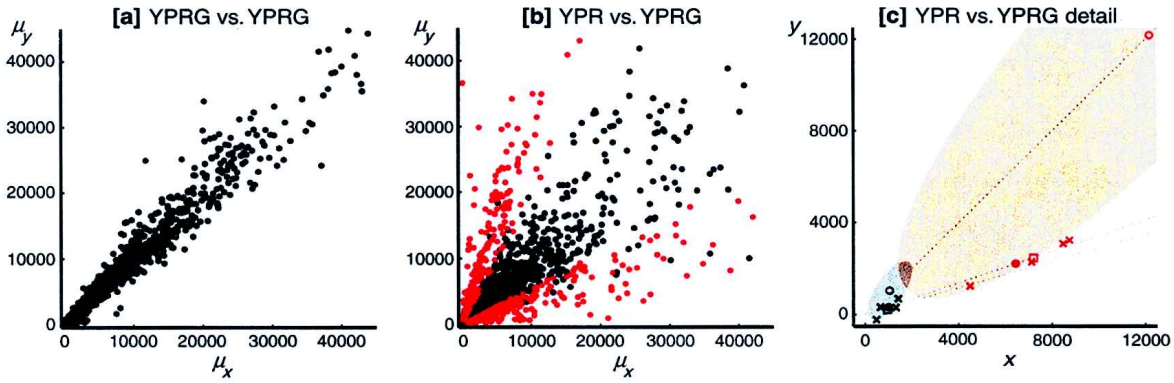

FIG. 2. (Caption on facing page.)

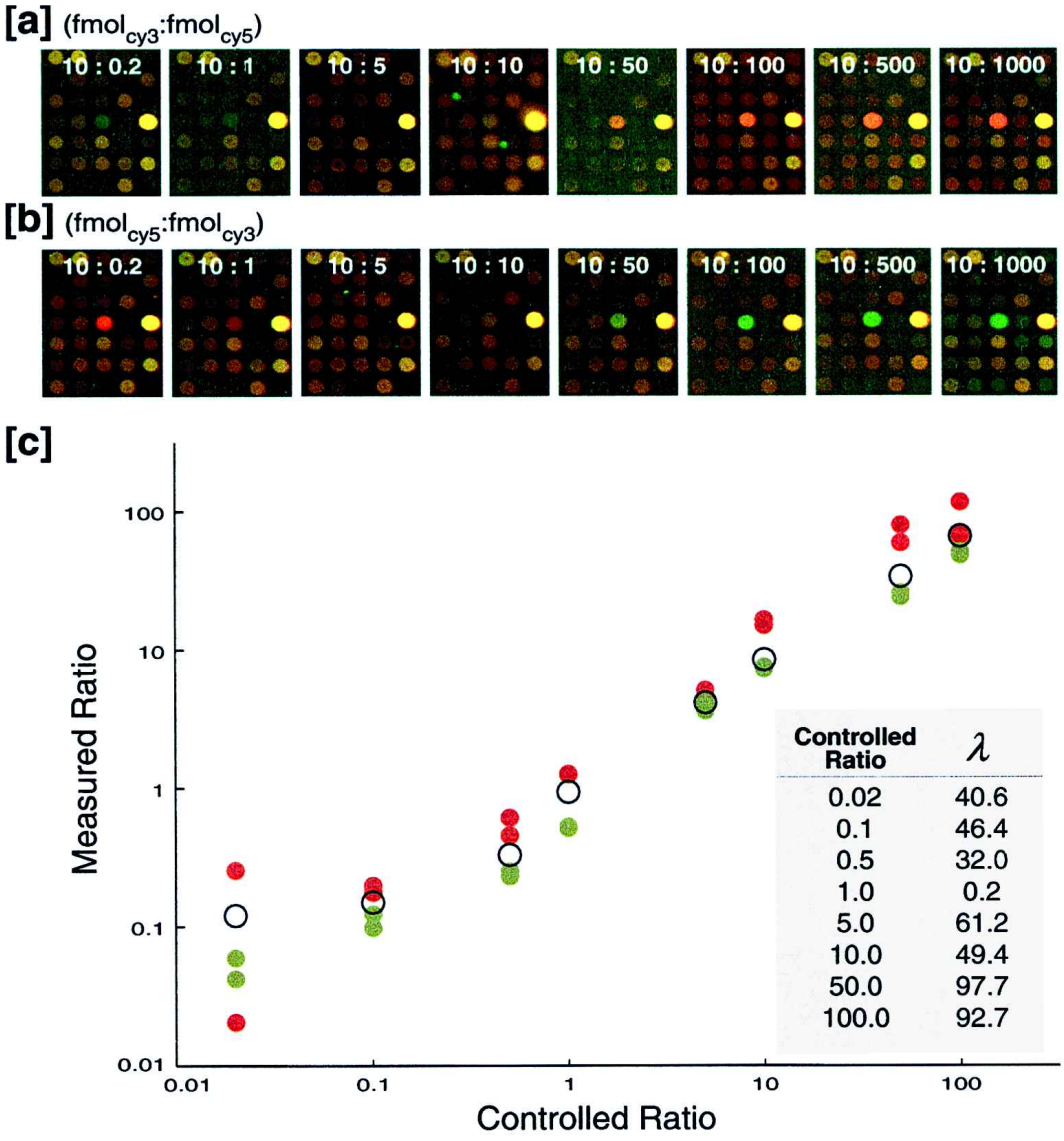

FIG. 3. (Caption on facing page.)

As shown in the table, the multiplicative errors are smaller (σ_{ε_x} and σ_{ε_y}) and more correlated (ρ_ε) within a slide than between slides. In addition, the within-slide measurements display less variability with regard to δ_x and δ_y, the additive error components. These findings are consistent with the expectation that several sources of error are unique to the between-slide measurements. For instance, although errors due to hybridization and quantitation occur in both types of experiment, mRNA extraction and labeling are performed prior to each slide hybridization and thus contribute to between-slide variability only. Because mRNA extraction and labeling are also performed independently for each condition (x and y), they may be responsible for the decrease in error correlation observed in the between-slide experiments.

We found that for these optimizations the parameter ρ_δ did not always converge: it was therefore set to zero during parameter estimation and does not appear in Table 2. We observed that in comparison with other data sets, the prenormalized x' and y' intensities of all 96 genes in the test data were moderate to relatively high (data not shown). We therefore postulate that ρ_δ was ill-determined because under the error model, ρ_δ is dominated by ρ_ε for larger intensities.

DISCUSSION

We have presented a mathematical model of the variability observed over repeated observations of intensities for genes represented on a DNA microarray. The model is motivated by empirical observations that x and y variances and x–y correlation increase with increasing values of x and y. Under this model, we have implemented a likelihood ratio test to identify genes whose true intensities μ_x and μ_y are unequal. Since we also provide evidence that ratios μ_y/μ_x are approximately proportional to ratios of molecular copy number, genes with unequal μ_x and μ_y are hypothesized to have different copy numbers of corresponding mRNA in the two cell populations under comparison, *i.e.*, are *differentially expressed.* This interpretation is tentative: we have not determined copy number per cell directly (only copy number in the labeled cDNA), we have fixed this quantity for the *GAL80* gene only, and our results may not apply to array experiments involving more complex RNA populations found in humans or other higher eukaryotes. However, our analysis does suggest that the yeast microarray is linear over a workable range and is sensitive enough to measure <2 copies/cell of message.

Comparison with ratio-based significance tests

Our maximum-likelihood approach has several important advantages over the currently accepted method based on expression ratios. In the ratio-based method, the expression ratio $r_i = y_i/x_i$ is computed for each gene i, and those genes are selected for which $r_i > r_c$ or $r_i < 1/r_c$. Because the ratio-based statistic combines x and y measurements into the single ratio r, absolute intensity information is necessarily lost. When repeated samples per gene are available, common practice is to compute the ratio of averaged x and y intensities, again discarding useful information.

Some genes are treated very differently by the λ and r statistics. For example, the two genes detailed in Fig. 2c (red versus black data points) have similar average expression ratios ($r = 2.9$ vs. $r = 3.5$ respectively), but the red-colored gene was determined to be significant between the YPR and YPRG cell populations by our method ($\lambda = 37.4$) while the black-colored gene was not ($\lambda = 13.8$). This difference in λ arises mainly because the samples corresponding to the red-colored gene are higher in intensity than those of the black-colored gene. In Equation 5, we compute λ for each gene by optimizing the model parameters (μ_x, μ_y) with and without the constraint $\mu_x = \mu_y$ then compare the likelihood of the (x, y) samples under the constrained and unconstrained models. The four red-colored samples are in the tail of the probability distribution for the error model with the constraint imposed (represented by a pink ellipse in Fig. 2c), resulting in a reduced likelihood L and thus a relatively high significance value λ. In contrast, the black-colored samples are relatively well explained by the constrained error model distribution (gray ellipse), resulting in a lower value of λ. Note that if the statistic r were applied with the commonly used threshold $r_c = 3.0$, the *black* gene would be accepted as significant while the *red* gene would not.

More generally, although the sets of genes chosen by the ratio- *vs.* likelihood-based thresholds are similar, an appreciable number of genes are chosen exclusively by one method or the other (see Fig. 2b).

For example, use of the ratio threshold $r_c = 3.0$ results in 283 significant genes in the YPR *vs.* YPRG experiment as opposed to the 456 genes selected using $\lambda_c = 23.8$, with 241 significant by both methods, 42 unique to the r_c threshold, and 215 unique to λ_c. Alternatively, use of a ratio threshold $r_c = 2.5$ would result in roughly the same number of significant genes for either method (456), 31% of which are unique to one of the methods.

Required number of samples per gene

Although many laboratories do not currently perform repeat array experiments, as the experimental process becomes more refined, inexpensive, and automated we believe the generation of repeated measurements will soon become standard practice. The ability to obtain repeated measurements has also been limited by the supply of available RNA, but this too is becoming less of a problem due to improvements in tissue dissection, RNA extraction, and single-cell PCR.

How many repeats are enough? It is known that maximum-likelihood methods can exhibit small-sample bias, and in simulations we have indeed observed biased estimates for small sample sizes. However, due to the large number of genes involved in a typical experiment, we have demonstrated that a likelihood ratio test performed with only four samples per gene chooses differentially-expressed gene candidates that are in good agreement with other experimental evidence. With two samples per gene, the number of genes identified as differentially expressed is reduced by more than one half, and the galactose-pathway structural genes, known previously to be the most highly induced between the two cell populations compared, are no longer among the most significant. Thus, although parameter estimates may be obtained with any sample size of two or greater, increasing the sample size from two to four has an appreciable effect on the behavior of this test.

Future work

The general framework presented here may be extended in several important directions. First, although we have chosen λ_c based on control experiments in which two cell populations are grown in identical conditions, a subsequent experiment of this type suggests that this cut-off can vary by 15% (resulting in a 19% change in the number of significant genes chosen in the YPG versus YPRG comparison). In order to reduce this variability in the future, it will be worthwhile to explore alternative methods for deriving λ_c.

Second, the error model does not currently distinguish between repeated samples drawn from multiple spots on a single array versus repeated samples drawn from multiple hybridizations to different arrays. Since we have observed that multiple spots within an array show less variability and more dye-to-dye correlation than do multiple spots observed over several arrays, it is reasonable that an error model which distinguishes between these two types of sampling would result in a more sensitive and/or accurate likelihood ratio test. Experimental systems which involve more than one level of sampling are well studied and can be addressed under the framework of a nested design model (Dunn and Clark, 1987).

Third, a maximum likelihood framework could be used not only to identify differentially-expressed genes but to place a confidence interval on their true expression difference. Instead of testing the hypothesis that $\mu_x = \mu_y$, we compute for each gene the range $l < (\mu_x - \mu_y) < h$. A detailed exploration of this calculation is left to a future publication.

Fourth, as the entire microarray process becomes more automated in the near future, we anticipate that error models such as the one presented here will be of prime importance for quantifying, comparing, and ultimately reducing the error introduced by each stage of the array process. As an example, we compared model parameters for two different levels of repeats: replicate spots on one array versus a single spot observed over multiple array hybridizations. In future work, this analysis could be greatly expanded to quantify several different levels of variation, such as variation due to cell culture, RNA preparation, labeling, or hybridization.

Finally, our method may be extensible to a wide range of biological data involving comparisons between multiple measurements. For instance, many laboratories perform gene expression experiments using radioactively labeled cDNA hybridized to gene clones spotted on membranes, and the use of oligonucleotide arrays is widespread. Apart from array data, technologies for comparing levels of protein expression between two cell populations have recently made dramatic improvements (Gygi *et al.*, 1999). The observed

quantities per gene are analogous to dye intensities observed in a microarray experiment and appear to be highly correlated. It appears likely that our error model is appropriate for describing measurements made using each of these technologies.

ACKNOWLEDGMENTS

We are indebted to Roger Bumgarner for his help in printing DNA microarrays and wish to thank Joe Felsenstein for inspiring discussions. T.E.I. is supported by an NIH Genome Training Grant, while V.T. is supported by a Sloan Foundation/DOE Fellowship in Computational Molecular Biology. A.F.S. holds the Grant I. Butterbaugh Professorship at the University of Washington.

REFERENCES

Ausubel, F., Brent, R., Kingston, R., Moore, D., Seidman, J., Smith, J., and Struhl, K. 1995. *Current Protocols in Molecular Biology*, vol 1, John Wiley, New York.

Buhler, J., Ideker, T., and Haynor, D. 2000. Improved techniques for finding spots on DNA microarrays. U.Washington Tech. Rep. 2000-08-05, www.cs.washington.edu/homes/jbuhler

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

Coleman, T., Branch, M.A., and Grace, A. 1999. *Matlab Optimization Toolbox User's Guide*. Third ed. Math Works, Natick.

DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A., and Trent, J.M. 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* 14, 457–460.

DeRisi, J., van den Hazel, B., Marc, P., Balzi, E., Brown, P., Jacq, C., and Goffeau, A. 2000. Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants. *FEBS Lett.* 470, 156–160.

DeRisi, J.L., Iyer, V.R., and Brown, P.O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686.

Dunn, O.J., and Clark, V.A. 1987. *Applied Statistics: Analysis of Variance and Regression*, Second ed. John Wiley, New York.

Greller, L.D., and Tobin, F.L. 1999. Detecting selective expression of genes and proteins. *Genome Res.* 9, 282–296.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. 1999. Correlation between protein and mRNA abundance in yeast. *Mol. Cell Biol.* 19, 1720–1730.

Hilsenbeck, S.G., Friedrichs, W.E., Schiff, R., O'Connell, P., Hansen, R.K., Osborne, C.K., and Fuqua, S.A. 1999. Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *J. Nat. Cancer Inst.* 91, 453–459.

Iyer, V., and Struhl, K. 1996. Absolute mRNA levels and transcriptional initiation rates in *Saccharomyces cerevisiae*. *Proc. Nat. Acad. Sci. USA* 93, 5208–5212.

Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson, J., Jr., Boguski, M.S., *et al.* 1999. The transcriptional program in the response of human fibroblasts to serum. *Science* 283, 83–87.

Kendall, M., and Stuart, A. 1979. *The Advanced Theory of Statistics*, vol 2, Fourth ed. Macmillan Publishing, New York.

Lander, E.S. 1999. Array of hope. *Nat. Genet.* 21, 3–4.

Lim, J.S. 1990. *Two-Dimensional Signal and Image Processing*, Prentice Hall, Englewood Cliffs.

Lohr, D., Venkov, P., and Zlatanova, J. 1995. Transcriptional regulation in the yeast GAL gene family: A complex genetic network, *Faseb Journal* 9, 777–787.

Ly, D.H., Lockhart, D.J., Lerner, R.A., and Schultz, P.G. 2000. Mitotic misregulation and human aging. *Science* 287, 2486–2492.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. 1992. *Numerical Recipes in C: The Art of Scientific Computing*, second ed. Cambridge University Press, Cambridge.

Roberts, C.J., Nelson, B., Marton, M.J., Stoughton, R., Meyer, M.R., Bennett, H.A., He, Y.D., Dai, H., Walker, W.L., Hughes, T.R., *et al.* 2000. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287, 873–880.

Wang, K., Gan, L., Jeffrey, E., Gayle, M., Gown, A.M., Skelly, M., Nelson, P.S., Ng, W., Schummer, M., Hood, L., and Mulligan, J. 1999. Monitoring gene expression profile changes in ovarian carcinoma using cDNA microarray. *Gene* 229, 101–108.

Address correspondence to:
Trey Ideker
The Institute for Systems Biology
4225 Roosevelt Way NE
Suite 200
Seattle, WA 98105-6099

E-mail: tideker@systemsbiology.org