JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 8, Number 1, 2001
Mary Ann Liebert, Inc.
Pp. 37–52

On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data

M.A. NEWTON,[1,2] C.M. KENDZIORSKI,[2] C.S. RICHMOND,[3] F.R. BLATTNER,[3] and K.W. TSUI[1]

ABSTRACT

We consider the problem of inferring fold changes in gene expression from cDNA microarray data. Standard procedures focus on the ratio of measured fluorescent intensities at each spot on the microarray, but to do so is to ignore the fact that the variation of such ratios is not constant. Estimates of gene expression changes are derived within a simple hierarchical model that accounts for measurement error and fluctuations in absolute gene expression levels. Significant gene expression changes are identified by deriving the posterior odds of change within a similar model. The methods are tested via simulation and are applied to a panel of *Escherichia coli* microarrays.

Key words: empirical Bayesian analysis, global gene expression, hierarchical modeling.

1. INTRODUCTION

TECHNOLOGY IS NOW BECOMING WIDESPREAD for measuring the simultaneous expression levels of thousands to tens of thousands of genes in a given cell type. There is mounting evidence that such data can yield significant insights into the underlying biology of the cell (e.g., Brown and Botstein, 1999; Lander, 1999). Coordinated expression patterns provide clues about gene function and shed light on complex biomolecular pathways; transcriptional profiles can characterize different cell types, thus potentially enabling improved cancer diagnosis and therapy, for example.

All high-throughput methods interrogate the population of mRNAs transcribed during gene expression in sampled cells, and they basically attempt to measure the abundance of each unique transcript. The methods rely on the highly specific process of hybridization to separate the complex pool of mRNA molecules. On a complementary DNA (cDNA) microarray, unique cDNA molecules are localized on a glass slide to act as probes against two different transcript samples. The two mRNA samples are prepared, separately labeled with distinct fluorescent dyes, and then cohybridized to the microarray. Fluorescence signal intensity in both channels is captured with a confocal microscope, and after some image analysis to localize each probe, expression levels are derived for each spot on the microarray within each original

[1]Department of Statistics, University of Wisconsin, Madison, WI 53792.
[2]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI 53792.
[3]Laboratory of Genetics, University of Wisconsin, Madison, WI 53792.

sample. Following convention, we use red to indicate the sample tagged with the Cy5 dye and green to indicate the Cy3 dye. Duggan *et al.* (1999) or Cheung *et al.* (1999), for example, provide further details on how to obtain cDNA microarray data. Similar expression data are obtained on oligonucleotide arrays (Lipshutz *et al.*, 1999), though we focus on the cDNA microarrays in the present development. There is reason to expect that the statistical methodology described here will apply in both domains.

To account for intrinsic differences between the hybridizing samples, intensity measurements are normalized in some fashion. One way is to compare signals at a set of house-keeping genes, i.e., genes thought to not present significant changes in expression between samples. Richmond *et al.* (1999) used a simple method on the *E. coli* microarrays reconsidered here; they normalized by the total signal intensity from all spots. Other possibilities include spiking the prepared sample with known concentrations of specific genes or combining measurements taken in both orientations of the dyes.

A component of each intensity measurement at a given spot is background fluorescence. An estimate of this component can be obtained from pixels near the spot, and the reported intensity is then the original measurement minus the background measurement. In what follows we consider the measurements to be normalized and to be adjusted for background intensity.

It is typical that inference about differential gene expression between two cell types is based on the ratio of measured expression levels. A basic statistical problem is to know when the measured differential expression is likely to reflect a real biological shift in gene expression. This depends on the amount of variation in the system, so it is difficult to justify a fixed rule, such as to focus on genes exhibiting more than a 3-fold shift, say. Certainly replication will be critical in applications; as more and more microarrays are measured, some confidence in the expression profiles will undoubtedly emerge. Our immediate concern is with data from a single microarray; we see some room for improvement in the initial signal processing which may have bearing on downstream tasks such as clustering or other forms of data analysis (e.g., Eisen *et al.*, 1998; Bassett *et al.*, 1999).

A given fold change in measured expression may have a different interpretation for a gene whose absolute expression is low as compared to a gene that is bright in both fluorescent channels (noted in Bassett *et al.* [1999] for example). We argue that any procedure which uses the raw intensity ratios alone to infer differential expression may be inefficient and thus may lead to excessive errors. Indeed, sources of variation are expected to be such that the absolute expression levels provide information on the variation of intensity ratios. This information is ignored in the standard treatment.

One solution is to ignore any genes whose transcripts are present at a low total abundance. We may have confidence about the differential expression of remaining genes, but at the price of throwing away potentially valuable data. In any case, the choice of a cutoff may be arbitrary. A gene may be deemed below the detection level in one channel but not in the other. Furthermore, the transcript abundance of many interesting genes may be very low, and so the strategy seems far from optimal.

The solution we describe is based on hierarchical models of measured expression levels in which we account for two obvious sources of variation. The first we call measurement error. In hypothetical repetitions of the experiment, the measured fluorescence signal will fluctuate around some mean value which is itself a property of the cell type, the particular gene, and other factors. These fluctuations are due to multiple sources of variation that arise in producing the measurement, such as variation in the preparation of the mRNA samples and in the incorporation of fluorescent tags, optical noise, and cross hybridization. Importantly, this variation may include, but is above and beyond, the background noise mentioned above. The second main source of variation we consider is due to the different genes spotted onto the microarray. The mean fluorescence value around which measured expression fluctuates changes from gene to gene and will serve as a random effect in our proposed model. The population of mRNAs from a given sample is composed of many distinct molecules, but the partition is not uniform; some mRNAs are abundant and others are rare.

As we see in Sections 3 and 4, by formally combining the two sources of variation we can readily obtain probabilistic statements about actual differential expression. We find that the observed ratios are not optimal estimators; we find that focusing on fold changes alone is insufficient and that confidence statements about differential expression depend on transcript abundance.

Perhaps the first statistical treatment of microarray data analysis is contained in Chen *et al.* (1997). These authors make the interesting argument that, although expected expression levels do fluctuate from gene to

gene across the microarray, the measurements are linked by having a constant coefficient of variation c, say. Then, the observed differential expression, say, $T_k = R_k/G_k$ (ratio of red to green intensity at gene k), has a sampling distribution dependent only on c under the null hypothesis that R_k and G_k have the same expectation, i.e., that there is no real differential expression, and computation is under a Gaussian model for both measurements. Using a set of house-keeping genes, which are thought to not present real differential expression, the maximum likelihood estimate of c is derived and "confidence intervals" for actual differential expression are computed from percentiles of the estimated null distribution of T_k. The intervals are easy to compute and are responsive to the intrinsic variation of data on the microarray because they use a data-dependent value of c. (By contrast, the procedure to call significant any gene presenting, say, a 3-fold or greater expression differential is not responsive to such variation.) The method has ignored ancillary information; $R_k \times G_k$ contains information about the variation of T_k. In other words, T_k is not independent of $R_k \times G_k$. There is the minor technical point, too, that R_k and G_k are modeled as Gaussian when in fact they must be positive. We consider a sampling model for measured intensities in Section 2.

One can view the Chen *et al*. (1997) method as producing a set of hypothesis tests, one for each gene on the microarray, in which the null hypothesis is that the expectation of both intensity signals is equal and the alternative is that they are unequal. When an observed T_k falls in the tails of the null sampling distribution, we reject the null and declare significant differential expression. As in other domains of application, we know that some benefit can be attained if the thousands of parameters are considered simultaneously, rather than in isolation (Efron and Morris, 1973, 1977; Carlin and Louis, 1996). The calculations presented here attempt to demonstrate the utility of treating the gene-specific parameters themselves as members of an array-specific population.

In Section 3 we consider the problem of estimating and possibly forming a confidence interval for the actual differential expression of a given gene. This involves calculations in a two-layer hierarchical model to produce a posterior probability distribution for the actual differential expression and an empirical Bayes estimate of the same. We illustrate with a panel of four *E. coli* microarrays, highlighting distinctions between the empirical Bayes estimates and the naive estimates. A simulation study shows that total estimation error can be reduced by using this procedure. The problem of testing for significant differential expression is the focus of Section 4, and here a third layer is added to the hierarchical model. We derive a function which gives the odds of actual differential expression as a function of measured intensities. This provides an effective summary, a sort of quality number, for each gene. To simplify our development, we focus on data from a single microarray, and we suppose that there is one spot for each gene. We address the issue of model validation in Section 5, where predictions from our hierarchical models are compared with features in the available data. Possible model extensions are also discussed.

2. A SAMPLING MODEL FOR MEASURED EXPRESSION

Measured intensity levels R (red) and G (green) approximate target mean values μ_r and μ_g at a given spot on the microarray. (We avoid using subscripts for distinguishing spots unless it is absolutely necessary.) The goal is to estimate $\rho = \mu_r/\mu_g$. The standard (naive) estimator is $\hat{\rho}_N = R/G$.

Measurement error depends on signal strength (Chen *et al*., 1997). To account for this explicitly, R and G are modeled as independent samples from distinct distributions with a common coefficient of variation. We find it convenient to work with Gamma distributions having a constant shape parameter. They exhibit constant coefficient of variation, they are Gaussian-like, they are supported on the positive line, and they are easy to manipulate. The Gamma model has a long history in statistical ecology (e.g., Fisher *et al*., 1943), not only for analytical convenience but also because it may possess a deeper biological interpretation. Dennis and Patil (1984) showed that the Gamma is the approximate stationary distribution for the abundance of a population fluctuating around a stable equilibrium. In Section 6 we comment on the sensitivity of our results to the Gamma model assumption. The probability density of a Gamma variable R with scale θ_r and shape parameter a is

$$p(r|\theta_r, a) = \frac{\theta_r^a r^{a-1} \exp(-r\theta_r)}{\Gamma(a)} \quad \text{for } r > 0 \tag{1}$$

where $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}\,dx$. We denote this density by Gamma(a, θ_r). Similarly, we model the measured intensity G as Gamma(a, θ_g), and we assume that R and G are independent. The expectations of R and G are a/θ_r and a/θ_g, respectively, and thus the target differential expression $\rho = \mu_r/\mu_g = \theta_g/\theta_r$. Both R and G have the same coefficient of variation $c = 1/\sqrt{a}$ even though they may have different scales.

By integrating the joint distribution of R and G, we can derive the sampling distribution of the measured differential expression $T = R/G$,

$$p(t|\theta_r, \theta_g, a) = \frac{\Gamma(2a)}{\Gamma^2(a)}\left(\frac{1}{\rho}\right)\frac{(t/\rho)^{a-1}}{(1+t/\rho)^{2a}} \tag{2}$$

for $t > 0$ where again $\rho = \mu_r/\mu_g$ is the parameter of interest. The tail of this distribution is asymptotic to $1/t^{a+1}$, so we restrict $a > 1$ to ensure a first moment. The form of this sampling distribution is well known; it is a scale multiple of a Beta distribution of the second kind (Kendall and Stuart, 1969, p. 151). At this point we could follow the development in Chen *et al.* (1997) using this Gamma model instead of the normal model used in that work. For instance, we see that when $\rho = 1$, i.e., no real differential expression, the distribution of T depends on the coefficient of variation $1/\sqrt{a}$ only. A problem with this approach is that we lose information when using the ratio T alone to assess differential expression. To see why, consider the conditional sampling distribution of $T = R/G$ given $S = RG$ within the Gamma model. The case of no actual differential expression, $\rho = 1$, is most simple. Denoting by θ the common value of θ_r and θ_g, we have

$$p(t|s, \theta, a) \propto \frac{1}{t}\exp\{-\theta\sqrt{s}(\sqrt{t} + 1/\sqrt{t})\}. \tag{3}$$

On the multiplicative scale of intensity measurements, S acts like total abundance from both channels. Inspection shows that the scale of this conditional distribution depends on S; i.e., the variation in T is smaller for larger S. The variability of differential expression is not constant, and so ignoring these changes can lead to inefficient statistical procedures. One way to modify the procedure from Chen *et al.* (1997), for example, would be to use something like (3) as a reference distribution instead of the analog to (1). We instead take a hierarchical modeling approach which enables direct parameter estimation and hypothesis testing.

3. ESTIMATING DIFFERENTIAL EXPRESSION

Except for test microarrays or house-keeping genes, we certainly expect real differences in gene expression between cell types; and clearly different genes can exhibit differences in actual expression within a given cell type. A key distinction of the present approach from earlier efforts is the formulation of a specific probability model to characterize these fluctuations. Among the range of possible specifications, we first consider a simple Gamma model for the scale parameters θ_r and θ_g. This form is conjugate to the Gamma sampling model and thus permits a detailed analysis. It entails independence among all the scale parameters on the microarray and assumes that they follow the common Gamma distribution Gamma(a_0, ν). Model fit can be improved slightly if we allow different scale parameters, say, ν_g and ν_r for the two dyes, but we take a common parameter in the present development. This model is reasonably flexible, skewed right, and presents increasing variation with increasing mean. It represents prior uncertainty in actual expression levels. An extension which allows correlation is described in Section 5.

Our main reasons for choosing a Gamma distribution to govern the latent scale parameters θ_r and θ_g are analytical tractability and model flexibility. We note in passing, however, that some theoretical justification may also exist. The target expression levels $\mu_r \propto 1/\theta_r$ and $\mu_g \propto 1/\theta_g$ each represent some kind of true abundance of the given transcript in the two mRNA pools. As such, their distribution concerns relative frequencies of frequencies and the size–frequency relation characteristic of the Zipf-Pareto law may obtain (Johnson *et al.*, 1994). If so, the relative frequency of genes with transcript abundance μ is proportional to $1/\mu^{a_0+1}$ for some power a_0. The reciprocal Gamma distribution has essentially the same density for moderate to large values μ.

With the two model components in place, we can derive some interesting consequences. Notably, we can compute the posterior distribution of the true differential expression at a given spot

$$p(\rho|R, G, \eta) \propto \rho^{-(a+a_0+1)} \left\{ \frac{1}{\rho} + \frac{(G+\nu)}{(R+\nu)} \right\}^{-2(a+a_0)} \tag{4}$$

where $\eta = (a, a_0, \nu)$ denotes the additional parameters yet to be specified. This is the distribution of the ratio of two independent Gamma variables, and it can be derived in the same way as the sampling distribution (2). Uncertainty about the true differential expression at a given spot is characterized by this distribution, and so, depending on our loss function, the Bayes estimate of ρ is some measure of its center. We note the mode and mean of this right-skewed distribution are

$$\text{mode} = \left(\frac{R+\nu}{G+\nu}\right)\left(\frac{a+a_0-1}{a+a_0+1}\right) \quad \text{and} \quad \text{mean} = \left(\frac{R+\nu}{G+\nu}\right)\left(\frac{a+a_0}{a+a_0-1}\right).$$

As it is in between these two values, and is somewhat simpler, we take as the Bayes estimate of differential expression

$$\hat{\rho}_B = \frac{R+\nu}{G+\nu}. \tag{5}$$

This has the classic form of a shrinkage estimator. For strong signals, $\hat{\rho}_B$ will be quite close to the naive estimator $\hat{\rho}_N = R/G$, but there is attenuation of the Bayes estimate, especially when the overall signal intensity is low. Thus, the Bayes estimate naturally accounts for the decreased variation in differential expression with increasing signal.

The amount of attenuation is governed by the parameter ν, which is yet to be specified. Being the scale parameter of the expression model component, we can represent it in more familiar terms. Consider the marginal expectation of signal intensity in the red channel, say, obtained by integrating uncertainty in θ_r,

$$E(R) = E[E(R|\theta_r)] = E(a/\theta_r) = a\nu/(a_0 - 1)$$

and so $\nu = (a_0 - 1)E(R)/a$. This simple formulation contains three key quantities. The coefficient of variation in the measurement error is controlled by a; the coefficient of variation describing fluctuations of actual expression among genes is controlled by a_0; and $E(R)$ is the overall average intensity measurement in the red channel. (We are assuming the data are normalized, and our model asserts that R and G have identical marginal distributions, so we could use G above instead of R.)

A pragmatic approach to dealing with the unknown parameters $\eta = (a, a_0, \nu)$ is to estimate them by marginal maximum likelihood. That is, by integrating uncertainty in both θ_r and θ_g we obtain the predictive probability density of each measurement pair (R, G) as

$$p_A(r, g) = \left\{ \frac{\Gamma(a+a_0)}{\Gamma(a)\Gamma(a_0)} \right\}^2 \frac{(\nu)^{2a_0}(rg)^{a-1}}{[(r+\nu)(g+\nu)]^{a+a_0}}. \tag{6}$$

Due to the independence assumption, this joint distribution is the product of the marginal distribution of R and the marginal distribution of G. Each margin is a scale mixture of Gamma distributions and hence is a compound Gamma distribution belonging to Type VI of Pearson's system (Johnson *et al.*, 1994, p. 381). The marginal loglikelihood $l(a, a_0, \nu)$ is the sum of contributions from all spots on the microarray

$$l(a, a_0, \nu) = \sum_k \log p_A(r_k, g_k)$$

where (r_k, g_k) are observed at the kth spot. We call $l(a, a_0, \nu)$ a marginal loglikelihood rather than an ordinary loglikelihood because the gene-specific θ parameters have been integrated away. We optimize

Table 1. Parameter Estimates in Gamma-Gamma Model[a]

Microarray	*NA*	a	a_0	ν
Control	37	2.06	2.10	12.84
Heat shock	82	2.12	1.61	7.13
IPTG-a	207	1.48	1.85	15.29
IPTG-b	149	1.19	1.57	14.71

[a]NA is the number of spots, out of 4,290, in which background is higher than spot intensity, and the remaining columns give maximum likelihood parameter values. The scale ν is for normalization to fractions then times 10^5.

$l(a, a_0, \nu)$ numerically using the **Splus** function **nlminb** (Statistical Sciences, 1993). In the examples worked so far, we find that a is estimated to be larger than $a_0 - 1$, so the estimate of ν is smaller than the mean intensity. The inference that results by estimating parameters as above is called empirical Bayes (EB) (Efron and Morris, 1973).

Efforts at whole genome expression analysis were pioneered in *E. coli* K-12 (Chuang *et al.*, 1993). Sequencing of the *E. coli* K-12 genome has enabled the fabrication of high resolution microarrays containing the entire complement of 4,290 open reading frames in this genome (Blattner *et al.*, 1997; Richmond *et al.*, 1999). To demonstrate our statistical methodology, we reanalyzed a panel of four microarrays from this *E. coli* project. One of these is a control microarray, two are replicates involving treatment with isopropyl-β-D-thiogalactopyranoside (IPTG) in which it is known that only a few transcripts should be induced, and one is from a heat shock treatment (HS) in which more global changes are expected. For the control microarray, total RNA was isolated from *E. coli* grown in a rich medium. This single pool was split in two, separately labeled, and then mixed and cohybridized to the microarray. In the IPTG replicates, control RNA (labeled Cy3) was cohybridized with RNA (labeled Cy5) from *E. coli* treated with IPTG. Similarly obtained was the heat shock microarray. Richmond *et al.* (1999) provide details. Following these authors, we invoke a simple normalization method; on each microarray, we first subtract background intensity from each spot. Then we divide each adjusted intensity by the total intensity obtained by combining all positive adjusted measurements. We omit from the estimation process any spots where the background is higher than the signal (column 1, Table 1). Maximum likelihood parameter estimates for the hierarchical Gamma–Gamma model are also given in Table 1.

The EB estimates of differential expression attenuate the naive estimates, as summarized graphically in Fig. 1. Points indicate the normalized, background-adjusted intensities in the two dyes, say, (G, R). Line segments run from each (G, R) to $(G + \nu, R + \nu)$. Owing to the logarithmic scale, of course, this shrinkage is most pronounced for low intensity spots. It is interesting that on the control microarray the shrinkage constant is fairly large in comparison to the heat-shock microarray, suggesting that the method can distinguish noise from significant signal.

The attenuation inherent in the EB estimates affects the ranking of the most highly differentially expressed genes. Figure 2 illustrates the ranking changes for two of the *E. coli* microarrays. We consider the top 100 most differentially expressed genes, as measured by $\hat{\rho}_N = R/G$. For each n in the range 1 to 100, we ask how many of these top n genes are ranked in the top n by the EB procedure. If the methods agree (i.e., if ν is very small), then the answer is n. About one quarter of the 100 most highly differentially expressed genes measured via $\hat{\rho}_N$ are not in the top 100 as measured by the EB procedure. Data analysis methods often focus on the most differentially expressed genes, and so it is quite possible that the use of the more efficient Bayes estimation procedure will have an impact on "downstream" computations using measured differential expression.

The EB estimate $(R+\nu)/(G+\nu)$ is attenuated compared to the naive estimator R/G, but is the estimate any better? We report the results of a small simulation to address this question. Green and red intensities were simulated for a synthetic microarray having 4,000 spots. Intensities arose from Gamma distributions with shape parameter $a = 2$ and Gamma distributed scales in which $a_0 = 2$ and $\nu = 8$. We incorporated

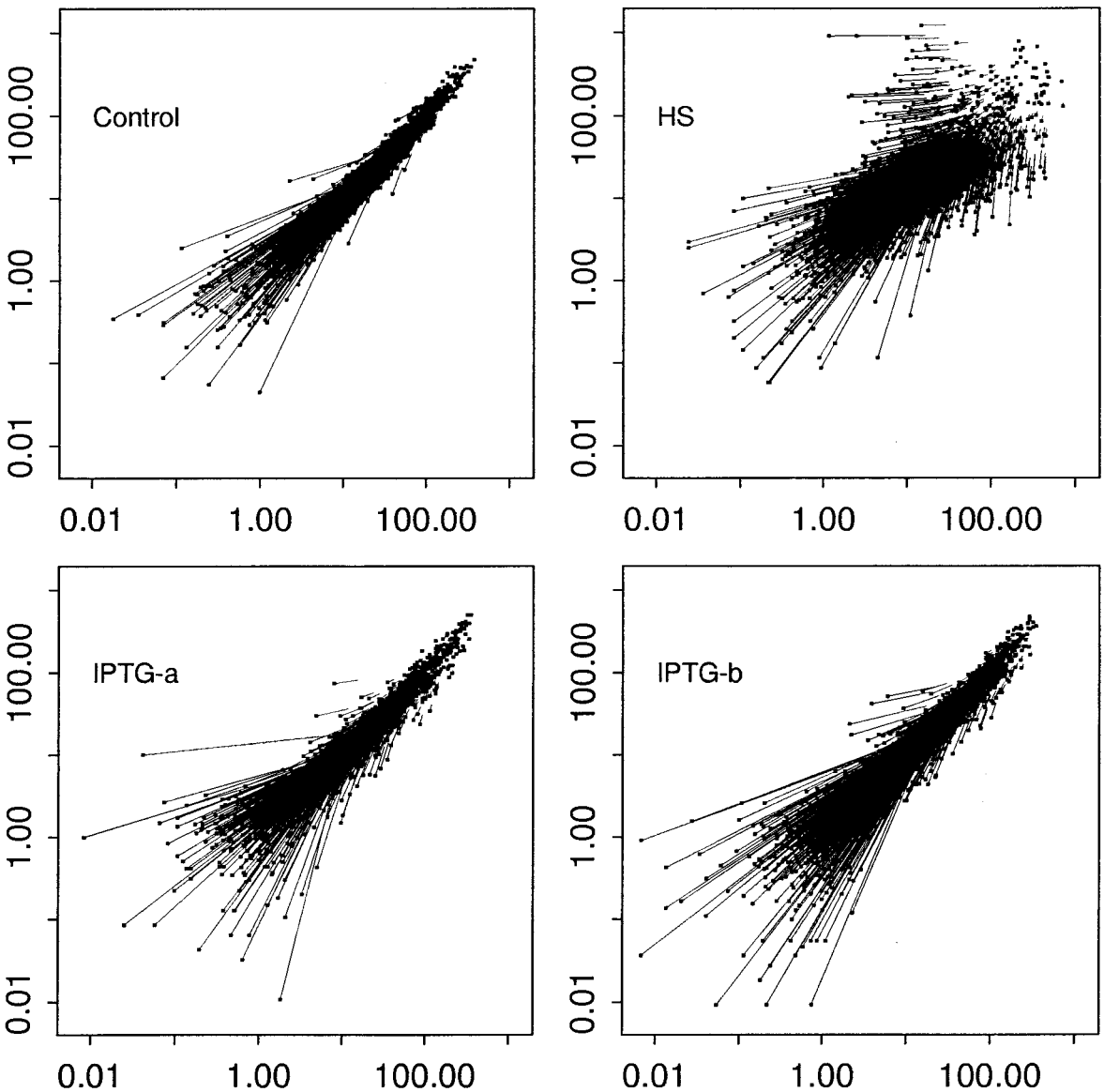

FIG. 1. Shrinkage estimation, *E. coli*: Points are plotted at measured intensities (G, R), and line segments extend to $(G + \nu, R + \nu)$. On lines parallel to the diagonal, fold change is constant.

some positive correlation between green and red scale parameters using the model described in Section 5 with $\kappa = 20$. Thus the simulation model differed from the model used for fitting. By keeping track of the simulated scales θ_g and θ_r, of course we can measure the error in estimating ρ for both the naive procedure and the EB procedure (Fig. 3). There is a fairly significant error reduction by the EB procedure in this case. Other cases not reported showed similar error reductions.

Beyond point estimates of differential expression, we can use (4) to obtain Bayesian confidence intervals (credible intervals). By a change of variables, $1/(1 + \rho/\hat{\rho}_B)$ is distributed a posteriori as a symmetric Beta distribution with shape parameter $a + a_0$, and so endpoints of a credible interval may be computed by back-transforming quantiles from this symmetric Beta. The credible interval provides a measure of uncertainty in ρ, but we find that inference beyond point estimation may be more accurate in a model, such as the one described next, in which it is recognized that there are no real changes in expression for some subset of genes.

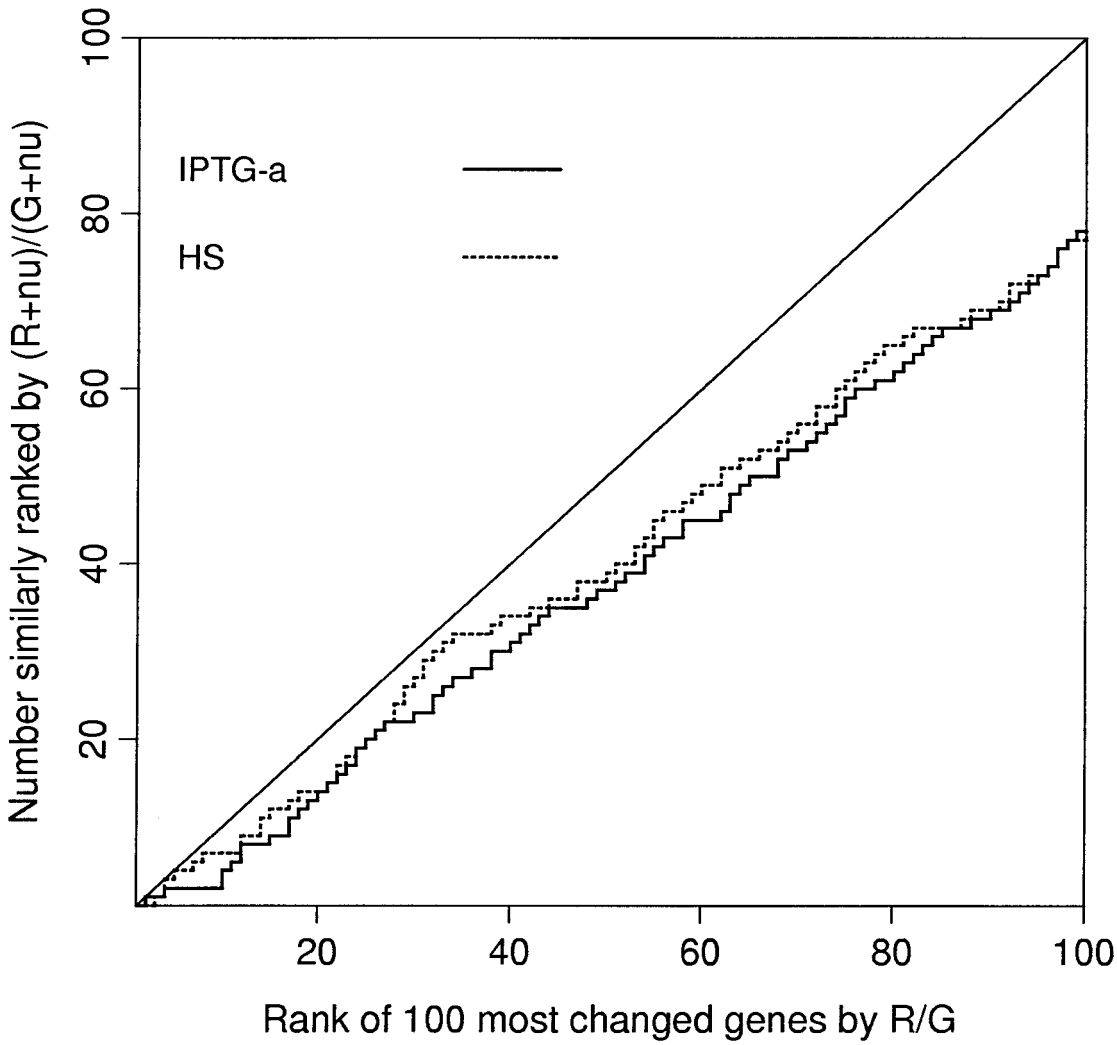

FIG. 2. Effect on ranking of genes.

4. IDENTIFYING SIGNIFICANT DIFFERENTIAL EXPRESSION

We turn attention to deciding whether or not the observed differences at a given gene are sufficiently large to assert significance. Having a third layer in the hierarchical model facilitates the calculations. The true mean intensities of some proportion p of spots change between conditions (i.e., $\mu_r \neq \mu_g$), while the others remain fixed ($\mu_r = \mu_g$). For spots which change, we use the previous model from Section 3. In other words, scale parameters θ_r and θ_g are independent Gamma variates with common shape a_0 and scale ν. For unchanged spots, the common scale parameter θ is deemed to arise from the same Gamma distribution.

A problem presented by the Gamma-Gamma-Bernoulli specification is that the identity of the changed spots is unknown. The likelihood calculations (which we would use to estimate a, a_0, ν, and p) appear impossibly complex, since they involve summation over all 2^n configurations of missing indicators, where n is the number of spots. Fortunately, the model fits well into the EM algorithm framework (Dempster *et al.*, 1977), and so we have a simple recursion to infer these parameters from the marginal likelihood.

The first ingredient in the calculation is the marginal probability of data at a spot if there is no real differential expression. This is obtained by integrating the Gamma model for the common scale θ. In

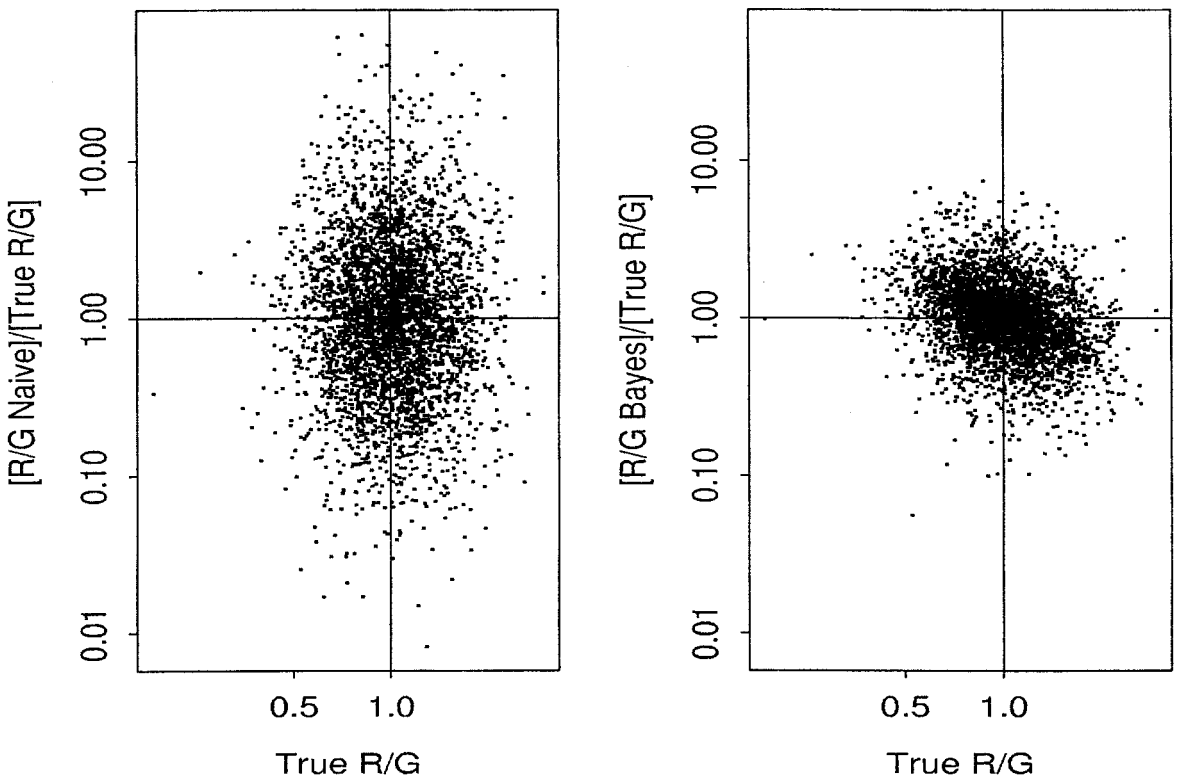

FIG. 3. Error reduction: average$|\log(\hat{\rho}_N/\rho)|$ = 0.88 whereas average$|\log(\hat{\rho}_B/\rho)|$ = 0.42. The empirical Bayes estimate gives a 50% reduction in error.

contrast to (6) in which there are different scales, the marginal probability in the null case is:

$$p_0(r, g) = \frac{\Gamma(2a + a_0)}{\Gamma^2(a)\Gamma(a_0)} \frac{\nu^{a_0}(rg)^{a-1}}{[(r + g + \nu)]^{2a+a_0}}. \tag{7}$$

Letting r_k and g_k denote the measured intensities at spot k and introducing the binary indicator variable z_k to be 0 unless there is true differential expression, the *complete* data loglikelihood is

$$l_c(a, a_0, \nu, p) = \sum_k \{z_k \log p_A(r_k, g_k) + (1 - z_k) \log p_0(r_k, g_k) + z_k \log(p) + (1 - z_k) \log(1 - p)\}.$$

The E-step is to obtain the conditional expectation of $l_c(a, a_0, \nu, p)$, which simply involves replacing z_k by the posterior probability of change

$$\hat{z}_k = P(z_k = 1|r_k, g_k, p) = \frac{pp_A(r_k, g_k)}{pp_A(r_k, g_k) + (1 - p)p_0(r_k, g_k)} \tag{8}$$

and with parameters a, a_0, ν, and p fixed at tentative values. The M-step is to maximize the resultant form in the four parameters. Having broken the mixture structure, this maximization is simplified. Immediately, we find the updated estimate of p is the arithmetic mean of $\{\hat{z}_k\}$. An off-the-shelf numerical procedure, such as **nlminb** in **Splus**, readily optimizes the remaining parameters in each iteration. Forty iterations of EM were used to obtain the estimates reported in Table 2, and results were checked from various starting configurations.

Placing a prior distribution over p stabilizes the computations and enables a nice interpretation of the output. We use a Beta(2,2) prior in what we report in Table 2, which amounts to a prior assumption of exchangeability of the $\{z_k\}$ and that $P(z_k = 1) = 0.5$ upon integrating uncertainty in p. It is convenient to

Table 2. Parameter Estimates in Gamma–Gamma–Bernoulli Model via EM Algorithm

Microarray	a	a_0	v	p
Control	22.90	0.94	0.28	0.003
Heat shock	2.75	1.37	4.12	0.052
IPTG-a	12.53	0.82	0.37	0.007
IPTG-b	9.69	0.68	0.28	0.004

fix other parameters, a, a_0, and v at their estimated values rather than integrating against a prior. Owing to the large sample size, there should not be significant error in the present examples.

Our goal is to compute posterior odds of change at each spot. The odds summarize our inference about actual differential expression at each spot using all the data on the microarray. With $D = \{r_k, g_k\}$ denoting expression measurements on the whole microarray, the posterior odds of change at spot k are:

$$\text{odds} = \frac{P(z_k = 1|D)}{P(z_k = 0|D)},$$

where

$$P(z_k = 1|D) = \int_0^1 P(z_k = 1|p, r_k, g_k) P(p|D)\, dp \tag{9}$$

by conditional independence of the data at different spots given the parameter p. The Bayes rule determines $P(z_k = 1|p, r_k, g_k)$ in terms of $p_0(r_k, g_k)$ and $p_A(r_k, g_k)$; see (8). Also, the EM-algorithm finds the posterior mode of $P(p|D)$, say, $\hat{p}$. To a first approximation, the integral (9) equals the integrand $P(z_k = 1|r_k, g_k, p)$ evaluated at its modal value $p = \hat{p}$. Therefore,

$$\text{odds} \approx \frac{p_A(r, g)}{p_0(r, g)} \frac{\hat{p}}{1 - \hat{p}}. \tag{10}$$

These posterior odds may also be called Bayes factors because the prior odds for change equal unity. An inspection of the profile loglikelihood curve for p indicated that $P(p|D)$ is highly concentrated for the *E. coli* examples, and so (10) provides a good approximation.

Figure 4 shows contours of the posterior odds of true differential expression computed in the Gamma-Gamma-Bernoulli model using (10). The contour map provides an interesting summary of significant change. The grey area in each panel indicates that the odds favor no change. An important feature of the map is that the contour lines are not straight on this log–log scale, indicating that to consider fold changes alone is not enough. As we expected, we are less confident about naive fold change at low signal intensity. Dotted lines on each panel correspond to the 99% rule from Chen *et al.* (1997). (We used all the spots to estimate the coefficient of variation, and this was so large in the heat-shock data that the Chen *et al.* approximation to the upper band broke down.) These lines are designed so that they will be exceeded only for 1% of spots which in fact have not changed. Within the present model, most of this 1% can be expected to occur at low total abundance, a sign of inefficiency.

The Bayes factor computation allows us to rank order genes by their probability of real differential expression. We have replicate microarrays for the IPTG treatment, so it is instructive to check the reproducibility of these assessments. As a technical note, we had removed in the estimation phase any spots for which the measured intensity was below the estimated background. For assessing changes, we include these spots and deem them to present zero signal in that channel. (Interestingly, the log Bayes factor is continuous at zero intensity so there is no computational problem in doing so.) On IPTG-a, 20 genes have odds better than 10:1 favoring changed expression. The replicate IPTG-b shows seven changed genes. Five valid genes are in common between the replicates and are induced by IPTG. These five are exactly the genes identified by Richmond *et al.* (1999) as being induced on the basis of the same microarray data and other radioactivity data.

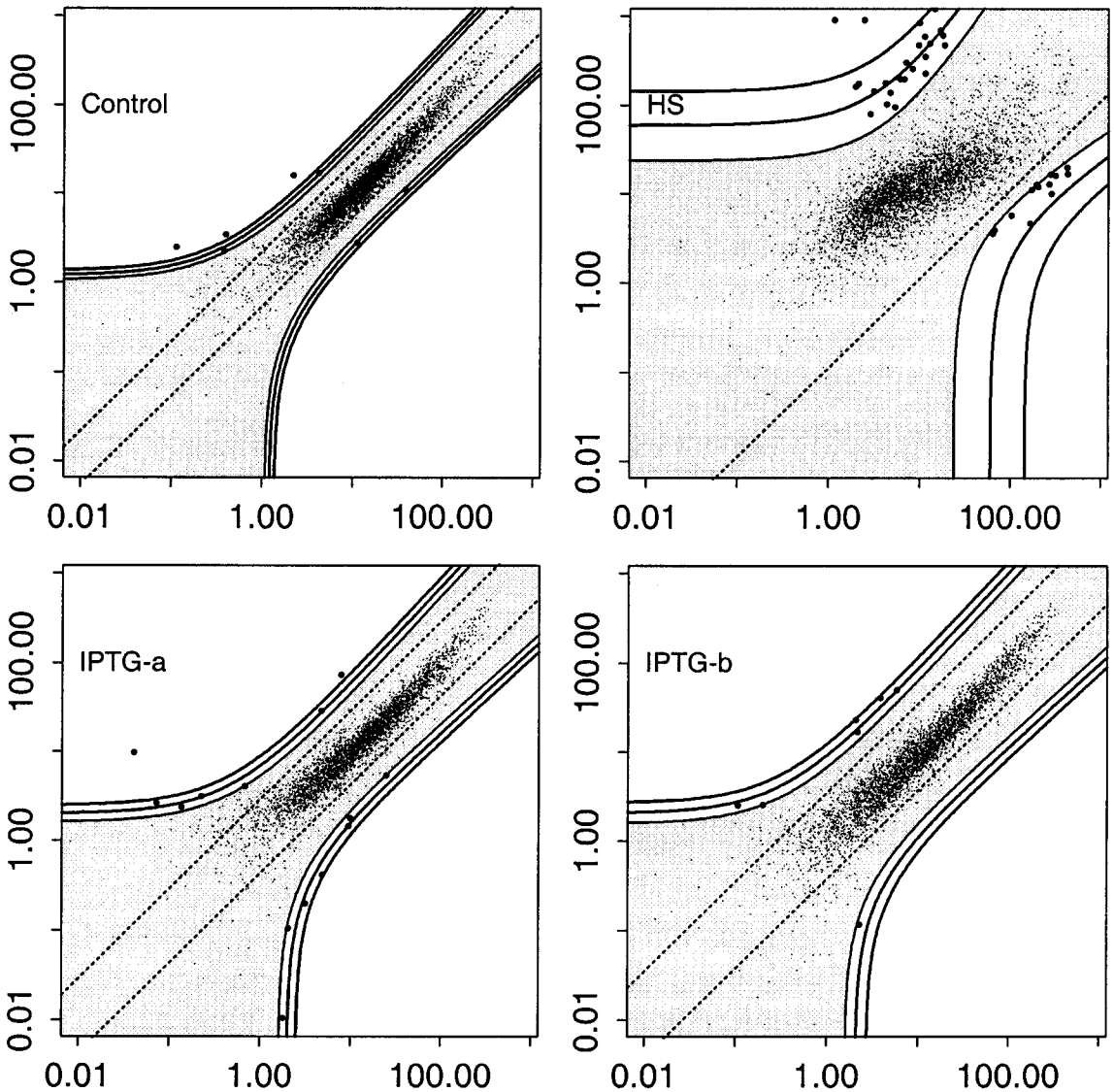

FIG. 4. Odds of real differential expression *E. coli* data: In shaded region odds are for no change. Contours are at odds of change of 1:1, 10:1, and 100:1, respectively.

The heat-shock data provide an interesting demonstration of the methodology. Using Bayes factors, we find 38 genes exhibit differential expression (25 induced and 13 repressed). This represents about 1% of the genes on the microarray, and prior work suggests that more genes have changed (Richmond *et al.*, 1999). If we look at the parameter estimates for this case (Table 2), we see that the estimated proportion of changed spots, $\hat{p}$, is about 5% (about 200 genes), and this number is in line with the earlier report. Being the optimal parameter value, $\hat{p}$ satisfies the equation $\hat{p} = [\sum_{k=1}^{n} \hat{z}_k(\hat{p})]/n$ where $\hat{z}_k(p)$ is the posterior probability of change at spot k, as in (8). At the same time, the Bayes factor (odds for change) is the ratio of $\hat{z}_k(\hat{p})$ to $1 - \hat{z}_k(\hat{p})$. Something interesting is going on. On average over spots, the posterior probability of change is about 5%, and this leads us to infer that about 5% of spots have changed. When it comes to deciding which spots have changed, however, only 38 spots, about 1%, have $\hat{z}_k(\hat{p}) > 0.5$ and thus have odds favoring change. Our inference about the proportion of changed spots does not need to be the same as the proportion of spots which we can confidently say have changed. In fact, by Markov's inequality, the proportion of spots in which $\hat{z}_k(\hat{p}) > 0.5$ is no greater than $2\hat{p}$, and some refinements to this bound may be possible.

TABLE 3. ERROR RATES, SIMULATED GAMMA–GAMMA–BERNOULLI MODEL[a]

	Odds-Bigger-Than-1 Rule		*Less Stringent Rule*	
	$\hat{z}_k > 0.5$	$\hat{z}_k \leq 0.5$	$\hat{z}_k > 0.236$	$\hat{z}_k \leq 0.236$
True change	577	423	694	306
No change	73	2927	311	2689
Total	650	3350	1005	3995

[a]Two decision rules are examined below for a 4,000 spot microarray in which 1,000 genes have changes in true gene expression. Significant change is inferred if the posterior probability of change, $\hat{z}_k$, exceeds a cutoff.

To study our methodology further, we performed a small simulation. We considered a single microarray with $n = 4000$ spots and in which data for 1,000 spots arose from the Gamma-Gamma model with $a = 12$, $a_0 = 1$, and $\nu = 1$. The remaining 3,000 spots had variable expression levels, but there was no change in true expression from green to red. The same Gamma model was used to generate these common expression levels and then also to generate the measured intensities. So basically we simulated the Gamma–Gamma–Bernoulli model, but we forced exactly 1,000 spots to change. Parameter estimates, obtained by the EM algorithm, were $\hat{a} = 12.5$, $\hat{a}_0 = 1.0$, $\hat{\nu} = 0.95$, and $\hat{p} = 0.26$, and thus we recovered the model parameters extremely well. Table 3 records error rates by two decision rules. Taking our standard rule, to call a spot as changed if the odds exceed unity, we inferred only 650 changed spots, much less than the 1,000 or so which we conclude have changed from $\hat{p}$. This underestimation mirrors what happened with the heat-shock data. In total, we make 496 incorrect calls using the odds-bigger-than-one rule. A second, much less stringent rule is to call a total of $n\hat{p}$ spots as changed, rank-ordered by their individual posterior probabilities. For these simulated data, we thus lowered the bar from a change probability of 0.5 to a change probability of 0.236, and by so doing we produced a list of about 1,000 genes. This rule has the advantage that our conclusion about the number of changed genes is in line with our reporting of particular genes. But, at least in this simulation, we amplified the overall error rate; with this rule, 617 spots were called incorrectly. The simulation highlights difficulties with inferring differential gene expression, but we point out that specific error rates may depend on the application.

5. MODEL VALIDATION AND DISCUSSION

Both the Gamma–Gamma model and the Gamma–Gamma–Bernoulli model attempt to capture some structural features expected in microarray data, but of course they are highly parameterized and it is important to check whether predictions implied by them are in line with available data. We have considered several simple checks. For instance, we can compare a histogram of measured intensities to the fitted marginal model (Fig. 5). Plotted on each histogram (on the log scale) is the fitted marginal density from the Gamma–Gamma model used for shrinkage estimation (dotted line) and the fitted density from the Gamma–Gamma-Bernoulli model (solid line). Clearly there is room for improvement in the fit, but the primary features of the data are captured.

A second interesting check is based on a well-known property of the Gamma distribution. If R and G are measurements on one spot, and there is no real differential expression, then both measurements arise from a common Gamma distribution with shape a and scale θ. The renormalized difference

$$B = \{1 + (R - G)/(R + G)\}/2$$

has a symmetric Beta distribution with shape parameters a and a; i.e., its density is proportional to $b^{a-1}(1 - b)^{a-1}$ for $b \in (0, 1)$. Notably, this density does not depend on the scale parameter, so it is the common distribution for *all* unchanged spots on the microarray. Figure 6 compares the histogram of B values to the fitted Beta density. For the treatment microarrays, we focused only on B values from spots deemed to be probably unchanged by the Bayes factor computation. This is a true model check; the fitting procedure does not attempt to capture variations on the shown scale. Indeed, the fit is poor in some respects, but again primary features of variation are captured.

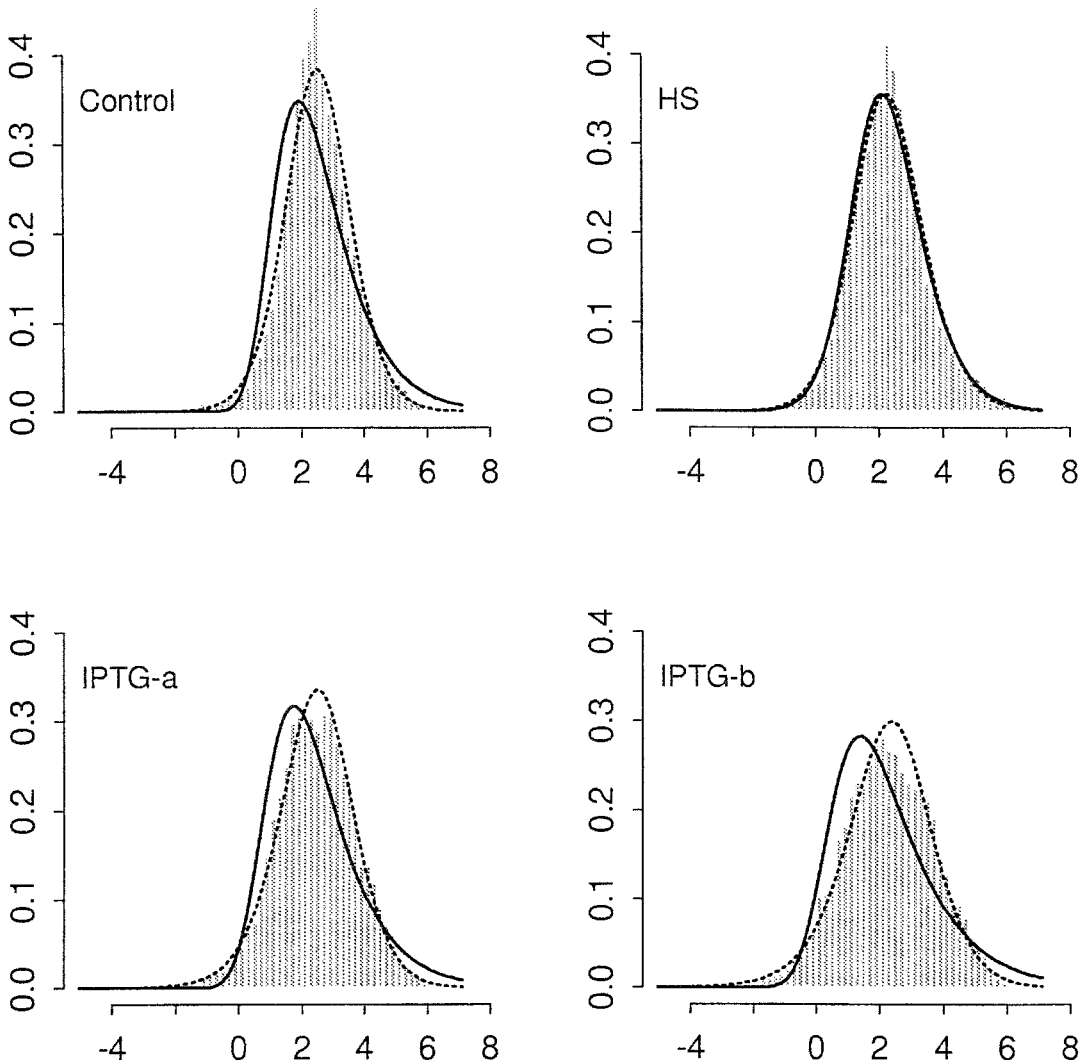

FIG. 5. Diagnostic check: Histograms are of intensities (both colors pooled) on the natural log scale. Dashed curve is fit from Gamma–Gamma model; solid curve is fit from Gamma–Gamma–Bernoulli model.

We are using just four parameters to describe marginal variation and dependence between the red and green channels, and improvements may come by increasing the number of parameters. We did let the scale parameter v be color specific, and this improved the fits somewhat, especially on the IPTG microarrays, as the estimated value of a goes up. We also let the shape parameter a be color specific, but this did not significantly improve fits. The use of additional scale parameters lead to problems with the Bayes factor contours in Fig. 4, and so currently we are investigating model elaborations and identifiability questions.

We conjecture that improvements may arise if some positive correlation is added to the expression model. We can do this by retaining the Gamma sampling model but adding correlation between θ_r and θ_g in the expression model. For example, we might say $\psi \sim \text{Gamma}(a_0, v)$ and two multipliers $m_r, m_g \sim_{iid} \text{Gamma}(\kappa, \kappa)$ for a positive dependence parameter κ. Then write $\theta_r = m_r\psi$ and $\theta_g = m_g\psi$. The multipliers are centered on unity, and will be close to unity if κ is large. This model is an intermediate between the null and alternative models so far studied, and it requires more sophisticated machinery to fit, but it may be effective at identifying subtle expression changes.

Our methods use Gamma distributions but other parametric forms can be considered. Various parametric models entail constant coefficient of variation on the positive line. The log-normal model has been used for this purpose, and a comparison of the different formulations will be useful (Wiens, 1999). On the

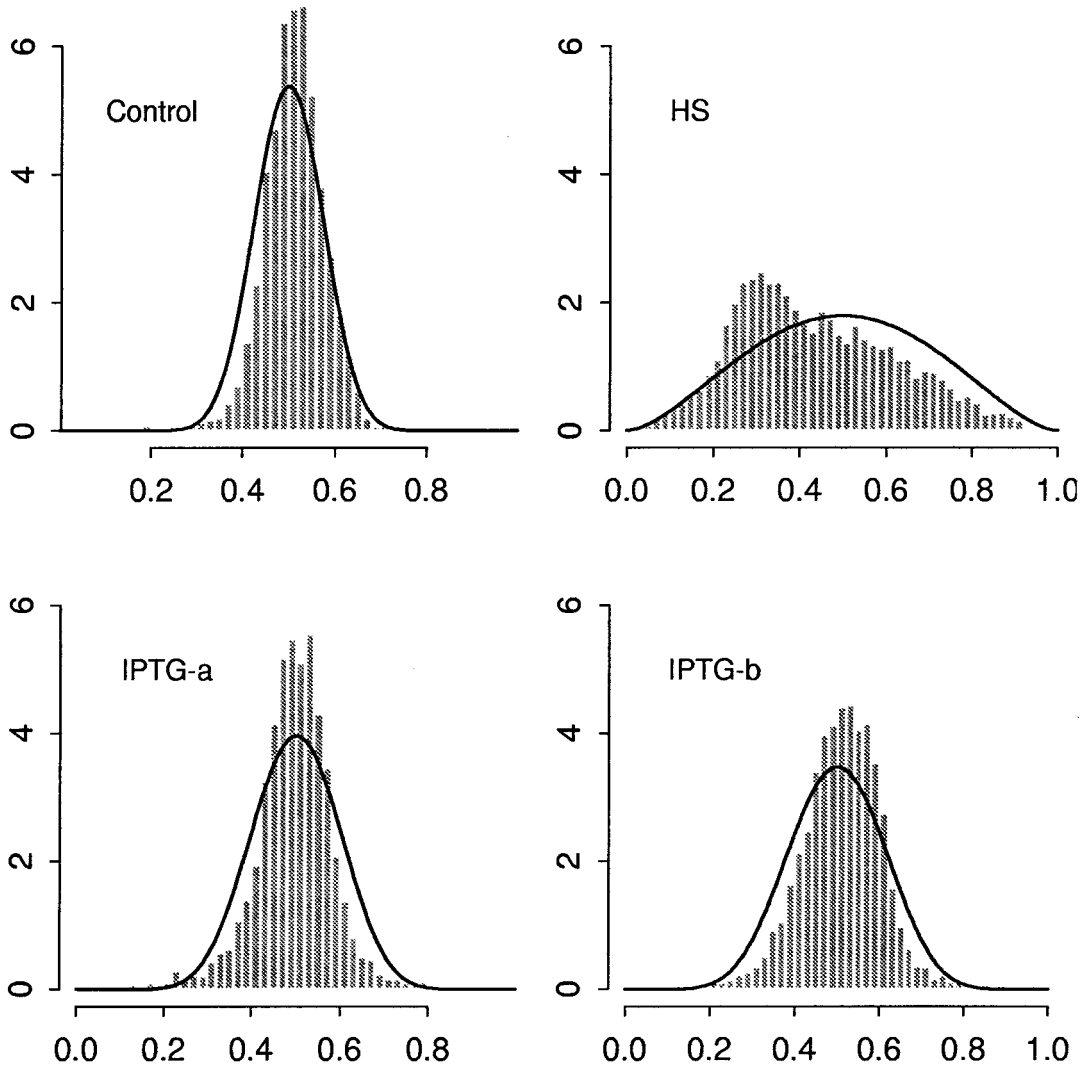

FIG. 6. Diagnostic check: Histograms are of renormalized differences $B = (1/2)[(R - G)/(R + G) + 1]$ for spots deemed to have not changed. Curves are predicted Beta densities.

basis of preliminary computations, we can say that the same qualitative analysis features carry over to the log-normal, in particular the shape of contours in the Bayes factor plots is similar.

We have used one particular method of normalization and background noise adjustment. Probably some advantage can be gained by combining these tasks with the present modeling method to better account for these sources of variation. For instance, on normalization, we could say that the scale parameters in one sample are a global constant multiple of those in the other sample, and then treat this constant as another model parameter to be estimated from unnormalized intensities. This is similar to the calibration procedure described in Chen *et al.*, (1997), but in the context of a hierarchical model.

Our methodology deals with a single microarray at a time, and does not attempt to combine data, though the modeling framework certainly allows this elaboration. One approach would be to decompose say $\log \theta$, the expression scale parameter, into contributions from different genes, different RNA preparations, and different growth conditions. Combining information from multiple microarrays may be an effective way to obtain accurate estimates of the contribution of different sources of variation. Kerr *et al.* (2000) provide details of a related method which expresses the expected value of log-transformed intensity measurements in terms of contributions from such factors.

Hierarchical statistical modeling allows for efficient data processing in large-scale expression studies. This provides more precise estimates of differential gene expression and more accurate assessments of significant changes than standard methods by accounting for differential variability in data. Calculations account for the measurement error process and for natural fluctuations in absolute expression levels. Preprocessing image data via these methods may reduce errors in downstream tasks, such as cluster analysis or classification.

ACKNOWLEDGMENTS

The authors thank Bob Mau for his critical review of an earlier draft and a referee for helpful comments. They also acknowledge interesting discussions with Gary Churchill who, independently, has considered the use of shrinkage estimation for differential expression, and they thank Alex Loguinov who identified a bug in one of the plotting programs. Code and data used in this article are freely available at the first author's web site *www.stat.wisc.edu/~newton/*. This research was funded in part by the National Cancer Institute, grant R29CA64364-01 to M.A.N. and training grant TA-CA 09565 for C.M.K. NIH grant R01 GM35682 supported C.S.R. and F.R.B.

REFERENCES

Bassett Jr., D.E., Eisen, M.B., and Boguski, M.S. 1999. Gene expression informatics—it's all in your mine. *Nature Genet. Suppl.* 21, 51–55.

Blattner, F.R., Plunkett, G.III., Bloch, C.A., Perna, N.T., Buland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, R., and Shao, Y. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277, 1453–1474.

Brown, P.O., and Botstein, D. 1999. Exploring the new world of the genome with DNA microarrays. *Nature Genet. Suppl*. 21, 33–37.

Carlin, B.P., and Louis, T.A. 1996. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, New York.

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

Cheung, V.G., Morley, M., Aguilar, F., Massimi, A., Kucherlapti, R., and Childs, G. 1999. Making and reading microarrays. *Nature Genet. Suppl*. 21, 15–19.

Chuang, S.-E., Daniels, D.L., Blattner, F.R. 1993. Global regulation of gene expression in *Escherichia coli*. *J. Bacteriol.* 175(7), 2026–2036.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Statistical Society, Series B* 39, 1–38.

Dennis, B., and Patil, G.P. 1984. The gamma distribution and weighted multimodal gamma distributions as models of population abundance. *Mathematical Biosciences*, 68, 187–212.

Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J.M. 1999. Expression profiling using cDNA microarrays. *Nature Genetics Supplement* 21, 10–14.

Efron, B., and Morris, C. 1973. Combining possibly related estimation problems (with discussion). *Journal of the Royal Statistical Society, Series B* 35, 379–421.

Efron, B., and Morris, C. 1977. Stein's paradox in statistics. *Sci. Am.* 236, 119–127.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95, 14863–14868.

Fisher, R.A., Corbet, A.S., and Williams, C.B. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Animal Ecology* 12, 42–58.

Johnson, N.L., Kotz, S., and Balakrishnan, N. 1994. *Continuous Univariate Distributions*, vol. 1, 2nd ed. Wiley, New York.

Kendall, M.G., and Stuart, A. 1969. *The Advanced Theory of Statistics*, vol. 1, 3rd ed. Hafner, New York.

Kerr, M.K., Martin, M., and Churchill, G.A. 2000. Analysis of variance for gene expression microarray data. Manuscript. *http://www.jax.org/research/churchill/*.

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

Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R., and Lockhart, D.J. 1999. High density synthetic oligonucleotide arrays. *Nature Genet. Suppl*. 21, 20–24.

Richmond, C.S., Glasner, J.D., Mau, R., Jin, H., and Blattner, F.R. 1999. Genome-wide expression profiling in *Escherichia coli* K-12. *Nucl. Acids Res.* 27(19), 3821–3835.

Statistical Sciences, 1993. *S-PLUS Guide to Statistical and Mathematical Analysis, Version 3.2*. StatSci, a division of MathSoft, Inc., Seattle.

Wiens, B.L. 1999. When log-normal and gamma models give different results: A case study. *Am. Statistician* 53, 89–93.

Address correspondence to:
Michael A. Newton
Department of Statistics
1210 W. Dayton Street
University of Wisconsin
Madison, WI 53706-1685

E-mail: newton@stat.wisc.edu