Bioinformatics Advance Access published March 7, 2006

Rosetta Error Model for Gene Expression Analysis

Lee Weng*, Hongyue Dai, Yihui Zhan, Yudong He, Sergey B. Stepaniants, and Douglas E. Bassett, Jr.

Rosetta Inpharmatics LLC
401 Terry Avenue North
Seattle, WA 98109, USA

Associate Editor: John Quackenbush

* To whom correspondence should be addressed.

ABSTRACT

Motivation: In microarray gene expression studies, the number of replicated microarrays is usually small because of cost and sample availability, resulting in unreliable variance estimation and thus unreliable statistical hypothesis tests. The unreliable variance estimation is further complicated by the fact that the technology-specific variance is intrinsically intensity-dependent.

Results: The Rosetta error model captures the variance-intensity relationship for various types of microarray technologies, such as single-color arrays and two-color arrays. This error model conservatively estimates intensity error and uses this value to stabilize the variance estimation.

We present two commonly used error models: the intensity error-model for single-color microarrays and the ratio error-model for two-color microarrays or ratios built from two single-color arrays. We present examples to demonstrate the strength of our error-models in improving statistical power of microarray data analysis, particularly, in increasing expression detection sensitivity and specificity when the number of replicates is limited.

Availability: Rosetta error models are available in the Rosetta Resolver® system and Rosetta Luminator™ system for gene-expression analysis. These technology-specific error models are designed and optimized for different microarray technologies, such as Affymetrix® and Agilent Technologies.

Contact: lee_weng@rosettabio.com

1 INTRODUCTION

DNA microarrays are widely used to study gene expressions (Hughes, et al., 2000). Fluorescent intensities of hybridizations from microarray spots measured by optical scanners provide indirect measurements of messenger-RNA abundances in biological samples of interest. It has been demonstrated that the hybridization intensity is approximately proportional to the RNA abundance (Lockhart, et al., 1996). Both single-color and two-color microarrays are commonly used in hybridization experiments. A single-color microarray, such as that from Affymetrix, Inc. provides intensity measurements from one hybridized biological sample. A two-color microarray, such as that from Agilent Technologies, Inc. measures expression ratios between two hybridized samples that are labeled with two different fluorescent dyes, such as Cy5 and Cy3.

The primary application of microarrays is to study changes in gene expression under different conditions, such as different phenotypes or different treatment perturbations. Microarray measurements are subject to many sources of variation, ranging from array-lot variability to washing conditions, and many different errors that can affect the measurement results. To improve the measurement precision, we typically hybridize several replicated microarrays in each condition group. Then we apply statistical hypothesis tests to analyze the change in measured intensities. When comparing two conditions, we may apply a *t*-test or Wilcoxon test. When comparing more conditions in a more complicated factorial design, we may use an ANOVA test. In these statistical tests, the underlying null-hypothesis is that intensity measurement has no change. A *p*-value is usually the result of the test, which indicates the probability of observing a discrepancy as large as, or larger than, the given observation under the null. When the *p*-value computed from a microarray measurement for a particular gene (or RNA sequence in general) is small, e.g., less than 0.01, we can reject the null and accept the alternative hypothesis that the gene expression is different in one or more conditions. One caution in interpreting test results is that intensity changes may not necessarily imply gene expression change. Because we cannot directly measure gene expression, the indirect measurement using hybridization intensities can have changes caused by factors other than gene expression. Proper experiment designs can help reduce the biases in the result of hypothesis tests; however, we are detecting intensity changes instead of expression changes in microarray measurements.

1.1 Variance Estimation

Most hypothesis tests explicitly or implicitly estimate the variance within the same condition group. For example, in a *t*-test, within-group variance comes from the standard deviation estimation. In an ANOVA test, variance is the within-group mean sum-of-squares. If the between-group difference (in a *t*-test) or between-group variance (in an ANOVA) is significantly larger than the within-group variance, we reject the null hypothesis. A reliable estimation of the within-group variance is critical in hypothesis tests.

Unfortunately, reliable variance estimations are not always available in microarray gene-expression studies. Limited by experimental material costs and biological sample costs,

we often cannot obtain a sufficient number of replicates in each group. Commonly, two or three replicates per group are all what we can expect. In a *t*-test example of two replicates of expression ratios, we only have 2-1=1 degree-of-freedom for the within-group variance estimation, and the result is unreliable. To improve reliability, some permutation-based methods have been developed to stabilize the variance estimation (Tusher, et al., 2001). However, when the number of replicates is very small (two or three, for example), permutation methods do not work properly either.

1.2 Errors in Microarray Measurements

We can categorize measurement errors as systematic and random. Systematic errors bias the measurements in a direction we may be able to approximately estimate. If we estimate the size and the direction of the bias, we can correct or reduce the systematic error. Some systematic errors common to microarray measurements include non-zero background intensity levels, differences between two labeling dyes, positional bias due to array production process, hybridization, or scanning, etc. Many data preprocessing methods, such as background-subtraction, normalization, detrending, and fluorescent-reversal combination have been developed to reduce the impact of the systematic errors on gene-expression analyses (Schadt, et al., 2002; Yang, et al., 2002; Quackenbush, 2002). Random errors are measurement fluctuations left after the systematic error correction. The exact value and direction of the random fluctuation is not predictable, but the variance of the random error may follow certain rules. Error models are built to capture the predictable behavior of the variance. Random error in microarray measurement is the focus of this paper.

The intensity variance of microarray measurement is intensity dependent. This phenomenon has been discussed in the literature (Tu, et al., 2002). Many different error models have also been developed to describe the microarray measurement variance (Chen, et al., 1997; Li and Wong 2001). Although error models may have different forms, the observation is that the absolute intensity-variances tend to be larger in higher intensities. In general, there are two types of approaches in modeling the intensity-dependent measurement error. One approach is to build a regression curve to fit the intensity versus variance relationship (Jain, et al., 2003). The form of the regression equation, for example, LOWESS regression, is purely data-driven. The other approach (the explicit approach) is to model the errors from various sources or types, such as additive and multiplicative errors (Rocke and Durbin, 2001; Theilhaber, et al., 2001; Dror, et al., 2003). The Rosetta error model uses the latter approach. There are several advantages of the explicit approach. It is based on understanding the actual cause of error, so that it is less susceptible to over-fitting, such as the variance underestimation problem caused by intensity saturations.

2 SYSTEM AND METHODS

We developed the Rosetta error model to improve variance estimation using a small number of replicates (Stoughton and Dai, 2002). Specifically, we explicitly model main

error sources in microarray measurements and then apply the model-predicted measurement error as the error floor to help stabilize the variance estimation. This approach has been leveraged across multiple array technologies and sample types to improve the specificity and the sensitivity in differential gene-expression detections (Agy, et al., 2003; Falls, et al., 2003; Geiss, et al., 2002; Geiss, et al., 2003; Helfrich, et al., 2003; Hori, et al., 2003; Liu, et al., 2002; Marini, et al., 2003; Richards, et al., 2003; Schirra, et al., 2002; Smith, et al., 2003; Thimmapaya, et al., 2003; Tonouchi, 2002; Wout, et al., 2003).

2.1 Intensity Error Model

In general, the hybridization intensity measurement I is a function of the RNA transcript abundance θ in the sample plus errors:

$$I(i, j) = f_j(\theta_i) + \varepsilon(i, j) \quad (1)$$

where i is the index of the microarray to which the sample is hybridized, j is the targeted sequence (the spot) index in the microarray, and ε is the random measurement error. Here we assume systematic errors have been removed during data preprocessing using available methods (Yang, et al., 2002; Quackenbush, 2002). Often, a linear relationship between the abundance and the intensity is assumed:

$$f_j(\theta_i) \approx \alpha_j \cdot \theta_i \quad (2)$$

where α_j is the binding efficiency of the given sequence probe. This simplified relationship may not be valid at high abundance. In the Rosetta error model method described in this paper, we do not model this intensity-abundance relationship and do not make an assumption of the linearity, which differs from some published intensity modeling approaches (Li and Wong, 2001). Here we are only interested in modeling the intensity error ε.

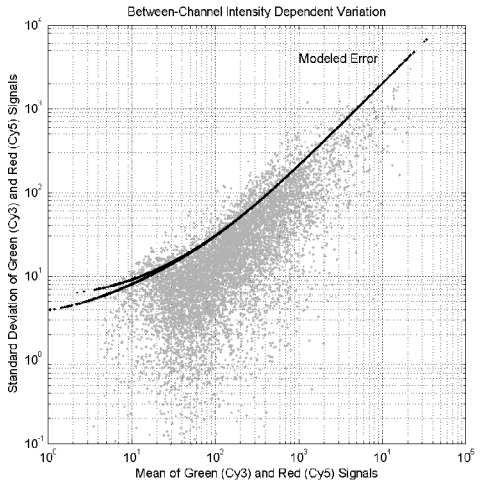


Fig.1. Demonstration of intensity-dependent measurement variations in microarrays. The data come from a same-versus-same two-color microarray where the same RNA sample is hybridized in both the red (Cy5) and the green (Cy3) channels. The horizontal axis is the mean of two measurements from the two channels. The vertical axis is the standard deviation of the two measurements. Each microarray spot is shown as one gray dot in the figure. For comparison, two overlapping black lines are modeled intensity errors in the red and the green channels computed from Equation (7), where POISSON=5 and FRACTION=0.2. The two lines become separated at low intensities because the red and the green channels have different levels of additive noise. In this example, the additive noise in the green channel is higher than in the red channel.

In replicated microarray experiments, we observe that the standard deviation of the repeated intensity measurement increases with the mean intensity (Fig. 1). We can model the intensity-dependent measurement error by breaking it into various error sources based on their different statistical characteristics and different causes:

$$\varepsilon(i,j) = \varepsilon_{add}(i,j) + \varepsilon_{Poisson}(i,j) + \varepsilon_{frac}(i,j) + \varepsilon_{spot}(i,j). \quad (4)$$

The first term is the additive error, which is independent from the specific binding intensity. It may include two components, background and cross-hybridization (non-specific binding):

$$\varepsilon_{add}(i,j) = \varepsilon_{bkg}(i,j) + \varepsilon_{xhyb}(i,j). \quad (5)$$

The background noise may come from array substrate contamination, fluorescent residuals on the microarray after washing, and electronic noise in the scanner amplifier. The background noise mean and variance are denoted by:

$$E\left(\varepsilon_{bkg}(i,j)\right) = \mu_{bkg}(i,j)$$
$$Var\left(\varepsilon_{bkg}(i,j)\right) = \sigma^2_{bkg}(i,j).$$

The cross-hybridization noise has its mean and variance as well:

$$E\left(\varepsilon_{xhyb}(i,j)\right) = \mu_{xhyb}(i,j) \approx \beta_j \cdot \phi_i$$
$$Var\left(\varepsilon_{xhyb}(i,j)\right) = \sigma^2_{xhyb}(i,j)$$

where β_j is the overall non-specific binding efficiency of the probe and ϕ_i is the overall non-specific hybridization concentration. Because it is often difficult to separately estimate the background noise and the cross-hybridization noise, we estimate them together as the total additive noise:

$$E\left(\varepsilon_{add}(i,j)\right) = \mu_{add}(i,j) \approx \mu_{bkg}(i,j) + \beta_j \cdot \phi_i$$
$$Var\left(\varepsilon_{add}(i,j)\right) = \sigma^2_{add}(i,j) = \sigma^2_{bkg}(i,j) + \sigma^2_{xhyb}(i,j).$$

For Affymetrix® GeneChip® arrays, mismatch probes provide information about the total additive noise including the cross-hybridization noise. The mean of the additive noise is estimated and removed during data preprocessing (Hubbell, et al., 2002; Liu, et al., 2002). We need only focus on the variance of the noise. For other microarrays that do not have mismatch probes, we usually can estimate only the background noise but not the cross-hybridization noise. The mean of cross-hybridization noise often cannot be removed during background subtraction in data preprocessing. The resulting expression signal intensity in Equation (8) is the combination of both specific and cross hybridizations.

The second term in Equation (4) is the Poisson error. A Poisson process has its standard deviation proportional to the square root of its mean, and it is modeled as:

$$E\left(\varepsilon_{Poisson}(i,j)\right) = 0$$
$$Var\left(\varepsilon_{Poisson}(i,j)\right) = POISSON \cdot I(i,j)$$

where *POISSON* is a fixed parameter in the current model for a given microarray technology. The Poisson error represents the randomness of the hybridization binding process, which is a stochastic Poisson process. The existence of Poisson noise in microarray measurements is reported in other publications as well (Tu, et al., 2002).

The third term in Equation (4) is the fractional (multiplicative) error. Fractional error is used to model the linear relationship at high intensities observed in microarray experiments, such as the one illustrated in Fig. 1. It has its mean and variance as

$$E\left(\varepsilon_{frac}(i,j)\right) = 0$$
$$Var\left(\varepsilon_{frac}(i,j)\right) = \left(FRACTION \cdot I(i,j)\right)^2$$

where *FRACTION* is a fixed parameter in the current model for a given microarray technology. The standard deviation of the fractional error increases linearly with an increase in measured intensities. There are many factors that cause fractional errors in microarray measurements. The variation of array spotting or probe synthesis introduces multiplicative error in intensity measurements. Scanned images can suffer from laser speckle noise that has a Raleigh distribution, of which the standard deviation is proportional to the mean.

The last term ε_{spot} in Equation (4) is called the spot defect noise. It describes spot imperfections caused by dust, physical damage, and contaminations. It is array and spot specific, and it is usually not predictable. Most microarray feature extraction software provides QC flags that identify those spots failed in image analysis. However, spots that have passed QC are not necessarily of equal quality. The error model should consider the quality differences among different spots in different arrays. However, this term may be less relevant to the two-color microarrays since the artifacts are usually shared by both

channels for a given spot. Their effects are usually canceled when computing expression ratios.

When data are analyzed from a probe sequence deposited multiple times on an array, the variation of these within-array replicates only comes from the intra-array error. When data from identical probe sequences are analyzed across multiple arrays hybridized with the same sample, the measurement suffers from both intra- and inter-array (or inter-channel) variations. We may decompose each of the first three terms in Equation (4) into its intra-array/channel and inter-array/channel components. For example, the additive error can be written as

$$\varepsilon_{add}(i,j) = \varepsilon_{\text{int}\,er\_add}(i) + \varepsilon_{\text{int}\,ra\_add}(j) \ . \qquad (6)$$

For intensities from two-color microarrays, there is an additional dye-related bias. The dye-bias is sequence dependent. It is usually difficult to predict. Rather than modeling it explicitly in Equation (4), we rely on the fluorescent-reversal (dye-swapping) method in experimental design to cancel out the dye-bias (Stoughton and Dai, 2002).

2.1.1 Model Development and Parameter Estimation

Our goal in error modeling is to estimate conservatively the measurement error for a particular microarray technology. We prefer a conservative error model to keep the false-positive detection rate low. Depending on available measurement information from the feature extraction software, we typically construct two error estimates and then pick up the larger one as final intensity error. These two error estimates are the modeled error and the measured error.

2.1.1.1 Modeled Error

The variance of the modeled error is estimated as

$$\sigma^2_{\text{mod}\,eled}(i,j) = \sigma^2_{add}(i,j) + POISSON \cdot I(i,j) + FRACTION^2 \cdot I^2(i,j) \ . \qquad (7)$$

The variance can be viewed as the Taylor-series expansion of the intensity-dependent variance. The technology-dependent parameters *POISSON* and *FRACTION* are estimated for a given microarray technology type (array, labeling/amplification, hybridization protocol, scanning equipment, and feature extraction software) during error model development. They are fixed as long as the technology remains unchanged. Poisson and fraction noises can be slightly different in different microarray technologies, but they are usually stable over time for a given technology. The additive noise can change significantly from one array to another so that the additive component in Equation (7) is estimated on an array-by-array basis.

There are many possible methods to model the additive component. Most feature extraction software provides some background estimations based on pixels that surround feature spots. For genome arrays, we may use features in the lowest percentiles of intensity distribution to estimate the background, and these estimates are carried out for a

region on the array or the entire array. Negative control features on the array can also provide information about the background and non-specific hybridization information. Based on our observations, background measured from surrounding pixels is usually not very reliable. It often has weak correlation with intensity inside the feature when the sequence of the feature has no expression. Regardless of what background measurement method we use, for a given feature, it is more reliable to use the averaged information from a surrounding region much larger than one spot to model the additive term in Equation (7).

When developing an error model, we estimate the value for parameters *POISSON* and *FRACTION* in Equation (7). For different microarrays, *POISSON* is typically in the range of 0-20 for data from a 16-bit scanner, and *FRACTION* is in the range of 0.1-0.25. We use the same-versus-same technical replicates to estimate proper values of these parameters. These replicated arrays come from the same RNA sample, and are processed (inverse-transcription, amplification, labeling, hybridization, and so forth) separately, so the data contain most of the technical variations in the microarray data acquisition process. One example is shown in Fig. 1. For two replicated intensity profiles, the x-axis plots their mean and the y-axis plots their standard deviation. Each gray dot represents a feature spot. The modeled error, the square-root of Equation (7), is plotted as a black line. We establish these two parameters in a technology-specific fashion so that the modeled error curve fits to the top of the standard deviation distribution. Details of error model parameter estimation are provided in Appendix-B. Using this method, our modeled error is conservative and includes both intra-array/channel and inter-array/channel variations. *FRACTION* controls the goodness-of-fit at the high-intensity range (the right side of the plot). *POISSON* controls the goodness-of-fit at the mid-level intensity range. After we find the proper value of these two parameters for a given microarray technology, we fix *POISSON* and *FRACTION* as constants in later applications. Table 1 lists some examples of the Rosetta error model parameters for several commonly used microarrays.

Table 1. Examples of the Rosetta error model parameters for some commonly used microarray technologies

Microarray Technology	***FRACTION***	***POISSON***	***RANDOM***
Affymetrix® GeneChip® 18-micron arrays (e.g., HG-U133A)	0.1	5	0.35
Affymetrix GeneChip 20-micron arrays (e.g., HG-U95A)	0.1	10	0.35
Affymetrix GeneChip 24-micron arrays (e.g., RG-U34A)	0.15	20	0.3
Agilent SurePrint® microarrays	0.12	3	0.05
GE Healthcare (formerly Amersham Biosciences) CodeLink arrays	0.18	5	0.05

The modeled error in Equation (7) has important applications. It makes variance standardization possible when deriving the parameter *xdev* [see (13)] for differential

expression *p*-value calculations. The intensity transformation method for intensity variance stabilization can also be derived from the error model (Weng, 2003).

2.1.1.2 Measured Error

Equation (7) does not include the spot defect term in Equation (4). Many feature extraction software applications provide summary statistics, such as standard deviations, of pixel variations within and surrounding each spot. Many spot defects, such as dust and scratches, can significantly increase the pixel standard deviations. Because we usually subtract the background from the signal intensity

$$I(i,j) = I_{signal}(i,j) - I_{bkg}(i,j)\,, \tag{8}$$

we can model the variance of the intra-array measured error as

$$\begin{aligned}\sigma^2_{measured}(i,j) &= \sigma^2_{measured\_bkg}(i,j) + \sigma^2_{measured\_signal}(i,j) \\ &= \frac{\sigma^2_{bkg\_pixel}(i,j)}{RANDOM \cdot N_{bkg}(i,j)} + \frac{\sigma^2_{signal\_pixel}(i,j)}{RANDOM \cdot N_{signal}(i,j)}\end{aligned} \tag{9}$$

where σ_{bkg_pixel} is the background pixel standard deviation, σ_{signal_pixel} is the signal pixel standard deviation, N_{bkg} is the number of pixels used in the mean background calculation, and N_{signal} is the number of pixels used in the mean signal calculation. Pixel standard deviations are provided from microarray feature extraction software. Here we want to convert the pixel standard deviation to the standard error of the spot mean. Limited by the scanner resolution and array spot non-uniformity, the pixel intensity measurements are usually highly correlated. When we convert the pixel standard deviation to the standard error of the spot (mean pixel intensity of multi-pixels), we need to discount the redundancy in pixel measurements. In Equation (9) *RANDOM* is an error model parameter that defines the statistically independent fraction of the number of pixel measurements. The *RANDOM* parameter has the interpretation that it is the fraction of pixels that is truly random. The typical range of *RANDOM* is about 0.05-0.35. During error model development, we usually need to adjust this parameter first. When the additive error in Equation (7) is estimated from the standard errors of the background pixels, this parameter *RANDOM* controls the error model fit at low intensities (the left side of the plot in Fig. 1). To determine the proper *RANDOM*, we examine the error model fit with data of technical replicates. More details of the parameter estimation are discussed in Appendix-B. After we find the proper value of this parameter for a given microarray technology during error model development, we fix *RANDOM* as a constant in later applications.

Pixel standard deviations also carry information about intra-array/channel additive, Poisson, and fractional variances. Measured errors are composed of

$$\sigma^2_{measured}(i,j) = \sigma^2_{\text{int } ra\_add}(j) + \sigma^2_{\text{int } ra\_Poisson}(j) + \sigma^2_{\text{int } ra\_frac}(j) + \sigma^2_{spot}(i,j)\,. \tag{10}$$

Fig. 2 shows the comparison between the modeled error as a black line and the measured error in gray dots. Because measured error includes only intra-array/channel variations and modeled error includes both intra- and inter-array/channel variations, the measured error is lower than the modeled error for most features. But for some small number of features, the measured error can be significantly higher than the modeled error. Those are the spots that have large spot-defect errors defined in the last term of Equation (4) and Equation (10). This term is not covered in the modeled error in Equation (7).

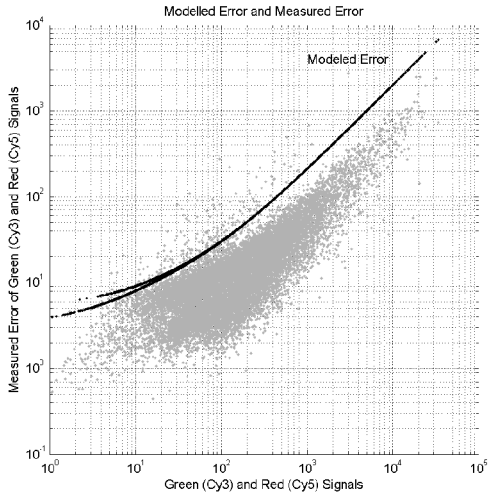


Fig.2. Comparison between the modeled error and the measured error (estimated from the pixel fluctuation as in Equation (9)). The data are the same as those in Fig. 1. The horizontal axis is the intensity measurements from the red and the green channel. The vertical axis is the measured error from Equation (9) of the red and the green channel. Each microarray spot is shown as two gray dots in the figure, one from the red channel and the other from the green channel. For comparison, the two overlapping black lines are modeled errors in the red and the green channels computed from Equation (7).

When combining the modeled error and the intra-array measured error together to cover all terms defined in the error model in Equation (4), we conservatively select the larger one as the final error estimation of intensity *I(i, j)*:

$$\sigma_I(i,j) = \max\left(\sigma_{\mathrm{mod}\,eled}(i,j), \sigma_{measured}(i,j)\right) \tag{11}$$

2.2 Ratio Error Model

One of the most common microarray applications is to compare gene expression in two different conditions. For two-color arrays, samples of these two conditions can be labeled

with two different fluorescent dyes, then mixed and hybridized to one array. The two fluorescent measurements (red and green) provide two intensities for comparison. For single-color arrays, these two samples are hybridized independently to two arrays. Then two separate intensity measurements are compared. Here we name the baseline intensity measurement as $I_1(i,j)$ and the experimental (perturbed) intensity as $I_2(i,j)$. Their estimated intensity errors are $\sigma_1(i,j)$ and $\sigma_2(i,j)$. Log-ratio of these two intensities is commonly used to measure their differences. We often use 2 or 10 based logarithm in computing the log ratio. We define the log-ratio as the 10-based logarithm of I_2 divided by I_1 in this paper:

$$lratio(i,j) = \log_{10}\left(\frac{I_2(i,j)}{I_1(i,j)}\right) \tag{12}$$

Often, one or both intensities can be zero or negative after background subtraction. To avoid the difficulty in computing log ratios with zero or negative intensities, we may set a positive floor to the intensity values in (12). Because the lowest positive scanner reading is one, we often set intensities below one equal to one before the log-ratio computation.

In differential expression analysis, besides the log-ratio, we are also interested in knowing the error of the log-ratio propagated from intensity error. The log-ratio error estimation is particularly important in the *p*-value estimation for differential expression calls. We define a new parameter *xdev* as

$$xdev(i,j) = \frac{I_2(i,j) - I_1(i,j)}{\sqrt{\sigma_2^2(i,j) + \sigma_1^2(i,j)}}. \tag{13}$$

Xdev is the intensity difference divided by the error of the difference. It standardizes the variance of the intensity difference. At the null condition (same-versus-same experiment), the parameter *xdev* has its distribution very close to normal N(0,1) distribution (Stoughton and Dai, 2002). This property simplifies the *p*-value calculation in differential expression analysis. When the *xdev* is small, the log-ratio error can be approximated as

$$\sigma_{lratio}(i,j) \approx \frac{lratio(i,j)}{xdev(i,j)}. \tag{14}$$

It can be proved (see Appendix-A) that Equation (14) is the first-order approximation of Equation (15) in a Taylor expansion. Equation (15) is the approximation of log-ratio error when intensity errors are relatively small:

$$\sigma_{lratio}(i,j) \approx \log_{10}(e) \cdot \sqrt{\frac{\sigma_1^2(i,j)}{I_1^2(i,j)} + \frac{\sigma_2^2(i,j)}{I_2^2(i,j)}}. \tag{15}$$

3 ERROR MODEL APPLICATIONS

The estimated error can be used in many applications together with measurements themselves to improve the analysis result. The possibility of leveraging the available error estimation from the error model is broad. Here we provide only a few examples.

3.1 Error-weighted Replication Combining

In microarray experiments, several replicated hybridizations for one treatment condition are usually generated to improve measurement precision and accuracy. We need to combine them to one averaged result. We prefer the averaged result that has the smallest error possible (the minimum variance estimates). The solution is error-weighted averaging, where the weighting factor is inversely proportional to the variance of the measurement (Stoughton and Dai, 2002). Assuming the measurement is $x(i)$ and the measurement error is $\sigma_x(i)$, and there are N replicates and $i=1:N$, we compute the error-weighted average as:

$$\bar{x} = \frac{\sum_i w(i) \cdot x(i)}{\sum_i w(i)}, \text{ where } w(i) = \frac{1}{\sigma_x^2(i)}. \tag{16}$$

Error-weighted averaging requires that the measurement error is not a monotonic function of measurement itself to avoid possible biases by weighting. In general, the log-ratio and log-ratio error defined in (12) and (15) meet this requirement. But intensity and intensity error in (8) and (11) do not, because intensity error is a function of intensity shown in Section 3. To apply error-weighted averaging to intensity data, we should transform the intensity to a new domain where the transformed intensity error is not related to the transformed intensity. We use an error-model based intensity transformation method (Weng, 2003). Some other published variance stabilization methods (Holder, et al., 2001; Durbin, et al.) can also be used.

The error of the averaged measurement in (16) is usually smaller than individual error $\sigma_x(i)$. There are two possible estimations of error in this averaged measurement (Stoughton and Dai, 2002). The first is called the propagated error, which is the population error of $\bar{x}$ derived from the individual measurement errors $\sigma^2{}_x(i)$ by invoking the formula for $Var(\bar{x})$:

$$\sigma_{\bar{x}p}{}^2 = \frac{1}{\sum_i \frac{1}{\sigma_x^2(i)}} = \frac{1}{\sum_i w(i)}. \tag{17}$$

The second is called the scattered error, which is the empirical standard deviation of $\bar{x}$ computed from the individual observations $x(i)$ using a weighted scheme:

$$\sigma_{\bar{x}s}^{\ 2} = \frac{1}{(N-1)\cdot\sum_i w(i)} \cdot \sum_i w(i)\cdot\left(x(i)-\bar{x}\right)^2 \ . \qquad (18)$$

The propagated error depends only on the error estimation from the error model. It is always available even when N=1. However, the propagated error can be biased. The scattered error is an unbiased estimation, but when N is small the scattered error estimation is very unstable and the error estimation itself can have a very large error.

There is another important difference between the propagated error and the scattered error. Different genes often behave differently in their expressions. Some genes are "quiet" and their expression levels show little change. Some other genes are "jumpy" and their expression levels have relatively large variation even in same-versus-same experiments. The Rosetta error model in (7) is not gene-specific; therefore, the propagated error in (17) is not gene-specific. However, the scattered error in (18) is gene-specific. When enough replicates are available, the quiet genes tend to have smaller scattered-error than the jumpy genes. Furthermore, if the repeated experiments are biological replicates, i.e., each individual measurement comes from an individual test subject (e.g., animal) under the same treatment, the scattered error carries information about both the technical measurement variation due to microarrays and the biological variation due to test subjects. But the propagated error only categorically carries information about the technical measurement variation (including the inter-array variations coming from the error model).

To obtain reliable error estimation for the weighted-average in (16), we combine the propagated error and the scattered error to balance their strength and weakness. When the number of replicates N is small, the scattered error is very unreliable and has large variation in itself. In this case, we want to use the propagated error to help stabilize the error estimation, in other words, to reduce the variation in the error estimation. The propagated error sets a lower bound or floor to prevent the error estimation from being smaller than the microarray technology can support. When N is large, we can trust the scattered error more because it is unbiased and also includes biological variations. One method, proposed by Stoughton and Dai, 2002, provides an optimal solution to combine these two error estimations. The combined error estimate of the weighted-average is:

$$\sigma_{\bar{x}} = \frac{\sigma_{\bar{x}p} + (N-1)\cdot\sigma_{\bar{x}s}}{N}. \qquad (19)$$

When N=1, the propagated error is the only estimate we have. When N increases, the scattered error estimate increasingly dominates the final error estimate and the contribution from the propagated error decreases, which is expected. This combination method has been verified extensively for the last a few years in many different microarray technologies. The results are satisfactory.

When applying error-weighted averaging to combine replicates, we penalize measurements with large estimated errors. Measurements that have much larger

measurement errors than most others, contribute minimally to the averaged results. The number of replicates contributing to average computation is effectively less than the total number of replicates available. The effective number of replicates can be computed as:

$$eN = \left(\frac{\min(\sigma_x(i))}{\sigma_{\bar{x}p}} \right)^2 = (\min(\sigma_x(i)))^2 \cdot \sum_i w(i) .$$

The number of replicates N in Equation (18) and (19) should be replaced with the effective number of replicates eN.

3.2 Present Expression Calls

In gene expression analysis, we often want to know whether a gene is present in transcripts. We can construct a hypothesis test to make the present call. The null hypothesis is expression absent. We may use negative-control sequences in a microarray to estimate the parameters of the null distribution. Negative-control sequences are typically selected that are dissimilar from the genome under study. Because negative controls should have very low intensity measurements and because the additive noise dominates at low intensities, we assume the null intensity has a Gaussian distribution with mean μ_{neg} and standard deviation σ_{neg}. For a particular sequence j in array i, its present call p-value can be computed as

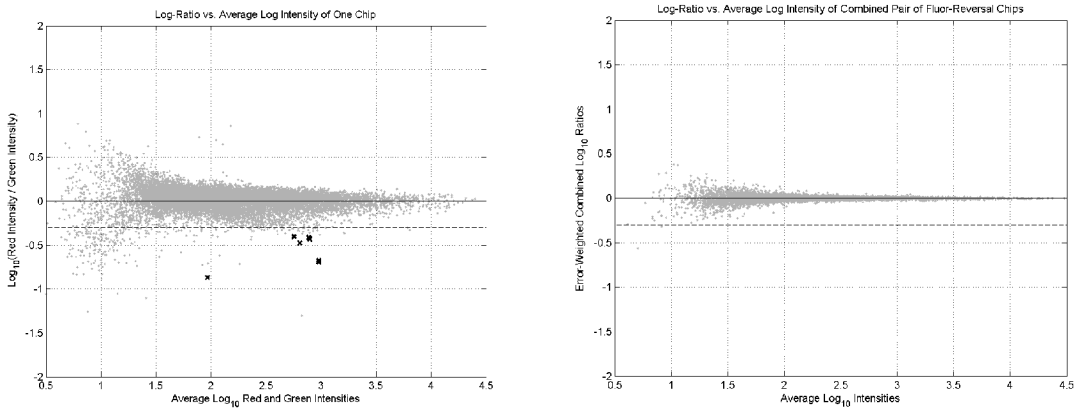
$$pvalue_{pres}(i, j) = 0.5 \cdot \left(1 - Erf \left(\frac{I(i, j) - \mu_{neg}(i)}{\sqrt{\sigma_I^2(i, j) + \sigma_{neg}^2(i)} \cdot \sqrt{2}} \right) \right) \qquad (20)$$

where *Erf* is the error function of a standard Gaussian distribution. This calculation provides a one-sided p-value. When p-value is small, e.g., less than 0.05, we reject the null hypothesis and accept the alternative hypothesis that the sequence transcript is present. For those sequences having large p-values we cannot simply call them absent because data analysis does not verify the null for any given experiment. They may be indeed biologically absent. Or they may be present but the measurements in the experiment are too noisy to make a confident call.

There may be two arguments against Equation (20). First, the intensity distribution is not Gaussian. Second, unlike the t-test, the p-value computation in (20) does not take degrees-of-freedom into consideration (see Discussion and Conclusion section). Overall intensity measurements in microarrays indeed do not have a Gaussian distribution. But the present call threshold is at very low intensity levels where the additive Gaussian noise dominates. The p-values near the commonly used threshold (<0.01-0.05) are still valid. At high intensities the Gaussian assumption is violated and p-value becomes less accurate. But there is no harm to the hypothesis test because those high-intensity sequences already have small enough p-values and are clearly present in the transcripts.

3.3 Differential Expression Calls

Traditionally, people use fold change (ratio) to make differential calls in detecting differential expressions. Often the call is made when the fold change is more than two. The fold change method assumes no additive errors in microarray measurements. When additive error exists, the variation in fold change increases when measured intensity value decreases. We can see this clearly in log-ratio (log of fold change) plots of same-versus-same experiments (Fig. 3). The fold change method can result in many false-positives at the low intensity end. At the same time, its sensitivity is low in detecting small fold changes at the high intensity end where the measurement variation in log-ratio is low. Overall the fold change method suffers from both low sensitivity and low specificity.



(a) (b)

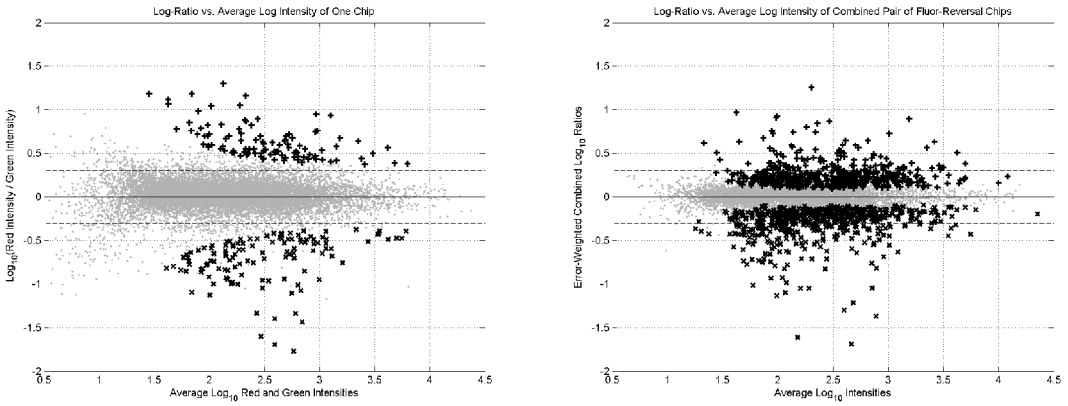
Fig.3. Log-ratio versus average log-intensity plots of a same-versus-same experiment (a) of one two-color array and (b) of a combined pair of fluorescent-reversal arrays. The same RNA sample is hybridized to the two channels of an array. The horizontal axis is the averaged log-intensities of the two channels. The vertical axis is the log-ratio of intensities of the two channels. The differential expression detection threshold is set at *p*-value<0.01 computed by Equation (21). Upregulated data, if any, are marked with a black "+." Downregulated data are marked with a black "x." Data that are not differentially expressed are shown as gray dots. In (a) each microarray spot is shown as one gray dot. In (b) two repeated spots in each array are error-weighted and combined and then the two fluor-reversal arrays are error-weighted and combined based on Equation (16). Any data called upregulated or downregulated are false positives in same-versus-same experiments. Two parallel dashed lines represent up and down two-fold change lines.

A statistical hypothesis test can be used in a differential expression study. The null hypothesis is that a gene is not differentially expressed. We can compute the *p*-value of the hypothesis test as

$$pvalue_{diff}(i,j) = 1 - Erf\left(\frac{|xdev(i,j)|}{\sqrt{2}}\right) = 1 - Erf\left(\frac{|lratio(i,j)|}{\sigma_{lratio}(i,j)\cdot\sqrt{2}}\right) \quad (21)$$

where *Erf* is the error function of a standard Gaussian distribution. This is a two-sided *p*-value including both up and down expressions. When the differential *p*-value of a sequence is small, e.g., less than 0.01, we call this sequence differentially expressed. It has been demonstrated that for same-versus-same experiments, *xdev* has a distribution close to Gaussian (Stoughton and Dai, 2002).

P-values provide us confidence information in making expression calls. *P*-values from error-model-based hypothesis tests set different fold change levels in differential calls for different measurement intensities. We can see in Fig 4 that for a given *p*-value threshold when intensities are low (the left side of the plot), only those having high log-ratios can be called differentially expressed. For the same *p*-value threshold, when intensities are high, sequences of low log-ratios can be confidently detected. In addition, we can see where some datapoints have similar intensities and similar log-ratios but some of them are called differentially expressed and some are not for the same given *p*-value threshold less than 0.01. Datapoints not detected have larger log-ratio errors from the error model estimations or from the replicates. These are improvements in sensitivity and specificity over the fixed fold change detection method.



(a) (b)

Fig.4. Log-ratio versus average log-intensity plots of a different-versus-different experiment (a) of one two-color array and (b) of a combined pair of fluorescent-reversal arrays. Two different RNA samples are hybridized to the two channels of an array. Any data called upregulated or downregulated in different-versus-different experiments are total positives for the given detection threshold, in this case, *p*-value<0.01.

Sometimes scientists use *t*-tests in differential expression detection. *T*-tests rely on replicates to estimate variance. P-value computed from the t-distribution is a proper gauge of the expected false-positive rate of the test under any given number of replicates. T-test does not assume any prior knowledge about the possible accuracy of the measurement technology itself. Sometimes, especially when the number of replicates is

small, for example two replicates, by chance some sequences may have differences between two repeated measurements close to zero. Estimated variances of these sequences are very small and much smaller than the inherent inaccuracy of the microarray technology. These small variances contribute partially to the overall false-positive rate. Fig 5 demonstrates false-positives in a *t*-test example. Many of those false positives have log-ratios very close to zero. In the error model method, the error model describes the intensity-dependent variation of the microarray technology. The error model helps to set a low limit in the variance estimation to avoid some of these false positives. This is similar to the penalized t-test (Tusher, et al., 2001) where a small constant is added to the variance estimation to set the lower bound. In stead of a constant, the lower bound in the error model varies according to the expected intensity-variance relationship of the microarray technology.

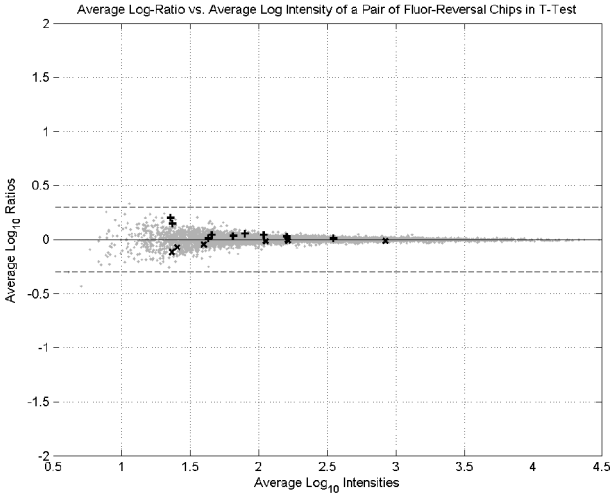


Fig.5. Average log-ratio versus average log-intensity plot of the same experiment shown in Fig. 3(b) but with *t*-test differential calls. The differential expression detection threshold is set at *p*-value<0.01. Upregulated data are marked with a black "+." Downregulated data are marked with a black "x." Data that are not differentially expressed are shown as gray dots. Any data called upregulated or downregulated are false positives in same-versus-same experiments.

Receiver Operation Characteristics (ROC) curves (He, et al., 2003) allow us to compare sensitivities (the total detection rate) and specificities (one minus the false-positive rate) from different analysis methods. In an ROC plot, the y-axis indicates the total positive rate, estimated by the number of datapoints (genes) in the different-vs-different experiments that passes a threshold divided by the total number of datapoints available. The x-axis indicates the false-positive rate, estimated by the number of datapoints in the same-versus-same experiments that passes the threshold divided by the total number of datapoints available. One datapoint in the ROC curve of the given analysis method is defined by these two rates, and the whole ROC curve is constructed by varying the detection threshold. An analysis method has higher statistical power if its ROC curve is closer to the upper-left corner. A higher ROC curve gives higher sensitivity for a given

false positive rate, or lower false positive rate for a given sensitivity. If the differential expression calls are made based on 50-50 chance, its ROC curve is the diagonal line.

In Fig 6 we plot ROC curves for our error model approach, the fold change approach, and *t*-test. Data are fluor-reversed two-color duplicates. The x-axis is the false-positive rate measured from same-versus-same experiments shown in Fig 3(b). The y-axis is the total positive-rate from different-versus-different experiments in Fig 4(b). The number of replicates N=2, so that there is only one degree-of-freedom for the *t*-test. For comparison purposes, the same data preprocessing is applied to the data before applying all three methods so that they have the same log-ratio as input. In this example, the error model approach is clearly superior to the other two.

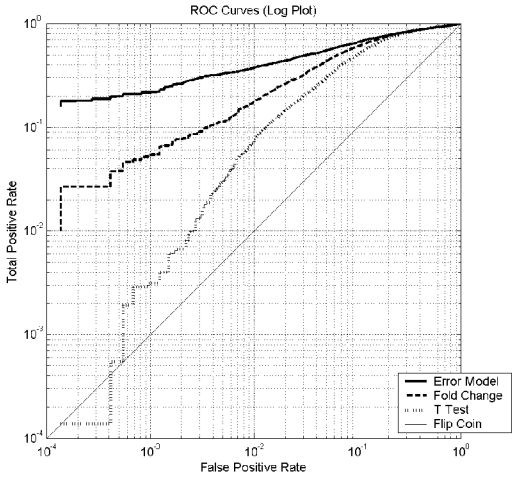


Fig.6. ROC curves of the three differential expression detection methods using a small number of replicates: the error model approach, fold change method, and *t*-test. Two pairs of fluorescent-reversal technical replicated arrays are used in generating the curves. The horizontal axis is the false positive rate computed from one pair of same-versus-same replicates shown in Fig. 3(b). The vertical axis is the total positive rate computed from one pair of different-versus-different replicates shown in Fig. 4(b). ROC curves closer to the upper-left corner of the plot have higher statistical powers in terms of sensitivity and specificity.

Technical replicates of the same RNA are used in the ROC curve study shown in Figure 6. To demonstrate the benefit of the error model in analyzing biological replicates, we design a study where the number of replicate increases from one to five. In this study samples from different animals (mice) are hybridized in both same-vs.-same and different-vs.-different experiments. In the same-vs.-same experiment, the animals are under the same drug vehicle treatment. In the different-vs.-different experiment, the comparison of differential expression are between vehicle-treated animals and drug compound treated animals. The ROC curves of the study are shown in Figure 7. In this particular study, when the number of replicates is small (n=1 and 2), the power of differential expression detection is not much better than the random chance of coin flipping. When the number of replicates increases, the detection power increases and the ROC curves rise in all three detection methods (error-model, t-test and fold-change). It is

noted that the discrimination among these three methods becomes more significant as the number of replicates increases. The ROC curves of the error model is always the highest from the number of replicates n=3 to 5. It is general in our observations that the error model ROC curve has the similar height as the t-test ROC curve of one more replicates. For example, the error model ROC curve of n=3 (or n=4) is similar to the t-test ROC curve of n=4 (or n=5).

The error-model approach has also been compared to some other published methods (Rajagopalan, D., 2003) where its advantages in improving detection sensitivity and specificity are clearly demonstrated.

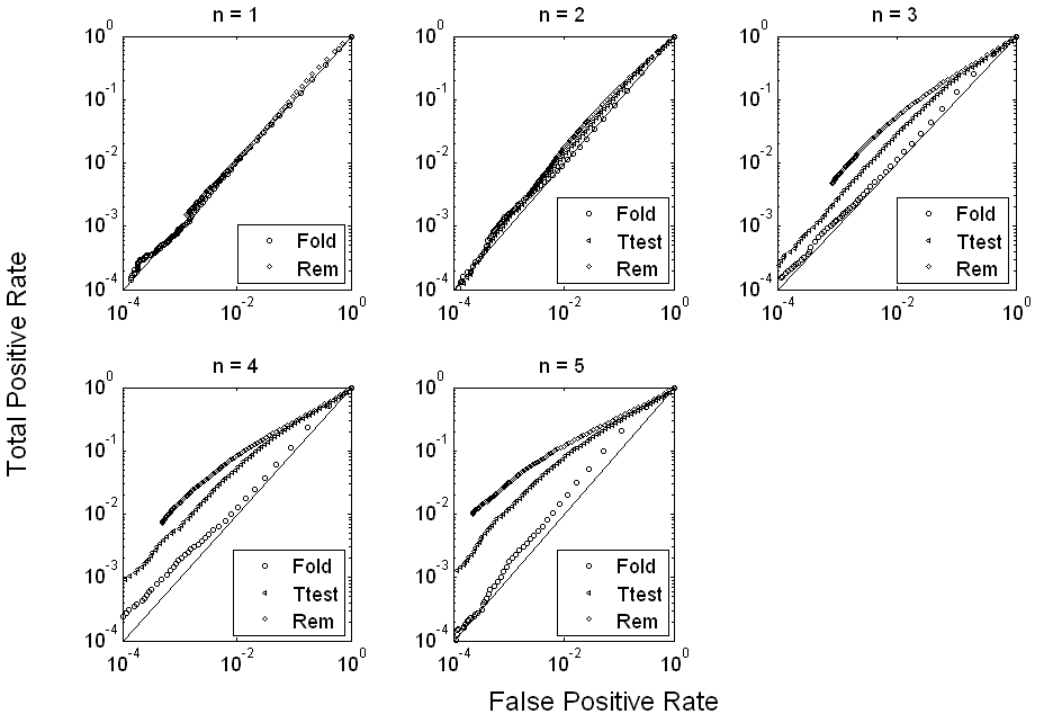


Figure 7. ROC curves of different number of biological replicates. Results of three differential detection methods (Rosetta error model (REM) , t-test, and fold change) are compared. The number of replicates n increases from 1 to 5. The REM method consistently provides higher detection power than the other two methods in this study.

4 DISCUSSION AND CONCLUSION

Microarray measurement errors are inevitable. They are significantly intensity-dependent. Properly designed error models provide estimates of the measurement error. In addition to the measurement itself, the measurement error carries important information that can be incorporated in microarray data analysis. For a given analysis method, sensitivity and specificity cannot be increased simultaneously unless additional information becomes available. Traditionally the only way to increase information is to request more replicates. But when more replicates are practically not possible, the measurement error information

becomes very valuable. Including this error information in data analysis increases statistical power (higher sensitivity and higher specificity) for a given small number of replicates, as indicated in the elevated ROC curves. This approach is the most significant contribution of the error model technology to microarray data analysis when the number of available replicates is limited.

Alternatively, to simplify the microarray error analysis, we can assume there are two main error sources, technical error of the measurement process and biological error of the subject variation (e.g., gene expression differences among animals under the same treatment):

$$\sigma_{total}^2 = \sigma_{tech}^2 + \sigma_{bio}^2 . \qquad (22)$$

When using the error information in hypothesis tests, we should clearly understand its implication on the test results. When the number of replicates is small, biological variation estimates σ_{bio}^2 are usually not separable from technological variations σ_{tech}^2 in microarray measurements. Technical variation can be a significant part of the uncertainty in the differential analysis particularly at low intensities. It defines the minimum amount of randomness we receive in the results. We provide a method to blend the predicted technical variations (the propagated error) together with the scattered error into the variance estimation in (19). The blended variance is used in hypothesis tests to make present calls and differential calls. The propagated error from the error model is a conservative estimation of σ_{tech}^2, which is not gene-specific and non-zero. It can be biased. The scattered error is an unbiased estimation of σ_{total}^2 that includes σ_{bio}^2, which is gene-specific in general. It is important to understand the meaning of the null hypothesis under a different number of replicates. For example, in differential calls, when the number of replicates is large, the null hypothesis states "no change in expression measurements between the two conditions is more significant than the variation caused by the microarray technology *and* differences among replicated subjects (animals)." Rejecting this statement is ultimately what we are interested in doing. However, when there are no replicates, the null hypothesis becomes "no change in expression measurements between the two conditions is more significant than the expected variation caused by the microarray technology." Often we are in between these two extremes when we have a few biological replicates. Biological variance cannot be reliably estimated unless we have enough biological replicates. However, using the conservatively predicted technical variance from error models, we can at least avoid some of the overall false positives. The predicted technical variance prevents the overall variance estimation from being too small and smaller than the inherent technical uncertainty. For many biological research projects, an under-estimated variance is a big problem because it results in a high false-positive rate. False positive results consume labor and resources in the form of follow-up studies.

In the current error model approach, when blending the propagated error and the scattered error in (19), we set a lower bound in the total variance estimation. Although the error model is not gene-specific, the combination process in (19) makes the estimated total variance $\sigma_{\bar{x}}$ gene-specific, as long as N>1. This phenomenon occurs because the scattered

error is gene-specific when biological replicates are used. In some of our early work (Hughes, 2000), we may be able to introduce a gene-specific correction factor to further adjust the lower bound based on large number of previously accumulated biological replicates. This factor is considered as a model of biological variance. But this biological model is often not available for most studies.

The Rosetta method described here has been validated (Rajagopalan, 2003) and successfully applied by users of the Rosetta Resolver system (Bassett, 2000). These customers conduct gene expression analysis with data ranging from different single-color and two-color microarray technologies, such as those from Affymetrix, Agilent Technologies, GE Healthcare (formerly Amersham Biosciences), and in-house cDNA arrays. Although there are critics (Dror, 2001), the benefit of the error model method in improving the detection power is clearly demonstrated in the ROC example, illustrated in the previous section. Compared with many other error model methods, the method described in this paper is simple. The reason such a simple method actually works well in real applications is that it captures the main error behavior in the microarray technology with minimum assumptions. It offers a solution to a practical problem of the variance estimation when the number of replicates is small.

The predicted measurement variance makes many novel microarray data analysis methods possible. For example, we provide error-weighted averaging to combine replicated microarray measurements. The error-weighting method can also be applied to compute a similarity matrix during clustering analysis where measurements with larger errors contribute less to similarity computations, such as correlation coefficient or Euclidian distance. We have also successfully incorporated the use of the predicted technical variance in ANOVA analysis, microarray intensity transformations, and other applications. Results will be published separately.

In summary, the Rosetta error model described in this paper provides us prediction about microarray measurement errors. The additional information gained from using the error model opens many new opportunities for us to improve the quality of microarray data analysis.

ACKNOWLEDGEMENTS

Authors would like to thank Roland Stoughton for reviewing and helping to revise the manuscript, Daniel Holder and Richard Raubertas at Merck Research Laboratory for helpful discussions and suggestions of error models, and Lisa Owen and Bill Kauffman at Rosetta Biosoftware for helping to prepare the manuscript. Authors are grateful to Roger Bumgarner at the University Washington for providing the microarray data demonstrated in this paper.

REFERENCES

Agy, M., Li, Y., Thomas, M., Korth, M. Geiss, G., Kwieciszewki, B., Bumgarner, R. Katze, M. (2003). Using cDNA microarrays for comparative genomics and for cellular gene expression profiling during SIV infection. *J. Medical Primatol*, 32 (4-5): 305.

Bassett, D. E., Jr. (2000). The Rosetta Resolver® system: a comprehensive storage, visualization, and analysis tool for high-volume gene expression data. Meeting poster, 12th International Genome Sequencing and Analysis Conference.

Casolari, J., Brown, C., Komili, S., West, J., Hieronymus, H., Silver, P. (2004). Genome-wide localization of the nuclear transport machinery couples transcriptional status and nuclear organization. *Cell*, 117, 427-439.

Chen, Y., et al., (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images, *J Biomed Opt*, p364-374.

Dror, R.O. (2001) Noise models in gene array analysis, Area exam report in MIT Department of Electrical Engineering and Computer Science, available at http://www.ai.mit.edu/people/rondror/

Dror, R.O., et al., (2003) Bayesian estimation of transcript level using a general model of array measurement noise, *J Comput Biol*, v.10, p.433-452.

Durbin, B.P., et al., (2002) A variance-stabilizing transformation for gene-expression microarray data, Bioinformatics, v.18, Suppl.1, p.S105-S110

Falls, G. ; Gemzik, B. ; Car, B. D. ; and Lehman-McKeeman, L. D. (2003). Hepatobiliary transporter induction in altered thyroid hormone homeostasis: a microarray analysis. *J Toxicol Sci*, 72 (S-1): p261.

Geiss, G., Salvatore, M., Tumpey, T., Carter, V., Wang, X., Basler, C., Taubenberger, J., Bumgarner, R., Palese, P., Katze, M., and Garcia-Sastre, A. (2002). Cellular transcriptional profiling in influenza A virus-infected lung epithelial cells: the role of the nonstructural NS1 protein in the evasion of the host innate defense and its potential contribution to pandemic influenza. *Proc Natl Acad Sci*, 99(16):10736-41.

Geiss, G., Carter, V., He, Y. Kwieciszewski, B., Holzman, T., Korth, M., Lazaro, C., Fausto, N., Bumgarner, R., Katze, M. (2003). Gene expression profiling of the cellular transcriptional network regulated by alpha/beta interferon and its partial attenuation by the hepatitis C virus nonstructural 5A protein. *J Virol*, 77(11):6367-75.

He, Y.D., et al., (2003) Microarray standard data set and figures of merit for comparing data processing methods and experiment designs, *Bioinformatics*, v.19, p.956-965.

Helfrich, R.; Davis-Taber, R. ; Choi, W. ; Zhu, C. ; Gauvin, D. ; Thimmapaya, R. ; Gubbins, E. ; Vos, M. ; Kage, K. ; Gopalakrishnan, M. ; Surowy, C. ; Honore, P. ; Kroeger, P. ; Faltynek, C. ; Daza, A. ; and Scott, V. (2003). Tranascriptional profiling of dorsal root ganglion in the chronic constriction injury model of neuropathic pain using microarray technology. Meeting abstract, 33rd Annual Meeting of the Society of Neuroscience.

Holder, D, et al., (2001). Statistical analysis of high density oligonucleotide arrays: a SAFER approach, available at http://oz.berkeley.edu/users/terry/zarray/Affy/GL_Workshop/SAFERv04.pdf.

Hori, Y. ; Spurr-Michaud, S. ; Argueso, P. ; and Gipson, I. K. (2003). Effect of retinoic acid on gene expression in human conjunctival epithelial cells as determined by microarray analysis. Meeting abstract, Annual Meeting of the Association for Research in Vision and Ophthalmology, Fort Lauderdale, FL, USA, May 4-8, 2003.

Hubbell, E.; Liu, W.M; Mei, R. (2002) Robust estimators for expression analysis, *Bioinformatics*, v18, n12, p1585-1592.

Hughes, T.R., et al., (2000) Functional discovery via a compendium of expression profiles, *Cell*, v102, p109-126.

Jain, N., et al., (2003) Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays, *Bioinformatics*, v.19, p.1945-1951.

Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucelotide arrays: expression index computation and outlier detection, PNAS, v.98, p31-36.

Liu, M. ; Richards, S. M. ; Schirra, F. ; and Sullivan, D. A. (2002). Sex and androgen influence on gene expression in lacrimal glands of normal and autoimmune mice. Meeting abstract, Annual Meeting of the Association for Research in Vision and Ophthalmology. Fort Lauderdale, FL, USA, May 5-10, 2002.

Liu, W.M., et al., (2002) Analysis of high density expression microarrays with signed-rank call algorithms, *Bioinformatics*, v18, n12, p1593-1599.

Lockhart, D.L., et al., (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nat Biotechnol*, vol. 14, pp. 1675-1680.

Lonnstedt, I. and Speed, T. (2002) Replicated microarray data, *Stat Sinica* v.12, p.31-46.

Marini, F. ; Reid, J.; Altieri, M. ; Carboni, L. ; Blakemore, S.; Cook, T.; Vitulli, G. ; Caldara, F. ; Arban, R. ; Corsi, M. ; Jansson, B.; and Domenici, E. (2003). Transcriptome analysis of rat brain after chronic and acute treatment with fluoxetine and selective antagonists of NK – 1 and CRF – receptors. Meeting abstract, 33rd Annual Meeting of the Society of Neuroscience.

Quackenbush, J. (2002) Microarray data normalization and transformation, *Nat Genet*, supplement, v.32, p.496-501.

Rajagopalan, D. (2003). A comparison of statistical methods for analysis of high density oligonucleotide array data. *Bioinformatics*, 19 (12), 1469-1476.

Richards, S. M. ; Schirra, F. ; Yamagami, H. ; Suzuki, T. ; and Sullivan, D. A. (2003). Sex-related differences in gene expression in the mouse meibomian gland. Meeting abstract, Annual Meeting of the Association for Research in Vision and Ophthalmology. Fort Lauderdale, FL, USA, May 4-8, 2003.

Roberts, C.J., et al., (2000) Supplementary Material in "Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles," *Science*, v.*287*, p.873-880.

Rocke, D.M. and Durbin, B. (2001) A model for measurement error for gene expression arrays, *J Comp Biol*, v.8, p.557-569.

Schadt, E.E., et al., (2002) Feature extraction and normalization algorithm for high-density oligonucleotide gene expression array data, *J Cell Biochem*, v84, p.120-125.

Schirra, F.; Liu, M. ; and Sullivan, D. A. (2002). Androgen regulation of gene expression in the mouse meibomian gland. Meeting abstract, Annual Meeting of the Association for Research in Vision and Ophthalmology. Fort Lauderdale, FL, USA, May 5-10, 2002.

Smyth, G. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, Statistical Applications in Genetics and Molecular Biology, v.3, n.1, available at http://www.bepress.com/sagmb/vol3/iss1/art3

Smith, M., Yue, Z., Geiss, G. Sadovnikova, N., Carter, V., Boix, L., Lazaro, C., Rosenberg, G., Bumgarner, R., Fausto, N., Bruix, J., and Katze, M. (2003) Identification of novel tumor markers in hepatitis c virus-associated hepatocellular Carcinoma. *Cancer Res*, 63 (4), 859-864.

Stoughton, R. and Dai, H.Y., (2002) Statistical combining of cell expression profiles, US Patent #6,351,712.

Theilhaber, J., et al., (2001) Bayesian estimation of fold changes in the analysis of gene expression: the PFOLD algorithm, *J Comp Biol*, v.8, p.585-614

Thimmapaya, R. ; Davis-Taber, R. ; Choi, W. ; Zhu, C. ; Gubbins, E. ; Vos, M. ; Helfrich, R. ; Kage, K. ; Daza, A. ; Donnelly-Roberts, D. ; Harris, R. ; Surowy, C. ; Honore, P. ; Kroeger, P. ; Faltynek, C. ; Goplalakrishnan, M. ; and Scott, E. (2003). Gene expression profiling in the spinal nerve ligation model of neuropathic pain using microarray technology. Meeting abstract, 33rd Annual Meeting of the Society of Neuroscience.

Tonouchi, M. (2002). Gene expression analysis tool accelerating post-genomic research: Resolver system of Rosetta Biosoftware. *Biobencha*, 2 (6), 61-66.

Tu, Y., et al., (2002) Quantitative noise analysis for gene expression microarray experiments," PNAS v.99, p14031-14036.

Tusher, V.G., et al., (2001) Significance analysis of microarrays applied to the ionizing radiation response, PNAS v.98, p5116-5121.

Weng, L. (2003). Methods for analysis of measurement errors in measured signals, US Patent Application #20030226098.

van 't Wout, A., Lehrman, G., Mikheeva, S., O'Keeffe, G., Katze, M., Bumgarner, R., Geiss, G., Mullins, J. (2003). Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4(+)-T-cell lines. *J Virol,* 77(2):1392-1402.

Yang, Y.H., et al., (2002) Normalization for cDNA microarray data: arobust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Res*, v.30, n.4.

6.0 APPENDIX-A: DERIVATION OF PARAMETER *XDEV*

Assuming two intensity measurements are I_1 and I_2, the log-ratio of intensities is

$$lratio = \ln(I_2 / I_1) = \ln(I_2) - \ln(I_1). \quad \text{(A1)}$$

When intensity error is small, the standard deviation of the log-ratio is

$$\sigma_{lratio} \approx \sqrt{(\sigma_1 / I_1)^2 + (\sigma_2 / I_2)^2}. \quad \text{(A2)}$$

When *differential* expression is weak, i.e., I_1 and I_2 are similar, we assume

$$I_1 \approx I_2 = a. \quad \text{(A3)}$$

In a Taylor expansion, (A1) can be written as

$$\ln(I_2 / I_1) = \ln(I_2) - \ln(I_1) = (\ln(a) + (I_2 - a)/a + ...) - (\ln(a) + (I_1 - a)/a + ...). \quad \text{(A4)}$$

With the weak differential expression assumption, (A4) can be approximated as

$$\ln(I_2 / I_1) = \ln(I_2) - \ln(I_1) = (\ln(a) + (I_2 - a)/a) - (\ln(a) + (I_1 - a)/a) = (I_2 - I_1)/a \quad \text{(A5)}$$

so that the log-ratio error in (A2) can be approximated as

$$\sigma_{lratio} \approx \sqrt{(\sigma_1 / I_1)^2 + (\sigma_2 / I_2)^2} \approx \frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{a}. \quad \text{(A6)}$$

Parameter *xdev* is defined as the ratio between log-ratio and log-ratio error in (15). From (A5) and (A6) we obtain

$$xdev = \frac{lratio}{\sigma_{lratio}} = \frac{\ln(I_2) - \ln(I_1)}{\sigma_{lratio}} \approx \frac{I_2 - I_1}{\sqrt{\sigma_2^2 + \sigma_1^2}}. \quad \text{(A7)}$$

This is the definition given in (14).

7.0 APPENDIX-B: ERROR MODEL PARAMETER ESTIMATION

There are different methods to estimate the error model parameters (RANDOM, POISSON and FRACTION) based on training data. One is to divide the intensity distribution percentile to three ranges, such as low (<2%), middle (>2% and <70%), and high (>70% and <99%). Based on Equation (21) we compute differential p-values of

pairwise comparisons between technical replicates. By varying the error model parameters, we calibrate the p-value at the default threshold 0.01. We first vary the parameter RANDOM and measure the actual false positive rate at the *low* intensity range. The false positive rate is the number of microarray features that have p-value below the threshold in the range divided by the total number of features in the range. To be conservative, we usually calibrate the mean pairwise false positive rate to be approximately the half of the p-value threshold. After the parameter RANDOM is calibrated, we then vary the parameter FRACTION and measure the actual false positive rate at the *high* intensity range. The parameter FRACTION is set at the level where the mean false positive rate is about half of the p-value threshold. With the estimated parameters RANDOM and FRACTION we calibrate the parameter POISSON at the last. We vary the parameter POISSON to get the actual false positive rate in the *middle* intensity range to about the half of the p-value threshold.

Error model parameters are estimated for a specific microarray technology, including sample preparation method, hybridization and scanning. For different microarray technologies, we usually have different sets of error model parameters. We keep the error model parameters as constants when processing the microarray data of the given technology.