

# Bayesian Estimation of Transcript Levels Using a General Model of Array Measurement Noise

RON O. DROR,<sup>1,\*</sup> JONATHAN G. MURNICK,<sup>1,\*</sup> NICOLA J. RINALDI,<sup>2</sup>  
VOICHITA D. MARINESCU,<sup>3</sup> RYAN M. RIFKIN,<sup>4</sup>  
and RICHARD A. YOUNG<sup>2</sup>

## ABSTRACT

Gene arrays demonstrate a promising ability to characterize expression levels across the entire genome but suffer from significant levels of measurement noise. We present a rigorous new approach to estimate transcript levels and ratios from one or more gene array experiments, given a model of measurement noise and available prior information. The Bayesian estimation of array measurements (BEAM) technique provides a principled method to identify changes in expression level, combine repeated measurements, or deal with negative expression level measurements. BEAM is more flexible than existing techniques, because it does not assume a specific functional form for noise and prior models. Instead, it relies on computational techniques that apply to a broad range of models. We use Affymetrix yeast chip data to illustrate the process of developing accurate noise and prior models from existing experimental data. The resulting noise model includes novel features such as heavy-tailed additive noise and a gene-specific bias term. We also verify that the resulting noise and prior models fit data from an Affymetrix human chip set.

**Key words:** gene arrays, microarrays, oligonucleotide arrays, noise model, Affymetrix, Bayesian estimation, statistical significance.

## 1. INTRODUCTION

GENE ARRAY TECHNOLOGIES INCLUDING OLIGONUCLEOTIDE ARRAYS and cDNA microarrays have recently allowed researchers to simultaneously measure the expression levels of thousands of genes in a cell population, providing a potentially powerful tool in biology and medicine (Brown and Botstein, 1999; Holstege *et al.*, 1998; Hughes *et al.*, 2000; Jelinsky and Samson, 1999; Tanaka *et al.*, 2000; Wodicka *et al.*, 1997). These experiments involve a large number of error-prone steps that lead to a high level of noise in the resulting data (Schuchhardt *et al.*, 2000). This noise raises practical questions in interpreting

---

<sup>1</sup>Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA 02139.

<sup>2</sup>Department of Biology, MIT, Cambridge, MA 02139, and Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142.

<sup>3</sup>Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139.

<sup>4</sup>Department of Operations Research and Center For Biological and Computational Learning, MIT, Cambridge, MA 02139.

\*The first two authors contributed equally to this work.

experimental results: How should one combine multiple observations of the same transcript level into a single estimate? How should one determine ratios of transcript levels under different conditions, given one or more observations of each? How should one handle the negative observations reported by many array technologies? How can one quantify the statistical significance of a result based on gene array data?

The Bayesian estimation of array measurements (BEAM) framework described in this paper addresses all of these problems rigorously and accurately. Given one or more gene array measurements, a statistical model of measurement noise, and any available prior information about the transcript levels, BEAM produces a statistically optimal estimate of an expression level or expression level ratio, as well as a measure of its uncertainty. BEAM quantifies the significance of expression level changes, either in a strict Bayesian framework or in terms of a  $p$ -value. We also describe a method for deriving accurate noise and prior models from existing experimental data so that BEAM can be tailored to any gene array system. Rather than assuming specific analytical forms for measurement noise, BEAM relies on computational techniques that apply to a wide range of noise and prior models.

Early studies using gene arrays recognized that measurement noise was a significant problem and that repetition can increase the significance of conclusions from gene array experiments (Lee *et al.*, 2000; Wodicka *et al.*, 1997). A number of authors have proposed models for measurement noise in oligonucleotide and cDNA arrays, as well as various methods to compute ratios of expression levels and to detect significant changes in expression level (Baggerly *et al.*, 2001; Chen *et al.*, 1997; Kerr and Churchill, 2000; Kerr *et al.*, 2000; Li and Wong, 2001; Mills and Gordon, 2001). Newton *et al.* (2001), Baldi and Long (2001), Long *et al.* (2001), and Theilhaber *et al.* (2001) have independently developed Bayesian approaches to computing expression level changes. Other recent work has proposed a combination of additive and multiplicative components to model array measurement noise, an approach we adopt here (Hughes *et al.*, 2000; Ideker *et al.*, 2000; Rocke and Durbin, 2001). However, these authors use noise and prior models chosen for their computational convenience.

In contrast, we propose numerical methods that allow us to handle general noise and prior models. We also present techniques to derive accurate models for a particular array technology. Our work is unique in its use of an experimental database to explicitly examine the forms of each component of the noise and prior models. BEAM is able to incorporate prior information about the experimental system without sacrificing computational elegance. We emphasize the computation of absolute transcript levels, ratios of transcript levels, and significance tests in a single coherent framework.

We illustrate our technique by developing a measurement noise model for Affymetrix yeast chips on the basis of a previously existing database of experiments performed over a one-year period in Richard Young's lab at the Whitehead Institute. Our noise model includes measurement-specific additive and multiplicative components as well as a randomly distributed gene-specific bias term. We find that the additive noise follows a heavy-tailed, non-Gaussian distribution. We also use the experimental database to derive a nonuniform prior model for the distribution of actual expression levels. We find that the same noise and prior models provide a good fit to data from an Affymetrix human chip set. Our general approach applies to other types of array data, although the details of the noise and prior models are likely to differ.

TABLE 1. EXAMPLES OF NOISE SOURCES IN AFFYMETRIX CHIPS

	<i>Multiplicative</i>	<i>Additive</i>
<b>Whole-chip</b>	<ul style="list-style-type: none"> <li>• Variation in hybridization time</li> <li>• Variation in reagent concentrations</li> <li>• Model term: <math>\nu</math></li> </ul>	<ul style="list-style-type: none"> <li>• Leak of external light during chip reading</li> <li>• Model term: (none)</li> </ul>
<b>Gene-specific</b>	<ul style="list-style-type: none"> <li>• Sequence-specific effect on hybridization efficiency</li> <li>• Model term: <math>\zeta</math></li> </ul>	<ul style="list-style-type: none"> <li>• Specific cross-hybridization by oligonucleotides that are typically present in the sample</li> <li>• Model term: <math>f</math></li> </ul>
<b>Measurement-specific</b>	<ul style="list-style-type: none"> <li>• Inhomogeneities in chip preparation</li> <li>• Variations in laser intensity during chip reading</li> <li>• Model term: <math>g</math></li> </ul>	<ul style="list-style-type: none"> <li>• Nonspecific hybridization</li> <li>• Model term: <math>e</math></li> </ul>

Our estimators agree qualitatively with popular heuristics that have been developed by experimentalists. However, the BEAM technique serves to quantify and extend these methods, improving the accuracy of gene array data analysis and providing measures of confidence in experimental results.

## 2. NOISE MODEL DERIVATION

The power of Bayesian estimation is dependent on the quality of the noise and prior models, so we derived models that provided a good fit to real data. For model construction, we used a data set consisting of measurements from 261 Affymetrix Ye6100 chip sets, recorded by a variety of researchers and reflecting a variety of experimental manipulations (see Appendix A).

The prior model  $p_x(x)$  is a probability distribution<sup>1</sup> over the quantity to be estimated—here a transcript level—describing any information available about this quantity prior to collection of the measurements. For example,  $p_x(x)$  will capture the fact that  $x$  cannot be negative ( $p_x(x) = 0$  when  $x < 0$ ) and that very high transcript levels are unlikely. The noise model specifies the conditional distribution  $p_{\mathbf{y}|x}(\mathbf{y}|x)$ , which indicates the probability of obtaining the measurements  $\mathbf{y} = (y_1, y_2, \dots, y_n)$  from separate arrays for a particular gene whose transcript level is  $x$ .

Noise in a chip or array measurement can be introduced from many sources, both during sample preparation and during hybridization and array reading (Schuchhardt *et al.*, 2000). A particular source of noise can be broadly categorized into either *multiplicative noise*, where the noise magnitude is proportional to the signal magnitude, or *additive noise*, where the noise is independent of the signal. In addition to considering both additive and multiplicative noise, we also distinguished between noise that affects every measurement on a chip similarly, noise that affects measurements of a particular gene on all chips similarly, and noise that is independent from measurement to measurement both within and between chips. Thus, we considered six types of noise terms for our model: additive and multiplicative, each either whole-chip, gene-specific, or measurement-specific. Table 1 lists examples of noise sources that we might expect to behave in each of these ways. We constructed the model based purely on the statistics of the data set; we did not attempt to correspond any particular term with a specific real-world noise source.

In deriving a noise model from our data set, we tried to balance the competing desires for a model complex enough to explain the data well yet simple enough to derive easily and use efficiently for estimation. Ultimately, we found the data was fit most closely by a model of the form

$$\tilde{y}_{ij} = v_i[g_{ij}(\zeta_j t_{ij}) + f_j + e_{ij}], \quad (1)$$

where  $\tilde{y}_{ij}$  is the unnormalized measurement for gene  $j$  on chip  $i$  computed from differences of match and mismatch probes by the Affymetrix Microarray Analysis Suite software,  $t_{ij}$  is the true transcript concentration of gene  $j$  on chip  $i$ ,  $\zeta_j$  is a constant factor specific to gene  $j$ ,  $f_j$  is an additive bias specific to gene  $j$ ,  $v_i$  is a noise term multiplying all genes on chip  $i$  (i.e., a “normalization” term), and  $g_{ij}$  and  $e_{ij}$  are multiplicative and additive noise terms, respectively, specific to gene  $j$  on chip  $i$ .

Because the amplification and hybridization steps are affected by the particular base sequence, the proportion between absolute transcript level and Affymetrix chip response differs for each gene (Lockhart *et al.*, 1996). We call this constant of proportionality  $\zeta_j$  and estimate the product  $x_{ij} = \zeta_j t_{ij}$ . Most experimental results depend on changes in expression level under different conditions or in different cells; such results are independent of the value of  $\zeta_j$ . To measure absolute expression levels, one could determine

---

<sup>1</sup>We use the following probabilistic notation throughout this paper. Latin characters without serifs ( $x$ ,  $\mathbf{y}$ ,  $f$ ) denote random variables, and the corresponding characters with serifs ( $x$ ,  $\mathbf{y}$ ,  $f$ ) denote sample values of those random variables. Boldface characters ( $\mathbf{y}$ ) denote vector quantities.

- $p_x(x)$ : probability density of  $x$ , evaluated at  $x$ .
- $p_{x|y}(x|y)$ : conditional probability density of  $x$  given an observation  $y$  of  $y$ .
- $E(x|y)$ : expected value of  $x$  given the observation  $y$  of  $y$ . This is shorthand for  $E(x|y = y)$ .
- $\hat{x}(y)$ : Bayes least squares estimate of  $x$  given the observation  $y$  of  $y$ .
- $\sigma_x^2(y)$ : variance of the posterior distribution of  $x$  given the observation  $y$  of  $y$ .

the constants  $\zeta_j$  through control experiments with known concentrations of gene  $j$ . Henceforth, we will use the term “true transcript level” to denote  $x_{ij}$ .

### 2.1. Whole-chip noise

Whole-chip noise is assumed identical for each measurement on a single chip. Thus, control transcripts are often added in known amounts to each chip for use in estimating the whole-chip noise and removing it from the other probes on the chip. The process of whole-chip multiplicative noise removal is known commonly as normalization, and consists of multiplying every measurement on the chip by some factor ( $1/v_i$  in our model).

Several authors have treated the problem of normalization of array data (Hartemink *et al.*, 2001; Schadt *et al.*, 2000; Schuchhardt *et al.*, 2000). We chose to apply the normalization algorithm of Hartemink *et al.* (2001), because it is based on assumptions about the measurement-specific noise that are consistent with a simplified form of the noise model developed here. After estimating  $v_i$  for each chip using their algorithm, we multiplied each chip’s data by  $1/v_i$ . Henceforth, we will use the terms “observed value” and “measured value” to denote the normalized observation  $y_{ij} = \tilde{y}_{ij}/v_i$ .

We doubted whether we would need to include a whole-chip additive term, since the Affymetrix chip reader performs background subtraction that would be expected to eliminate this type of noise. Additionally, individual transcript levels are calculated from differences between measurements of match and mismatch probes, a procedure that would likely remove any additive noise affecting every probe equally.

Despite these considerations, we performed a further check for whole-chip additive noise by examining the correlations between unspiked control transcripts. Unspiked control transcripts represent sets of chip hybridization spots for which no corresponding mRNA should have been present in the samples (see Appendix A.) Thus, all readings from these probes represent entirely noise. There were nine such unspiked controls on each chip in our data set. Whole-chip additive noise would cause all nine of these to vary up or down together, resulting in consistent positive correlations between them. However, when we calculated the correlation coefficients for all pairs of unspiked controls, 19 of 36 were negative, and none had a magnitude larger than 0.4. We concluded that it was unnecessary to include a whole-chip additive noise term in our model.

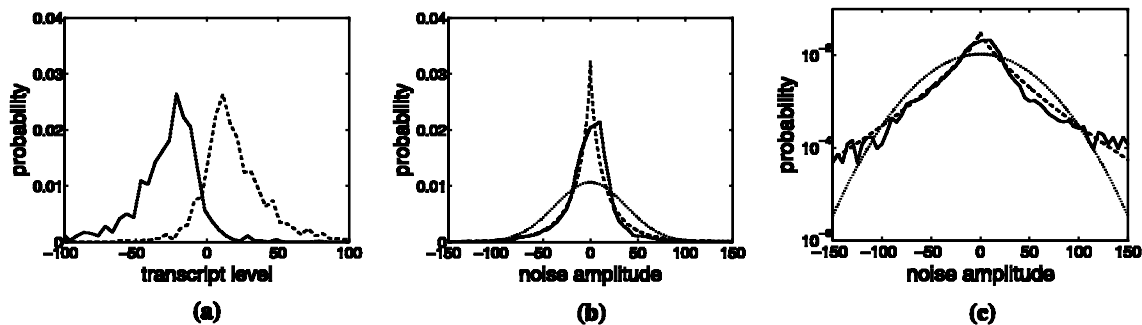
### 2.2. Gene-specific and measurement-specific noise

Measurement-specific noise must be treated fundamentally differently from whole-chip noise. Since the same normalization factor  $v_i$  multiplies every gene on chip  $i$ , each of the thousands of measurements from the chip contains some independent information about  $v_i$ . We can use some or all of these measurements to explicitly estimate  $v_i$ , as we do above. In contrast, only the single observation  $y_{ij}$  contains information about  $g_{ij}$  and  $e_{ij}$ , so we cannot determine the values of these noise terms through measurement of control transcripts. We can, however, learn their distributions, summarized by the probability distribution functions  $p_e(e)$  and  $p_g(g)$ .

We also find that the measurements contain gene-specific additive noise (i.e., bias). Even though this bias  $f_j$  is a constant for every measurement of a given gene, it is typically unknown. Such an unknown additive constant prevents accurate determination of the transcript level ratio for a particular gene under different conditions, an important goal of microarray experiments. To predict the behavior of the bias, we estimate its distribution,  $p_f(f)$ . We can then determine the probability distribution of the observed transcript level  $y$  given the true transcript level  $x$ :

$$y = gx + f + e. \quad (2)$$

To determine the distributions of the noise terms  $e$ ,  $f$ , and  $g$ , we made use of the twenty-four control probes on each Affymetrix chip that correspond to RNA sequences not normally present in yeast. Transcripts corresponding to fifteen of these probes (the “spiked controls”) are added in a cocktail in fixed, known quantities at the start of sample preparation in every experiment. Since these controls were added in the same concentration every time, any variation in the corresponding measurements must be due to some source of noise. In addition, any nonzero reading for the remaining nine controls (the “unspiked controls”) must represent entirely noise.



**FIG. 1.** Additive noise. (a) Distribution of observed transcript levels for two of the nine unspiked controls (solid: *bioC 3'*; dashed: *bioC 5'*). (b) Solid: average distribution of observations for unspiked control probes, with their medians subtracted. Dashed: generalized Laplacian distribution fit using a maximum likelihood criterion ( $s_e = 13$ ,  $\alpha_e = 0.76$ ). Dotted: best Gaussian fit. (c) Same as (b), shown on a  $\log_{10}$  vertical axis.

Several authors have used a combination of additive and multiplicative terms to describe noise in microarray data (Hughes *et al.*, 2000; Ideker *et al.*, 2000; Rocke and Durbin, 2001). We found strong evidence for both of these types of noise in the Young lab data set, as well. Briefly, we found that the variance of measurements of each control probe increases with increasing transcript level, indicating the presence of multiplicative noise. At the same time, the variance does not drop to zero for transcript levels at or near zero. This indicates the presence of additive noise as well.

*2.2.1. Additive measurement-specific and gene-specific noise.* After establishing that we needed to model a combination of additive and multiplicative noise, we estimated the additive noise distribution from the unspiked controls. For these transcripts,  $y = f + e$ , and any nonzero reading is entirely additive noise. The measurements for each of the nine unspiked controls were similarly distributed, but they had distinct and generally nonzero means (Fig. 1a). This suggests the presence of gene-specific biases that are identical from chip to chip, as well as measurement-specific noise that is independent from chip to chip.

We subtracted these means from each of the unspiked controls and plotted their average distribution (Figs. 1b and 1c). We found that this distribution had much heavier tails than a Gaussian. We instead chose to fit it with a generalized Laplacian distribution<sup>2</sup>:

$$p_e(\cdot) = \frac{1}{C} e^{-|\frac{\cdot}{s_e}|^{\alpha_e}} \tag{3}$$

where  $s_e$  and  $\alpha_e$  are fit parameters and  $C$  is a normalization constant.

We also needed to derive a distribution for the gene-specific bias  $f$ . Since there were only nine unspiked controls in our data set, we had only nine independent samples of bias available (e.g., the means of each of the unspiked controls). Since there were insufficient samples to define a more complex distribution, we chose a zero-mean Gaussian as the distribution for  $f$  ( $\sigma_f = 36$ ). The means of the unspiked controls were consistent with being drawn from this distribution.

*2.2.2. Multiplicative measurement-specific noise.* To characterize the measurement-specific multiplicative noise  $g$ , we used the spiked controls, for which  $x_j$  is nonzero, but constant from chip to chip. Thus,

$$y_j = gx_j^* + f_j + e, \tag{4}$$

where  $y_j$  is a measurement of spiked control  $j$  and  $x_j^*$  is its true transcript level. The constant additive bias specific to control  $j$  is  $f_j$ . The distribution of  $y_j$  is a convolution of the distribution of  $gx_j^* + f_j$

<sup>2</sup>This distribution is alternately known as the generalized Gaussian distribution. If  $\alpha_e = 2$ , it represents a Gaussian distribution. When  $\alpha_e < 2$ , it falls off more slowly at large values than a Gaussian with the same variance and therefore describes a heavy-tailed distribution that is more likely to produce extreme values.



and that of  $e$ . We derived the distribution of  $e$  in the previous section. We can, therefore, approximate the distribution of  $gx_j^* + f_j$  by deconvolving the distribution of  $e$  from the observed distribution of  $y_j$ . In order to enforce positivity constraints, we implement this deconvolution using a quadratic programming algorithm (Appendix B).

Note that for each control, the bias  $f_j$  is simply a constant. Therefore, the shape of the distribution of  $gx_j^* + f_j$  is unaffected; it merely gets shifted left or right by an unknown constant. If the true value  $x_j^*$  is large compared to the bias, this shift is negligible. We therefore determined  $p_g(g)$  by examining only those nine probes corresponding to the control sequences spiked in at the highest levels. All of these probes had median observations 10–50 times larger than typical bias values we determined above.

By deconvolving the  $p_e(e)$  distribution from each of these nine spiked controls and dividing by the median, we produced nine approximations of the distribution  $p_g(g)$ . The average of these nine distributions, shown in Fig. 2, is our best estimate of the distribution of  $p_g(g)$ . We found that a log normal fit this distribution well.

As a visual check of the noise model, we compared our model’s predicted distribution to the actual distribution for each of the 15 spiked controls (Fig. 3). For each control, we assume that the median of the observations is the true value  $x_j^*$ ; we apply our noise model to find the expected distribution of measurements. The noise model fits the observed data well, providing evidence that the model describes most of the noise present in the data, over a wide range of transcript concentrations.

### 2.3. Prior distribution

To derive a prior distribution, we used all of the normalized measurements from every chip in the dataset, with the exception of the control probes. The resulting dataset represents a typical distribution of chip readings for yeast data across all genes and hundreds of experiments under a wide variety of experimental conditions (see Fig. 4). The distribution of these measurements represents the prior distribution we seek, corrupted by the noise we wish to remove.

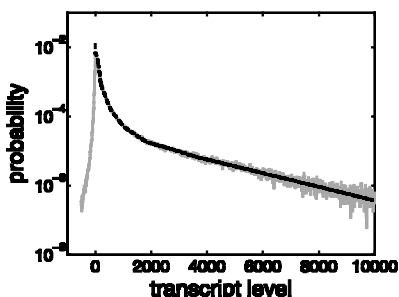
We inferred a noise-free version of the prior through two deconvolution steps. We first deconvolved the additive noise distribution from the distribution of observations. In this case, we deconvolved the distributions both of the independent noise  $e$  and of the bias  $f$ . Unlike individual spiked controls, for which  $f$  takes on a single value across all measurements, the full dataset shown in Fig. 4 represents observations of more than 6,000 genes. Each of these genes can be expected to have a different bias  $f_j$ . The full bias distribution is therefore contained within the data set.

After deconvolving the additive noise, the prior distribution is corrupted only by multiplicative noise  $g$ . To remove it, we convert the distribution to the log domain, noting that  $\log(gx) = \log g + \log x$ . The prior distribution is then computed by deconvolution with  $p_{\log g}$  followed by conversion back to the linear domain. The denoised prior is shown in Fig. 4.

There are relatively few samples in the data set with readings of 2,500 and above. The sparsity of the data at these high levels introduces artifacts into the shape of the prior. For use in estimation, we fit a closed-form function to the prior distribution at transcript levels above 4,000:

$$p_x(x) = ke^{-\lambda x}, \tag{5}$$

with  $k = 3.4 \cdot 10^{-5}$  and  $\lambda = 4.6 \cdot 10^{-4}$ .



**FIG. 4.** Solid gray line: distribution of all the readings from all the chips in the Young lab data set, with the exception of controls. Dashed black line: calculated prior distribution.

## 2.4. Summary of model

We began this section seeking a prior distribution  $p_x(x)$  on true transcript levels and a conditional distribution  $p_{y|x}(y|x)$  on observed levels given a true level  $x$ . We derived  $p_x(x)$  from data in Section 2.3 (Fig. 4). The conditional distribution  $p_{y|x}(y|x)$  is completely defined by Equation (2) and the distributions of  $e$ ,  $f$ , and  $g$ . The distribution  $p_g(g)$  is a log normal distribution such that  $\log g$  is zero-mean with standard deviation  $\sigma_g = 0.21$ ;  $p_f(f)$  is a zero-mean Gaussian distribution with standard deviation  $\sigma_f = 36$ ; and  $p_e(e)$  is a generalized Laplacian distribution with parameters  $s_e = 13$  and  $\alpha_e = 0.76$ . Thus, the noise model has four parameters in total:  $\sigma_g$ ,  $\sigma_f$ ,  $s_e$ , and  $\alpha_e$ .

## 3. ESTIMATION

### 3.1. Estimation of absolute transcript levels

To estimate an absolute transcript level of a gene, we begin with one or more observations  $\mathbf{y} = (y_1, y_2, \dots, y_n)$  corresponding to that gene on different arrays. We use Bayes's rule to write the probability distribution  $p_{x|\mathbf{y}}(x|\mathbf{y})$  of the true transcript level  $x$  given the observations as

$$p_{x|\mathbf{y}}(x|\mathbf{y}) = \frac{p_{\mathbf{y}|x}(\mathbf{y}|x)p_x(x)}{p_{\mathbf{y}}(\mathbf{y})}, \quad (6)$$

where the prior distribution on transcript levels determines  $p_x(x)$ , the noise model determines  $p_{\mathbf{y}|x}(\mathbf{y}|x)$ , and the normalization constant  $p_{\mathbf{y}}(\mathbf{y})$  may be computed by integrating the numerator of Equation (6) over all  $x$ .

We can summarize this probability distribution by computing a single estimate for the true transcript level. We use the Bayes least squares estimator, which is simply the expected value (mean) of the probability distribution of  $x$  given the observations  $\mathbf{y}$ :

$$\hat{x}(\mathbf{y}) = E[x|\mathbf{y}] = \int x p_{x|\mathbf{y}}(x|\mathbf{y}) dx = \frac{1}{p_{\mathbf{y}}(\mathbf{y})} \int x p_{\mathbf{y}|x}(\mathbf{y}|x) p_x(x) dx. \quad (7)$$

A measure of the uncertainty in the estimate is given by the variance of the posterior distribution:

$$\sigma_{\hat{x}}^2(\mathbf{y}) = E[(x - \hat{x})^2|\mathbf{y}] = \frac{1}{p_{\mathbf{y}}(\mathbf{y})} \int (x - \hat{x})^2 p_{\mathbf{y}|x}(\mathbf{y}|x) p_x(x) dx. \quad (8)$$

Analytical evaluation of Equations 7 and 8 proves difficult for most noise and prior models, including those developed in Section 2. We therefore evaluate these quantities through numerical integration. Efficient methods to perform these computations are described on our Web page at [www.ai.mit.edu/projects/BEAM](http://www.ai.mit.edu/projects/BEAM). We store lookup tables of estimates and uncertainties over a range of plausible measurements. For a given noise model, these tables need be computed only once. Subsequently, one can interpolate the results to find an estimated transcript level and an uncertainty measure corresponding to any given observation.

Figure 5a displays the Bayes least squares estimate of the transcript level as a function of the value of a single observation. The estimate approximates the observation for large observed values. As the observation falls below zero, however, the estimated level remains positive, because the prior distribution rules out negative transcript levels. As the observed level becomes increasingly negative, the estimate flattens out, even increasing slightly after reaching a minimum value of 23 at an observed level of  $-75$ . This behavior reflects the heavy-tailed nature of the additive component of the noise model. A highly negative observation does not necessarily indicate that the true transcript level is zero; according to our model, it is more likely to be the result of a positive true level with a slightly greater noise contribution. Figure 5b shows the uncertainty in the estimate, measured by the standard deviation of the posterior distribution. The multiplicative component of the noise model results in an increase in the uncertainty at high observed levels.

Figure 6 shows estimated transcript levels as a function of two repeated observations of the same quantity, typically corresponding probes from two gene arrays run separately. The figure also includes the

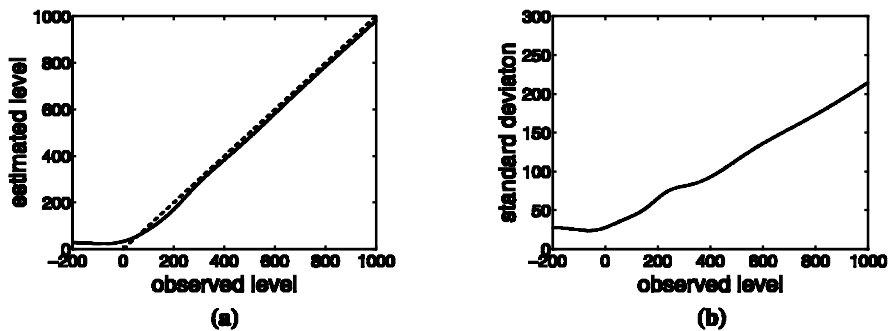


FIG. 5. Transcript level estimates based on a single observation. (a) Bayes least squares estimate of true transcript level from a single observed value. The dashed line corresponds to  $y = x$ . (b) Uncertainty of the estimate, measured by the standard deviation (square root of the variance) of the posterior distribution of transcript values given the observation.

uncertainty of the estimates. These estimates were computed numerically according to Equation (7); the results are stored in a lookup table with two independent variables.

### 3.2. Estimation of transcript level ratios

Analysis of gene array data in practice often involves the ratios of two different transcript levels, usually corresponding to the same mRNA under two different conditions. Given one or more observations  $\mathbf{y}_a$  under condition  $a$  and one or more observations  $\mathbf{y}_b$  under condition  $b$ , the optimal estimate of the log ratio  $r = \log_{10} \frac{x_a}{x_b}$  is given by the expectation of the posterior distribution on  $r$ ,

$$\hat{r}(\mathbf{y}_a, \mathbf{y}_b) = E \left[ \log_{10} \frac{x_a}{x_b} \mid \mathbf{y}_a, \mathbf{y}_b \right]. \tag{9}$$

We estimate logarithms of ratios, so swapping the two observations simply negates the estimate. This estimated log ratio differs from the log ratio of the estimates of absolute transcript levels. To quantify the uncertainty in the log ratio estimate, we compute the variance of the posterior distribution over  $r$ ,

$$\sigma_{\hat{r}}^2(\mathbf{y}_a, \mathbf{y}_b) = E[(r - \hat{r})^2 \mid \mathbf{y}_a, \mathbf{y}_b]. \tag{10}$$

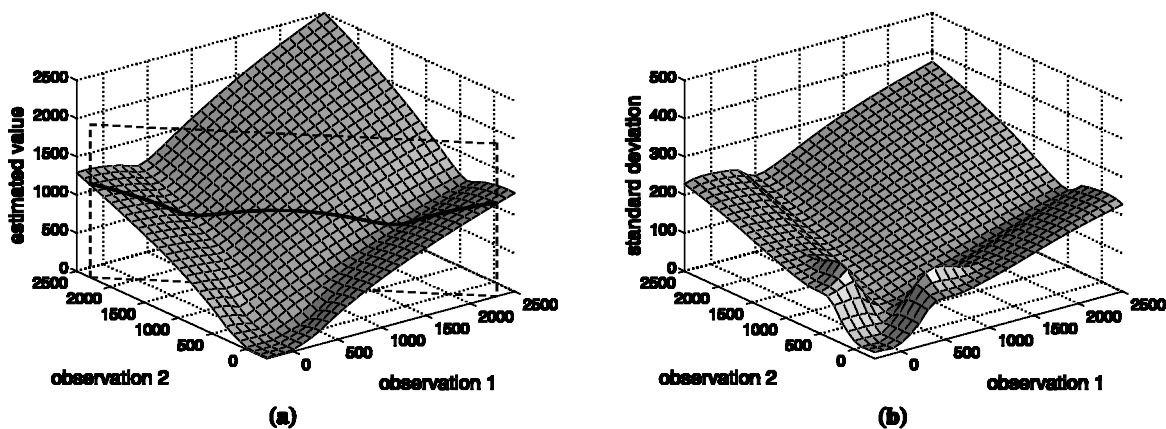


FIG. 6. Transcript level estimates based on repeated observations. (a) Bayes least squares estimate of true transcript level as a function of observations in two independent, repeated experiments. The cross-section indicated by the plane in this figure is illustrated in Fig. 11b. (b) Uncertainty of the estimate, measured by the standard deviation of the posterior distribution of the transcript level given the two observations.

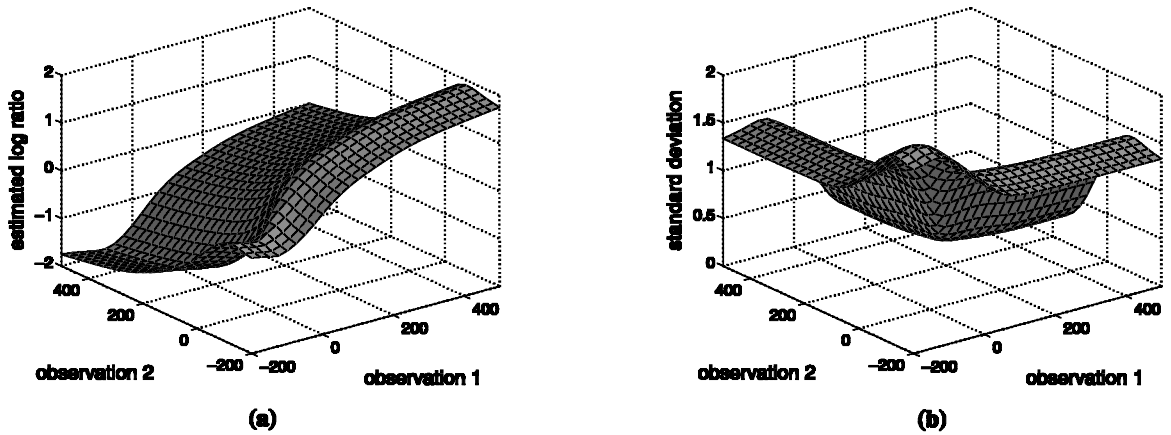


FIG. 7. Ratio estimates. (a) Bayes least squares estimate of base 10 log ratio of transcript levels of a particular gene observed under conditions a and b. (b) Uncertainty of the log ratio estimate, measured by the standard deviation of the posterior distribution of the ratio given the two observations.

The computation of  $\hat{r}(\mathbf{y}_a, \mathbf{y}_b)$  and  $\sigma_r^2(\mathbf{y}_a, \mathbf{y}_b)$  must take into account the fact that the observations  $\mathbf{y}_a$  and  $\mathbf{y}_b$  are influenced by the same, unknown bias term and are therefore not independent. Figure 7 shows the resulting log ratio estimates as well as their associated uncertainties, for the case of a single observation under each condition.

### 3.3. Significance tests

One often wishes to determine whether two sets of measurements  $\mathbf{y}_a$  and  $\mathbf{y}_b$  represent a significant difference in actual transcript levels. Using a strict Bayesian approach, one can test significance by calculating the probability that the true ratio of transcript levels is sufficiently different from unity. In particular, one might compute  $P(|r| > k | \mathbf{y}_a, \mathbf{y}_b)$ , the probability that the absolute value of the log ratio of the true transcript levels is greater than some threshold value  $k$ . Evaluation of this quantity given noise and prior models is a straightforward extension of the computation of our log ratio estimator  $\hat{r}$ .

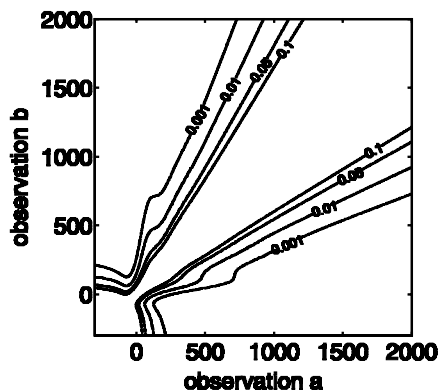
In the biological and medical literature, however, statistical significance is almost always quantified as a  $p$ -value. We therefore derived a method for computation of  $p$ -values given noise and prior models, which is consistent with our estimation-theoretic approach to transcript estimation. Once we have estimated the ratio of two transcript levels measured under different conditions, we associate a  $p$ -value with the estimated ratio. This  $p$ -value quantifies the probability that the estimated ratio would be at least this extreme by chance, under a null hypothesis that the true mRNA level of a given gene is the same under the two conditions.

Given two sets of observations  $\mathbf{y}_a$  and  $\mathbf{y}_b$  of the same gene under different conditions, we compute the estimated log ratio of  $x_a$  to  $x_b$  using Equation (9) and designate it as  $\hat{r}^*$ . We seek to determine the probability under the null hypothesis of finding a log ratio  $r$  such that  $|r|$  is greater than or equal to  $|\hat{r}^*|$ .

If our null hypothesis is that the true underlying transcript level is identically  $x^*$  under both conditions, it is straightforward to determine the probability of observing a ratio larger than  $\hat{r}(\mathbf{y}_a, \mathbf{y}_b)$ . For a known true transcript level  $x^*$ , the probability of observing any set of chip readings  $\mathbf{y}$  is given by the noise model. For any two sets of observations  $\mathbf{y}_1$  and  $\mathbf{y}_2$ , the estimated ratio is  $\hat{r}(\mathbf{y}_1, \mathbf{y}_2)$ , so we are interested in the total area for which  $|\hat{r}(\mathbf{y}_1, \mathbf{y}_2)|$  is greater than  $|\hat{r}^*|$ :

$$p = \int_{|\hat{r}(\mathbf{y}_1, \mathbf{y}_2)| \geq |\hat{r}^*|} p_{\mathbf{y}|x}(\mathbf{y}_1, \mathbf{y}_2 | x^*) d\mathbf{y}_1 d\mathbf{y}_2. \quad (11)$$

However, we do not know  $x^*$  exactly but can only estimate it from our two observations  $\mathbf{y}_a$  and  $\mathbf{y}_b$ . In fact,  $x^*$  could have a continuum of values with varying probabilities. For the most accurate  $p$ -value, we



**FIG. 8.** The  $p$ -value contours showing the probability that the ratio of two independent observations of the same gene, or a more extreme ratio, could have resulted if the underlying mRNA level had been the same in both cases. Contours for  $p = 0.1, 0.05, 0.01,$  and  $0.001$  are shown.

should consider all possible values of  $x^*$ , weighted by the probability of each. Ultimately, we calculate  $p$  as follows:

$$p = \int_{x^*=0}^{\infty} \left[ \int_{|\hat{r}(y_1, y_2)| \geq |\hat{r}^*|} p_{y|x}(y_1, y_2 | x^*) dy_1 dy_2 \right] p_{x|y}(x^* | y_a, y_b) dx^*. \tag{12}$$

Contours for several commonly considered values of  $p$  are shown in Fig. 8.

### 4. VALIDATION

We first verified that different subsets of the Young lab data implied similar noise and prior models. We fit the noise model parameters and prior distribution to 100 different random selections of 87 chip sets (one-third of our full complement of 261). The noise model parameters are extremely stable (Table 2), all having a coefficient of variation less than 0.08. The calculated prior distribution proved equally stable (not shown). This behavior suggests that our model describes general features of the dataset and does not overfit.

To test the broader applicability of our model to data from other labs and from other types of Affymetrix chips, we examined the fit to data publicly available from Affymetrix in their human Latin square data set (see Appendix A). This data set, consisting of 59 U95A human chips, was constructed by spiking in 14 different human genes at various concentrations on top of a constant background of mRNA derived from human pancreas. It is this constant background that interested us, as it provided more than 10,000 genes, each repeated at the same concentration on all 59 chips.

TABLE 2. STABILITY OF PARAMETER FITS TO 100 DIFFERENT SUBSETS OF THE YOUNG LAB DATABASE, EACH CONSISTING OF 1/3 OF THE EXPERIMENTS IN THE DATABASE<sup>a</sup>

	<i>Model value</i>	<i>Std. Dev.</i>	<i>CV</i>
$\alpha_e$	0.76	0.031	0.041
$s_e$	13	1.0	0.076
$\sigma_f$	36	1.6	0.044
$\sigma_g$	0.21	0.0081	0.039

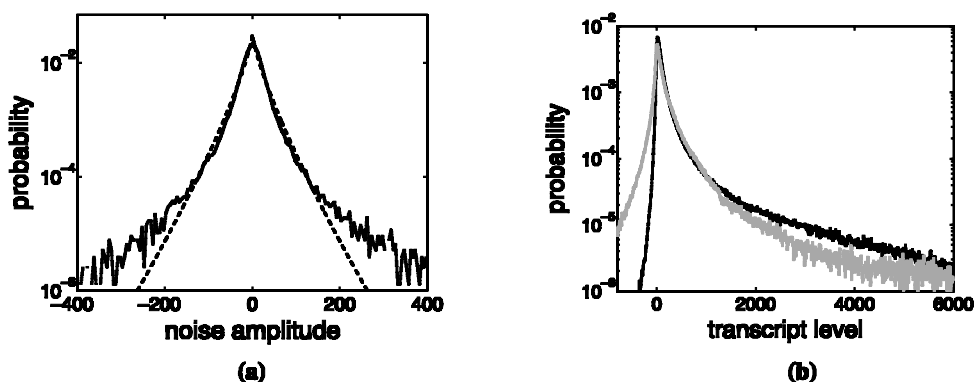
<sup>a</sup>The first column shows the parameter value fit to the full data set, the second shows the standard deviation of parameters fit to the 100 subsets, and the third shows their coefficient of variation, defined as the ratio of the standard deviation to the mean.

We checked whether these data exhibited the same heavy-tailed additive noise that we observed in the Young lab data set. Since the Affymetrix data set did not contain any unspiked controls, we selected all genes with a median very close to zero. Specifically, we examined all those with a median falling in the range  $-20$  to  $+20$ . For these, we would expect additive noise to be the predominant noise source. After subtracting the medians, we combined all of these observations to see the shape of the additive noise distribution (Fig. 9a). This distribution was symmetric and heavy-tailed, just as we observed in the yeast data set. The parameters of the generalized Laplacian best fitting these data ( $s_e = 13$ ,  $\alpha_e = 0.78$ ) are very similar to those in our noise model ( $s_e = 13$ ,  $\alpha_e = 0.76$ ), even though the tails of the actual distribution are heavier than those predicted by the generalized Laplacian fit. Given that our model was fit using yeast chips and that it describes well data recorded on newer-generation human chips, we believe that heavy-tailed additive noise may be a general feature of Affymetrix gene chip measurements.

We found strong evidence for gene-specific biases in the Affymetrix data as well. We considered first what we would observe if every single one of the 12,519 genes on the chip had a true transcript level of zero. If the noise was symmetric and independent from one sample to the next, we would expect about half of the observations to fall above zero and half to fall below. Gross deviations from half-above and half-below would be rare. We can define our expectations more precisely using a binomial distribution: for the 59 repeated experiments in the data set, we would expect that only 24 of the 12,519 total genes would have more than 40 measurements below zero. Since in reality many of the true transcript levels are greater than zero, we might expect even fewer than 24. However, in the Affymetrix dataset, 2,114 of 12,519 genes exhibited greater than 40 observations less than zero. This proportion is far greater than chance, suggesting the presence of noise that both is correlated between experiments and takes on negative values. The gene-specific bias,  $f$ , in our model satisfies both these criteria. These calculations, along with features of the Young lab data shown in Fig. 1a, indicate that some form of additive gene-specific bias is an important contributor to the noise in Affymetrix array data.

We also plotted the standard deviation versus the median for each of the genes and gene fragments in the Affymetrix dataset. These points fall roughly along a positively sloping line over much of their range, which is a hallmark of multiplicative noise (data not shown). The slope of this line should be roughly equal to the standard deviation of the multiplicative noise (Rocke and Durbin, 2001). The slope of this line was 0.20, fitted to points with median transcript levels between 500 and 5,000. This value is in close agreement with the  $\sigma_g = 0.21$  in our noise model, providing more evidence that our model may be applicable to other Affymetrix chip sets.

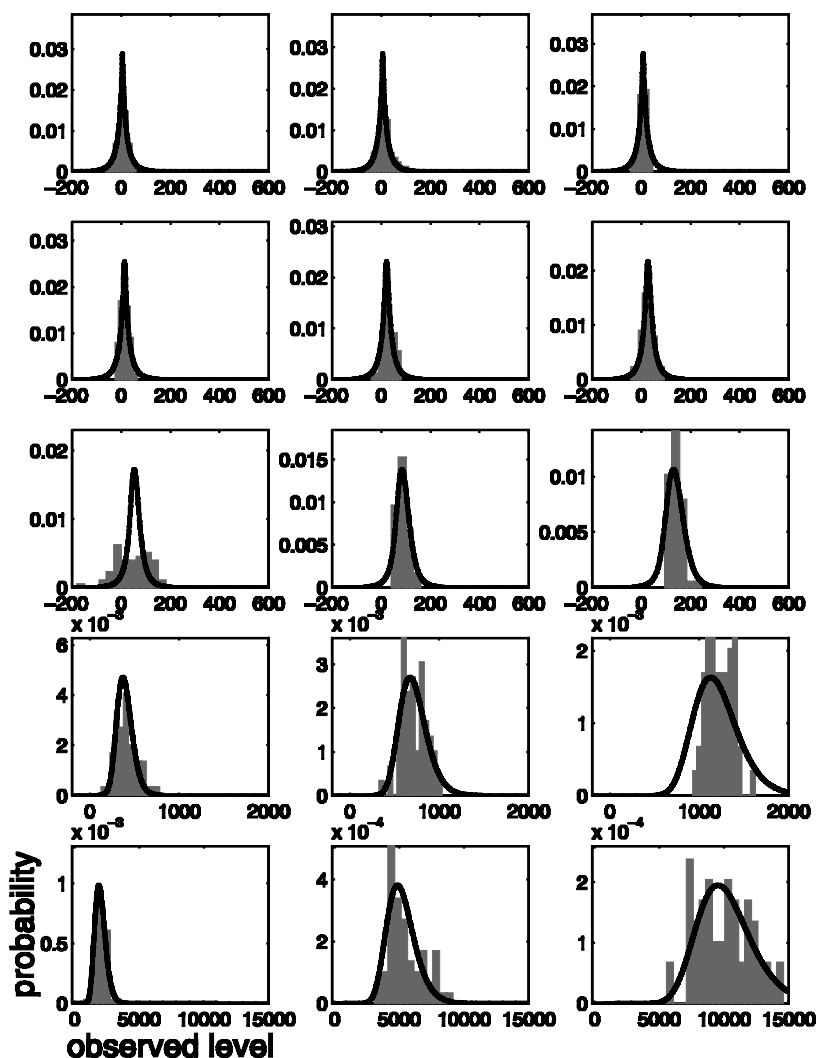
The prior distribution  $p_x(x)$  in our model was fit using yeast chip data, and we were interested in comparing it to the prior distribution of human genes. To derive a strictly correct prior distribution for the



**FIG. 9.** (a) Heavy-tailed nature of noise in Affymetrix human data. Solid line: combined distribution of all genes in the Affymetrix Latin square human data set with median observed transcript levels between  $-20$  and  $+20$ . Each gene had its median subtracted to remove bias. Dashed line: generalized Laplacian, fit using a maximum likelihood criterion ( $s_e = 13$ ,  $\alpha_e = 0.78$ ). On a log vertical axis, a Gaussian distribution would appear as an inverted parabola. (b) Comparison of human and yeast transcript level distributions. Distribution of complete data set, excepting control genes, for the Young lab yeast data set (black) and Affymetrix Latin square human data set (gray). Both data sets were normalized using the method of Hartemink *et al.* (2001).

Affymetrix human gene data set, we would need to perform the same model fitting procedures that we used in Section 2. However, we can get a reasonable understanding of the prior much more simply by examining the distribution of observed values across the dataset. Figure 9b highlights some similarities and differences between the Young lab and Affymetrix U95 data sets. Both distributions peak at a transcript level of zero and initially drop off similarly, especially in the positive direction. The differences between the distributions may occur because humans have a higher fraction of their genes expressed at a low level than do yeast. Such changes in the prior lead to only minor changes in our estimates (Section 5.2).

As a further check of our model, we examined the fit of our model to 15 randomly selected genes from the Affymetrix dataset. We took the median of the 59 repeated measurements to be the value of the true transcript level. We plotted the distribution of measurements expected according to our model against the distribution of the actual observations (Fig. 10). The expected and actual observations appear generally to agree, without any adjustment of our model parameters. These fits further justify specific features of our model as well, such as heavy-tailed additive noise and a log-normal multiplicative noise distribution.



**FIG. 10.** Predicted and actual distributions for human gene data. Each panel shows the observed distribution of transcript levels for one gene across the 59 chips in the Affymetrix Latin square data set (gray bars). Three genes were randomly selected from each of five ranges, in order to span the full range of typical transcript levels (0–10, 10–50, 50–250, 250–1600, 1600–10,000). The solid line is the predicted distribution according to our noise model. The median of the observed distribution is assumed to be the true level for each gene. The random selection procedure was performed only once, without regard to the goodness of the model fit, so this should be a representative sample.

As these are randomly selected genes from the Affymetrix data set, chosen without regard to goodness of model fit, we would expect that most of the genes on the U95 human chip would fit equally well.

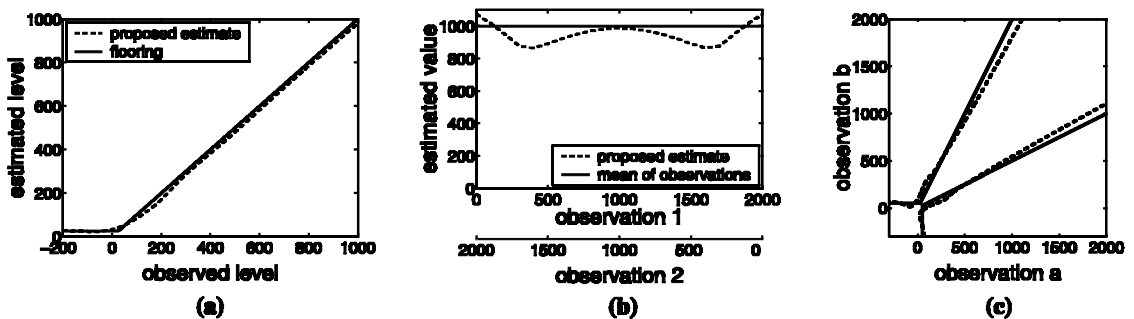
As a test of the applicability to human chips of the  $p$ -values computed from Equation (12) using the noise and prior models of Section 2, we examined the  $p$ -values calculated for pairs of experiments where the actual transcript levels were known to be identical. In this case, the null hypothesis that the measurements represent identical transcript levels is indeed true, so we expect 5% of the measurement pairs to result in  $p$ -values less than 0.05. For each of the 12,519 genes in the Latin Square data set that were present in the same concentration in all 59 experiments, we randomly selected pairs of experiments and computed the  $p$ -value associated with the two measurements in each pair. Across the entire data set, 5.9% of the pairs gave  $p$ -values below 0.05. When we restricted the set of genes to those whose median measurement was below 50, 5.5% of the pairs gave  $p$ -values less than 0.05. When we examined genes whose median values were above 500, 4.9% of the pairs gave  $p$ -values less than 0.05. Therefore, our method performs accurately for genes expressed at both high and low levels. On the other hand, when we randomly selected pairs of measurements corresponding to different genes in the data set, over 54% of the gene pairs resulted in  $p < 0.05$ .

## 5. DISCUSSION

We have shown that one can use Bayesian methods to optimally estimate transcript levels from one or two repeated experiments and to optimally estimate ratios of two expression levels, given a model for measurement noise and for the true distribution of expression levels. The BEAM technique also produces a statistical measure of confidence associated with each estimate. Given measurements of the same transcript on two different chips, the BEAM technique can generate a  $p$ -value for the hypothesis that the two expression levels are identical. We illustrated the application of this technique using detailed noise and prior models that we developed from a database of Affymetrix yeast chip experiments.

### 5.1. Comparison to heuristic approaches

Rather than contradicting current practice, our results quantify and extend the heuristics commonly applied in gene array analysis today. For example, investigators often deal with negative values by applying a technique known as “flooring,” in which observations below some small, positive threshold are set to that threshold while observations above that threshold are accepted as accurate (Holstege *et al.*, 1998). Figure 11a illustrates the mapping implied by a flooring method together with our estimator for a transcript level based on a single observation, from Fig. 5a. The two estimates are similar. Flooring can be viewed as a fairly accurate and simple approximation to the Bayes least squares estimator based on our prior and noise models.



**FIG. 11.** Comparisons to heuristic methods. (a) Estimated transcript levels from a single observation, using our method or the heuristic of flooring all values below the threshold of 25 to 25. (b) BEAM estimated transcript level given two observations whose mean is 1,000. This panel corresponds to the cross-section shown in Fig. 6a. (c) The dashed line shows the  $p = 0.05$  contour predicted by our method. The solid line shows the significance boundary determined using a factor-of-two criterion after flooring to a value of 25.

Our approach offers several benefits over the simple flooring heuristic even when only one observation is available. First, our approach provides a principled method to determine the value of the flooring threshold. Second, it provides a more accurate estimate in regions near the threshold, where the piecewise linear estimate due to flooring falls short. Third, our technique produces not only an estimate but also an associated measurement of uncertainty.

A similar comparison is possible for estimation from two or more repeated experiments or estimation of transcript ratios. The common practice of flooring followed by taking the mean or ratio produces results similar to the BEAM method for high transcript levels. However, when one or both observations are small or negative, BEAM estimates differ from these heuristic approaches. For example, Fig. 11b illustrates estimates based on two repeated experiments. When one measurement is smaller than the other, our method generally produces an estimate smaller than the mean of the two measurements. This results from the presence of a multiplicative noise component. The smaller observation, which would produce a posterior distribution of lower variance if used individually, receives the larger weight when the two observations are combined. If the larger observation remains constant while the smaller one becomes negative, the estimate begins to increase. The decreased weight on the smaller observation results from the heavy-tailed nature of the additive noise and corresponds to the intuition that a highly negative observation is known to be highly noisy. BEAM additionally provides an estimate of uncertainty in the transcript level or ratio, which decreases as the number of repetitions increases.

Current practice typically dictates that two transcript levels of the same gene are significantly different if one exceeds the other by a factor of two, after flooring to some value (Holstege *et al.*, 1998; Tanaka *et al.*, 2000). Figure 11c compares this practice to our calculation of  $p < 0.05$ , the most commonly encountered significance criterion. Interestingly, the ad hoc rule of factor of two defines a significance region of the same general shape as our statistical technique. At high transcript levels, significance at a given  $p$ -value occurs when one observation exceeds the other by a constant factor.

However, the linear portions of the BEAM contours do not correspond to lines that go through the origin. As transcript levels become smaller, the offset from the origin is increasingly important, and the factor by which one observation must exceed the other increases. For two observations  $y_a$  and  $y_b$ , where  $y_a$  is the lesser, we can roughly approximate the  $p < 0.05$  contour as

$$p < 0.05 \text{ if } \begin{cases} y_b > 1.8y_a + 45 & \text{if } y_a \geq 15 \\ y_b > 72 & \text{if } y_a < 15. \end{cases} \quad (13)$$

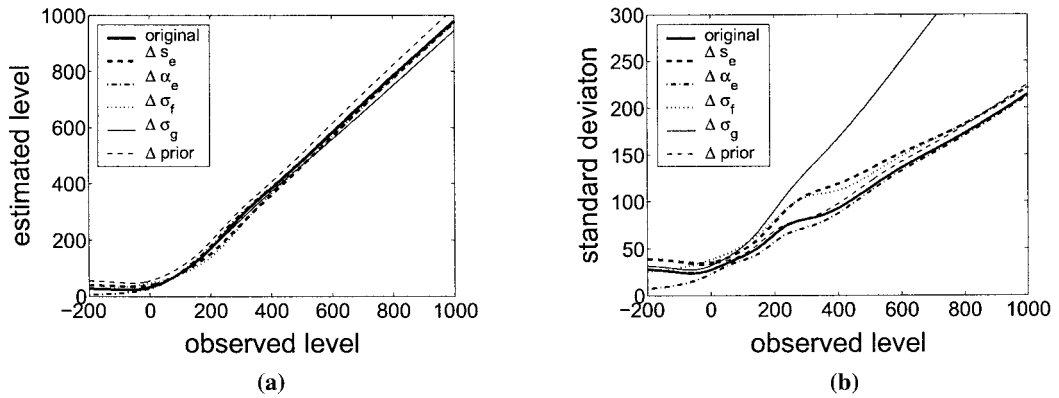
A factor of two criterion is most inaccurate at small positive values, just above the floor. With a floor of 25, chip readings of 35 and 80 would be considered significantly different by the factor-of-two rule. Yet, using our analysis, we would find that  $p = 0.26$ .

### 5.2. Effects of noise model components

The importance of including both additive and multiplicative measurement-specific noise in an error model for gene arrays is already established in the literature (Rocke and Durbin, 2001). The omission of the measurement-specific additive noise term ( $\epsilon$ ) leads to exaggerated ratio estimates, false identification of significant differences, and understated uncertainty measures when the observations are small. The omission of the multiplicative noise term ( $\gamma$ ) leads to similar problems when the observations are large.

The unique features of our noise model also play an important role in producing estimators that agree with experimental intuition and experience. The heavy-tailed additive noise term implies that as a measurement becomes increasingly negative, the associated estimate will remain above a finite floor; if that term were Gaussian, both the estimate and the associated uncertainty would approach zero. The heavy-tailed noise term will also produce more robust behavior when combining information from repeated experiments. The estimator will tend to ignore a single very small or negative measurement when the remaining measurements are all much larger, rather than favoring the small value.

In the absence of a gene-specific bias term, a sufficient number of experimental repetitions would produce an estimate with arbitrarily low uncertainty. The inclusion of the bias term captures the intuition that some of the measurement errors associated with gene arrays are systematic. Finally, our use of a prior that not only rules out negative transcript levels but also assigns a reduced probability to high transcript levels captures the intuition that true transcript levels tend to be larger than the corresponding measurements when the



**FIG. 12.** Effects of changing the noise and prior models on (a) transcript level estimates and (b) associated uncertainty measures from a single observation. The thick solid lines correspond to the model parameters of Section 2.

measurements are very small, and smaller than the corresponding measurements when the measurements are very large.

We found that the estimators and uncertainty measures described in Section 3 continue to display reasonable behavior as we vary the parameters of our noise and prior models. Figure 12 illustrates the effect of doubling each noise model parameter in turn on the estimated absolute transcript level and the associated uncertainty, computed according to Equations (7) and (8). The changes in  $s_e$ ,  $\sigma_f$ , and  $\sigma_g$  amount to quadrupling the variance of the additive noise, the bias, and the multiplicative noise, all of which have a relatively small effect on the estimated level. The uncertainty measure associated with the estimate depends primarily on the multiplicative noise at large observed levels and primarily on the additive noise at low observed levels. Figure 12 also shows the effect of replacing the prior model  $p_x(x)$  with  $xp_x(x)$ . The effect of this rather dramatic change in the prior is limited.

### 5.3. Application to datasets

The BEAM estimates of means, ratios, and  $p$ -values described in this work can be applied directly to normalized average difference measurements recorded from Affymetrix Ye6100 chips. We have computed lookup tables providing the BEAM estimates of means, ratios, variances, and  $p$ -values for measurements from one or two chips. These tables and software for automating the table lookups are available on-line at [www.ai.mit.edu/projects/BEAM](http://www.ai.mit.edu/projects/BEAM).

### 5.4. Potential extensions

Our noise and prior models were derived using data from Affymetrix yeast chips. Our validation study suggests that the unmodified model can describe U95 human chip data, as well. Still, noise and prior models specific to the system under study can only improve performance on that system. For example, Fig. 9b suggests that a prior model specific for human gene transcript levels should be weighted more heavily toward small values. One could also develop and use prior models specific to particular genes or to particular experimental conditions.

With appropriate noise and prior models, the BEAM framework could also be applied to other chip technologies, such as spotted arrays. One can develop a noise model for a particular technology using measurements of one or more transcripts present in the same constant amount in a large number of experiments, as we did here with the spiked controls.

Because our noise model was derived from data that were normalized using spiked controls, the resulting estimation lookup tables are most appropriate for normalized datasets. Unnormalized data will contain additional whole-chip multiplicative noise for which our model does not account. The resultant estimates will therefore be less precise. One could improve the results in the case where no spiked controls are available by deriving a noise model from unnormalized data or from data normalized using other methods.

Like most previous authors, we chose to work with the average difference measurement computed for each gene by the Affymetrix software, rather than using directly the 40 individual match and mismatch probe values for each gene on the Affymetrix chip. These lowest level data are typically not included in public data sets. Li and Wong (2001) have shown, however, that when the raw probe values are available, they can be used to model noise at the probe level, resulting in more accurate transcript measurements for each gene. For computational convenience, Li and Wong used an additive Gaussian noise model with a uniform prior distribution. A promising potential extension of our work would be the application of BEAM at the raw probe level. This would allow the incorporation of a more accurate noise model and of prior information on expression levels to the problem of estimating individual transcript levels addressed by Li and Wong.

The noise model developed in this paper represents measurement noise, the variation from array to array when the actual mRNA concentrations are identical. It includes variation introduced by the mRNA extraction, amplification, and labeling steps required before the sample is placed on the chip. Our noise model does not include intrinsic biological noise, the variation in mRNA concentrations between different cultures or animals grown under the same conditions. In order to judge whether the difference in expression levels observed between two populations is significant relative to inherent biological variation, one may incorporate a model of biological variation for a particular organism into BEAM.

## APPENDIX A. DATASETS

### A.1. Young lab yeast dataset

The data used in determining our noise and prior models were obtained from experiments performed by researchers in Richard Young's lab. These experiments used the Affymetrix Ye6100 chip sets, each set containing four chips comprising the 6,135 open reading frames in the yeast *Saccharomyces cerevisiae* genome. The experiments were all performed as previously described (Holstege *et al.*, 1998), by seven different researchers, between February 1999 and March 2000. The experiments used a number of different yeast strains under a variety of conditions. For a complete listing, see [www.web.wi.mit.edu/young/expression](http://www.web.wi.mit.edu/young/expression). The measurements in the database were the average difference values for each gene on each chip, as reported by version 4.0 of the Affymetrix Microarray Analysis Suite.<sup>3</sup>

Of the 271 chip sets with which we began, 10 were eliminated based on significant abnormalities in the data: some or all of the control probes on all four chips in these sets had observed transcript levels far outside the typical range. We used data from the remaining 261 chip sets in our analyses.

Each individual chip contained the same 15 spiked control probes. RNA corresponding to these probes was added at a known constant concentration during the mRNA purification stage of each experiment. The control probes consist of twenty 25-base-pair "perfect match" oligonucleotides contained in the 5' end, 3' end, and middle of five *Escherichia coli* genes, as well as twenty 25-base-pair "mismatch" oligonucleotides, which are identical to the perfect match oligonucleotides except for a one-base-pair mismatch. The *E. coli* genes used for controls are *trp*, *thr*, *lys*, *phe*, and *dap*. We considered the controls on each of the four chips in a set independently in this work, leading to  $261 \times 4 = 1,044$  measurements of each spiked control.

For some of our analyses, we were interested in the unspiked controls, i.e., control probes for which the corresponding mRNA was never added to the sample. Nine probes are included on each chip for use as hybridization controls; that is, the corresponding mRNA is spiked into the mix which is hybridized to the chip before scanning. These controls consist of 20 25-base-pair "perfect match" and 20 "mismatch" oligonucleotides for *E. coli* genes *bioB*, *bioC*, *bioDn*, and *cre*. The oligonucleotides for *bioB* were contained in the 5' end, 3' end, and middle of the gene, while the others consisted of only oligonucleotides for the 5' end and 3' end. The mRNA for these controls was spiked into only a subset of the experiments in

---

<sup>3</sup>The recently released version 5.0 of this software computes transcript level measurements from the probe-level data using a different statistical algorithm that no longer reports negative measurements. Although this new software does not represent a change in the actual chip technology, Affymetrix reports that it reduces the noise associated with the measurements (Affymetrix Inc., 2001).

the data set (32/261), leaving 229 chip sets where these probes were unused. Again, we considered the controls on each chip independently, leading to  $229 \times 4 = 916$  measurements of each unspiked control which were used for analysis of additive noise.

### A.2. Affymetrix Latin square human data set

We verified our noise model using the Affymetrix human Latin square data set, available at [www.affymetrix.com/analysis/download\\_center2.affx](http://www.affymetrix.com/analysis/download_center2.affx). This data set consists of 59 U95A human chips.

We computed average difference measures for each gene using Affymetrix's Microarray Analysis Suite, version 4.0. We then normalized the Latin square data set as we did the Young lab data set, using the spiked controls with the method of Hartemink *et al.* (2001). The only other manipulation we performed on the Latin square data was to scale the entire data set by a constant factor (4.79). This factor made the mean of the entire Latin square data set equal to the mean of the Young lab data set, on which we fit our noise model.

## APPENDIX B. CONSTRAINED DECONVOLUTION PROCEDURE

We posed the deconvolutions that arose in deriving our noise model as quadratic programming problems so that we could enforce linear constraints on the solutions. Given functions  $g$  and  $h$ , suppose we wish to find  $f$  such that  $f * g = h$ . Assuming that these are all functions of a discrete variable  $k$ , we can expand the convolution operation as

$$f * g = \sum_{j=-\infty}^{\infty} f(j)g(k-j).$$

To perform the deconvolution, we find the  $f$  which minimizes the sum-squared difference between  $f * g$  and  $h$ , namely

$$\hat{f} = \min_f \sum_{k=-\infty}^{\infty} \left[ \left( \sum_{j=-\infty}^{\infty} f(j)g(k-j) - h(k) \right)^2 + \lambda (f(k) - f(k-1))^2 \right],$$

where the minimization is over  $f(k)$  for all  $k$ . The regularization term  $\lambda (f(k) - f(k-1))^2$  penalizes the derivative of  $f$ . For each deconvolution, we chose the smallest  $\lambda$  which produces a reasonably smooth solution. If  $f$  represents the multiplicative noise distribution or the prior distribution, then we perform the minimization subject to the constraints

$$\begin{aligned} f(k) &= 0 \text{ when } k < 0, \\ f(k) &\geq 0 \text{ when } k \geq 0. \end{aligned}$$

Because this problem involves minimizing a quadratic function of the  $f(k)$  under linear constraints, we can solve it using a standard quadratic programming package.

One weakness of our deconvolution technique is the lack of an automatic method for selecting  $\lambda$ . This may be addressed through further investigation, although we have found that the choice of  $\lambda$  does not have a strong effect on the BEAM estimates derived using our noise model.

## ACKNOWLEDGMENTS

We would like to thank Alexander Hartemink and David Gifford for helpful discussions and advice, Christine Ladd at the MIT/Whitehead Genome Center for aid with the Affymetrix Microarray Analysis Suite, Yoseph Barash for suggesting validation data sets, Erik Marchese of Affymetrix for clarifying the

experimental procedure used to produce the Latin square data set, and Laura Baldwin for help in manuscript preparation. Several anonymous reviewers provided helpful comments that led to substantial improvement of the paper.

## REFERENCES

- Affymetrix Inc. 2001. New statistical algorithms for monitoring gene expression on GeneChip probe arrays. [www.affymetrix.com/support/technical/technotes/statistical\\_algorithms\\_technote.pdf](http://www.affymetrix.com/support/technical/technotes/statistical_algorithms_technote.pdf).
- Baggerly, K.A., Coombes, K.R., Hess, K.R., Stivers, D.N., Abruzzo, L.V., and Zhang, W. 2001. Identifying differentially expressed genes in cDNA microarray experiments. *J. Comp. Biol.* 8, 639–659.
- Baldi, P., and Long, A.D. 2001. A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17, 509–519.
- Brown, P.O., and Botstein, D. 1999. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* 21, 33–37.
- Chen, Y., Dougherty, E.R., and Bittner, M.L. 1997. Ratio-based decision and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* 2, 364–374.
- Hartemink, A.J., Gifford, D.K., Jaakkola, T.S., and Young, R.A. 2001. Maximum likelihood estimation of optimal scaling factors for expression array normalizations. *SPIE Bios.* 2001.
- Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95, 717–728.
- Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburty, K., Simon, J., Bard, M., and Friend, S.H. 2000. Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126.
- Ideker, T., Thorsson, V., Siegel, A.F., and Hood, L.E. 2000. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comp. Biol.* 7, 805–817.
- Jelinsky, S.A., and Samson, L.D. 1999. Global response of *Saccharomyces cerevisiae* to an alkylating agent. *Proc. Natl. Acad. Sci. USA* 96, 1486–1491.
- Kerr, M.K., and Churchill, G.A. 2000. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. USA* 98, 8961–8965.
- Kerr, M.K., Martin, M., and Churchill, G.A. 2000. Analysis of variance for gene expression microarray data. *J. Comp. Biol.* 7, 819–837.
- Lee, M.L., Kuo, F.C., Whitmore, G.A., and Sklar, J. 2000. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA* 97, 9834–9839.
- Li, C., and Wong, W.H. 2001. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* 98, 31–36.
- Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E.L. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* 14, 1675–1680.
- Long, A.D., Mangalam, H.J., Chan, B.Y.P., Toller, L., Hatfield, G.W., and Baldi, P. 2001. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. *J. Biol. Chem.* 276, 19937–19944.
- Mills, J.C., and Gordon, J.I. 2001. A new approach for filtering noise from high-density oligonucleotide microarray datasets. *Nucl. Acids Res.*, 29, e72.
- Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W. 2001. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comp. Biol.* 8, 37–52.
- Rocke, D.M., and Durbin, B. 2001. A model for measurement error for gene expression arrays. *J. Comp. Biol.* 8, 557–569.
- Schadt, E.E., Li, C., Su, C., and Wong, W.H. 2000. Analyzing high-density oligonucleotide gene expression array data. *J. Cellular Biochem.* 80, 192–202.
- Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzog, H. 2000. Normalization strategies for cDNA microarrays. *Nucl. Acids Res.* 28, e47.
- Tanaka, T.S., Jaradat, S.A., Lim, M.K., Kargul, G.J., Wang, X., Grahovac, M.J., Pantano, S., Sano, Y., Piao, Y., Nagaraja, R., Doi, H., Wood, W.H., Becker, K.G., and Ko, M.S. 2000. Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray. *Proc. Natl. Acad. Sci. USA* 97, 9127–9132.

- Theilhaber, J., Bushnell, S., Jackson, A., and Fuchs, R. 2001. Bayesian estimation of fold-changes in the analysis of gene expression: The PFOLD algorithm. *J. Comp. Biol.* 8, 585–614.
- Wodicka, L., Dong, H., Mittmann, M., Ho, M.H., and Lockhart, D.J. 1997. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnol.* 15, 1359–1367.

Address correspondence to:

*Ron O. Dror*  
*D.E. Shaw Research and Development*  
*120 W. 45th Street*  
*New York, NY 10036*

*E-mail: rondror@ai.mit.edu*