Gene expression

Modeling nonlinearity in dilution design microarray data

Xiuwen Zheng¹, Hung-Chung Huang², Wenyuan Li², Peng Liu⁴, Quan-Zhen Li⁵ and Ying Liu₂,³,*

¹Department of Mathematical Sciences, ᵃDepartment of Computer Science, ᵃDepartment of Molecular and Cell Biology, University of Texas at Dallas, Richardson, TX 75083-0688, ⁴Department of Statistics, Iowa State University, Ames, IA 50011 and ⁵Microarray Core facility, University of Texas Southwestern Medical Center, Dallas, TX 75390 USA

Received on September 25, 2006; revised on January 2, 2007; accepted on January 8, 2007

ABSTRACT

Motivation: Dilution design (Mixed tissue RNA) has been utilized by some researchers to evaluate and assess the performance of multiple microarray platforms. Current microarray data analysis approaches assume that the quantified signal intensities are linearly related to the expression of the corresponding genes in the sample. However, there are sources of nonlinearity in microarray expression measurements. Such nonlinearity study in the expressions of the RNA mixtures provides a new way to analyze gene expression data, and we argue that the nonlinearity can reveal novel information for microarray data analysis. Therefore, we proposed a statistical model, called proportion model, which is based on the linear regression analysis. To approximately quantify the nonlinearity in the dilution design, a new calibration, beta ratio (BR) was derived from the proportion model. Furthermore, a new adjusted fold change (adj-FC) was proposed to predict the true FC without nonlinearity, in particular for large FC.

Results: We applied our method to one microarray dilution dataset. The experimental results indicated that, to some extent, there are global biases comparing with the linear assumption for the significant genes. Further analysis of those highly expressed genes with significant nonlinearity revealed some promising results, e.g. ‘poison’ effect was discovered for some genes in RNA mixtures. The adj-FCs of those genes with ‘poison’ effect, indicate that the nonlinearity can also be caused by the inherent feature of the genes besides signal noise and technical variation. Moreover, when percentage of overlapping genes (POG) was used as a cross-platform consistency measure, adj-FC outperformed simple fold change to show that Affymetrix and Illumina platforms are consistent.

Availability: The R codes which implements all described methods, and some Supplementary material, are freely available from http://www.utdallas.edu/~ying.liu/BetaRatio.htm

Contact: ying.liu@utdallas.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Microarray technology, one of the latest breakthroughs in molecular biology, is shifting experimental approaches from single gene studies to genome-level analysis. It allows monitoring of gene expression for tens of thousands of genes in parallel. One of the key aims of microarray is to determine a list of differentially expressed genes (DEGs) that distinguish biological change of states such as disease from normal.

The success of microarray technology has led to the production of multiple array platforms (Barnes et al. 2005). Although different microarrays are based on hybridization of nucleic acid strands, they differ in the types and composition of probes used (short-oligonucleotide, long-oligonucleotide, cDNA, etc.), the deposition technologies, the hybridization and labeling protocols (Barnes et al., 2005; Draghici et al., 2006; Thomson et al., 2005) The reproducibility and cross-platform consistency have drawn a lot of concerns on microarray technology (Barnes et al., 2005; MAQC Consortium, 2006; Marshall, 2004). To determine the cross-platform reproducibility, Barnes et al. (2005) proposed a study design based on a dilution series of two human tissues (blood and placenta) and tested in duplicate on Affymetrix (AFX) and Illumina (ILM) platforms. The dilution design is to mix two different RNA samples (from blood and placenta) at known proportions (Fig. 1) and the same RNA mixture is analyzed in duplicate on each platform. Microarray measurements normally assume a linear relationship between the quantified intensity and the amount of mRNA in the sample. However, as pointed out by Ramdas et al. (2001), there are sources of nonlinearity in microarray expression measurements and the true intensity may not follow exact linear relationships; also, ratio underestimation is a well-known phenomenon of microarray data (Dudley et al., 2002; Yuen et al., 2002). Some studies dealing with nonlinearity in microarray data have been reported, which mainly focused on correcting the nonlinear relationship between intensity data from different photomultiplier tube (PMT) gains. (Bengtsson et al., 2004; Dodd et al., 2004; Dudley et al., 2002 Lyng et al., 2004).

In our study, the different dilutions of the mixed RNA samples in the same PMT setting were utilized to quantify and correct the nonlinearity, so that we can predict the true fold

*To whom correspondence should be addressed.
change for improving reproducibility and consistency between platforms. When analyzing the dilution design gene expression data (Barnes et al., 2005), we found that the linear assumption does not always stand. In one extreme case, the intensity of the gene 208295_x_at in the sample with 5% placenta is almost as high as its intensity in the sample with 100% placenta (Fig. 2a). This is a significant nonlinear phenomenon in the intensities of a gene. Another example is shown in Figure 2b. Two genes with almost the same fold change (FC) are plotted and they exhibit different inclination of nonlinearity. Although those genes have the same FC, they should be considered differently when determining the preference of DEGs because of their different nonlinearity.

In an effort to model the nonlinearity in dilution design, we first estimated the ‘actual’ proportion (AP) according to the quantified intensities detected in the mixed RNA samples. Based on dilution proportion (DP) and AP, we defined the nonlinearity via the curve in the AP versus DP plot. Then we examined the nonlinearity in the gene expression data and developed a statistical model called proportion model, based on the linear regression analysis. According to proportion model, we define fold change (FC) as the estimated FC derived from all dilution samples, rather than the traditional FC (only pure samples). Moreover, a calibration, beta ratio (BR), derived from regularized beta function was calculated to approximately quantify the nonlinearity. We also discussed the effect of log2 transform on BR, and proposed a new adjusted fold change (adj-FC) based on the BR versus FC plot, to predict the true FC without nonlinearity.

We applied our model to the microarray dilution dataset (Barnes et al., 2005). The DEG lists were determined by ANOVA (Zar, 1999), and 0.001 was selected as P-value threshold. The experimental results indicated that, to some extent, there are global biases comparing with the linear assumption for the significant genes. Further analysis of those highly expressed genes with significant nonlinearity revealed some promising results as seen from the genes 208295_x_at (Fig. 2a) with ‘poison’ effect. The adj-FCs of such genes are much larger than normal genes, which indicates that the nonlinearity can be also caused by the inherent feature of the genes besides signal noise and technical variation. Two representative genes, CSH1 and CSH2, were identified to be such genes with ‘poison’ effect. They behave with the feature of ‘multiple initiation sites of transcription’, and are translated to become human placental lactogen which has the highest concentration known in pregnancy. Moreover, when percentage of overlapping genes (POG) was used as a cross-platform consistency measure (Shi et al., 2005), adj-FC outperformed simple fold change to show that Affymetrix and Illumina
platforms are consistent, via 0.001 P-value threshold, in particular for large FC.

2 MATERIALS AND METHODS

2.1 Gene expression dataset

The datasets, consisting of 54675 genes for Affymetrix platform and 24136 genes for Illumina platform were made publicly available by the original authors (Barnes et al., 2005). Briefly, Affymetrix ‘U133plus2.0 GeneChip’ and Illumina ‘Sentrix HumanRef-8 Expression BeadChip’ were utilized by Barnes et al. (2005) to get the intensity data for the RNA mixtures from whole blood and placenta samples. These expression data were retrieved from the web site ‘http://benzer.ubic.ca/platformCompare’. All these data were derived by the assay of six RNA mixtures (100%:0%, 95%:5%, 75%:25%, 50%:50%, 25%:75%, 0%:100%; blood : placenta) (Fig. 1) with two technical duplicates for each mixture (Barnes et al., 2005). AFX data were normalized with the RNA matrix from Bioconductor’s ‘affy’ package and ILM data were normalized using the ‘normalize.quantiles’ function from ‘affy’ package (Barnes et al., 2005). No further filtering or normalization on these data was performed. The common gene list for the inter-platform comparisons were represented in terms of a linear model for each gene:

\[ Y = X\beta + \epsilon \] (4)

where the design matrix \( X \) is as the following:

Let \( \beta_g \) be the true gene expression (transcription) level, \( \mu_g \) be the true signal intensity. Ideally, if we could observe the true intensity \( \mu_g \), then we may infer the true expression level \( \gamma_g \). There is a measurement function \( \mu_g = f(\gamma_g) \) to describe the relationship between \( \gamma_g \) and \( \mu_g \), which may not be linear (Bengtsson et al., 2004).

Let \( \theta_1, \ldots, \theta_m (0 = \theta_1 < \theta_2 < \ldots < \theta_m = 100\%) \) denote the dilution proportions (DP) of placenta from the RNA mixture experimental design, where \( m \) is the number of dilution series (\( m = 6 \) and \( \theta_1 = 0\%, \theta_2 = 5\%, \theta_3 = 25\%, \theta_4 = 50\%, \theta_5 = 75\% \) and \( \theta_6 = 100\% \) in our study). Since the RNA mixture with the proportion \( \theta_i (1 \leq i \leq m) \) was created from sample blood and placenta, we can formulate the true signal intensity. Actually, the constraint \( 0\% \leq \theta_i \leq 100\% \) is as the following:

\[ \gamma_g = [(1-\theta_i)\gamma_g^{blood} + \theta_i\gamma_g^{placenta}] \] (1)

However, the linear relationship in Equation (1) does not always hold for the true intensities. Let \( \mu_{g,\theta_i} = f(\gamma_g^{\theta_i}) \), and define the corresponding actual proportions (APs) \( p_{g1}, \ldots, p_{gm} \) as

\[ p_{gi} = \begin{cases} \mu_{g,\theta_i}/\theta_i & \mu_g^{blood} \neq \mu_g^{placenta} \\ \theta_i & \mu_g^{blood} = \mu_g^{placenta} \end{cases} \] (2)

which probably are different from DP. Please note that \( p_{g3} = 0 \) and \( p_{gm} = 1 \).

Our model estimates AP, \( p_{gi} \), using the weighted least square method, which is described in more details in section 2.2.1. Actually, the constraint \( 0\% \leq \theta_i \leq 100\% \) is enforced, because the true intensity from the mixed sample (e.g. 50% placenta) should be between sample A (0% placenta) and sample B (100% placenta) in general. In the perfect linear situation, AP is the same as DP, i.e. \( p_{gi} = \theta_i (1 \leq i \leq m) \). Hence, to some extent, the information from \( p_{gi} \) can be used to describe the nonlinearity. Then, we calculate BR value from the regularized beta function to combine all values of \( p_{gi} \) together in order to describe the curvature of the nonlinearity.

2.2.1 The estimation of actual proportion Let \( n_i (1 \leq i \leq m) \) denote the number of replicates with the \( ith \) dilution series, and \( Y_{g,\theta_i} (1 \leq j \leq n_i) \) for the \( jth \) quantified intensity in the \( ith \) dilution series. Hence, the quantified intensity \( Y_{g,\theta_i} \) can be expressed as

\[ Y_{g,\theta_i} = (1-p_{gi}) \cdot \mu_g^{blood} + p_{gi} \cdot \mu_g^{placenta} + \epsilon_{g,\theta_i} \] (3)

where \( \epsilon_{g,\theta_i} \) is the error with \( E(\epsilon_{g,\theta_i}) = 0 \). The design of microarray experiments can be represented in terms of a linear model for each gene:

\[ Y_g = X\beta_g + \epsilon \] (4)

where the design matrix \( X \) is as the following:

\[
X_g = \begin{bmatrix}
1 - p_{g1} & p_{g1} \\
\vdots & \vdots \\
1 - p_{gm} & p_{gm}
\end{bmatrix}
\] (5)

There are two unknown types of variables: the true intensity vector \( \beta_g \) from the sample A and B, the actual proportions \( p_{gi} = (p_{g1}, \ldots, p_{gm}) \) in Equation (2). Based on the definition of AP in Equation (2), we should determine whether \( \mu_g^{blood} = \mu_g^{placenta} \) first via other statistical methods (e.g. ANOVA), and then analyze further those genes lower than critical P-value.

Considering the different variances in the same dilution replicates, we introduce a diagonal matrix of weights \( W_g \). The weights to be used in weighted regression are inversely proportional to the sample variance \( s_g^2 \) of the \( ith \) dilution series (Rocke and Durbin, 2001). A positive number (we chose \( c_g = \max(1/s_g^2, \ldots, 1/s_m^2) \)) was added in deriving the weights so that the weights do not become too large or too small. Hence, we get:

\[ W_g = \text{diag}(c_g + 1/s_g^2, \ldots, c_g + 1/s_m^2) \]

We need to minimize the residual sum of squares (RSS) to estimate these variables. Actually, RSS can be expressed as a function of \( \beta_g \) and \( \rho_g \):

\[ \text{RSS}(\beta_g, \rho_g) = (Y_g - X\beta_g W_g Y_g)^T (Y_g - X\beta_g W_g Y_g) \] (6)

The best estimator for \( \beta_g \) is a function of \( \rho_g \):

\[ \hat{\beta}_g = (X_g' W_g X_g)^{-1} X_g' W_g Y_g \] (7)

We substitute \( \hat{\beta}_g \) with \( \hat{\beta}_g \) so that Equation (6) can be rewritten as a function with one parameter \( \rho_g \) in the form of \( \text{RSS}(\rho_g) \).

To solve the weighted least square problem becomes to solve the following optimization problem:

\[ \min_{\rho_g} \text{RSS}(\rho_g) \quad \text{s.t.} \quad 0 \leq \rho_{g1} \leq \cdots \leq \rho_{gm} \leq 1 \] (8)

We use ‘Nelder–Mead’ method with the constrained conditions to get the numerical results, which is provided by the function ‘constrOptim’ in R (2006). The DPs are used for the initial points.

2.2.2 Beta ratio: an approximate measure of nonlinearity We applied Method (8) to normalized AFX data without log2 transform. The AP of each gene was calculated based on both of the replicates. The relationship between DP and AP, DP-versus-AP curve,
was shown in Figure 3a (only some of the genes were shown). From Figure 3a, several facts can be observed: for some genes, only 5% DP behaves as if there was around 90% AP reflected on the intensity, i.e. DP ≪ AP, (e.g. 208295_x_at); while for some other ones, 75% DP only corresponds to around 20% AP (e.g. 237729_at). Furthermore, some genes follow the linear pattern that the DP is approximately equal to the AP (e.g. 213478_at). In order to measure the curvature of the nonlinearity of Figure 3a, we employed a smooth curve to approximate AP of each gene.

Assume that there is a curve family $\Psi$ with the parameter $\phi$, satisfying $f_0(\theta_i) = p_0$ for $1 \leq i \leq m$ if $f_0 \in \Psi$. We substitute $p_0$ with $f_0(\theta)$ and get $\text{RSS}(\phi) = \text{RSS}(f_0(\theta_1), \ldots, f_0(\theta_m))$. Hence, Equation (8) becomes an optimization problem about the parameter $\phi$:

$$\min_{\phi} \text{RSS}(\phi) = \min_{\phi} \text{RSS}(f_0(\theta_1), \ldots, f_0(\theta_m))$$

Due to the constraint we mentioned before ($0 = p_{16} \leq p_{20} \leq \ldots \leq p_{237729-1} = p_{237729} = 1$), to determine an appropriate curve family, the following conditions (for $\forall \phi \in \Psi$) should be satisfied:

1. $f_\phi(0) = 0$ and $f_\phi(1) = 1$
2. $f_\phi(x)$ should be non-decreasing function in $[0, 1]$.

The curve patterns in Figure 3a lead us to choose the regularized beta function as the curve family. The regularized beta function $L_{a,b}(x)$, with the parameters $a, b$, is:

$$L_{a,b}(x) = \int_{0}^{x} \frac{e^{a} - (1 - y)^{a-1} dy}{a b} \left( a > 0, 0 \leq x \leq 1 \right)$$

Beta Ratio can be defined as $BR = \frac{\log_{2}(\hat{\mu}_{B, blood}) - \log_{2}(\hat{\mu}_{B, placenta})}{\log_{2}(\hat{\mu}_{A, blood}) - \log_{2}(\hat{\mu}_{A, placenta})}$, which considers each of the dilution series. It is more reasonable than the traditional fold change derived from the pure samples, and some discussions were shown in Figure S1 of Supplementary Material. If the intensity data have been log2 transformed, define FC as $\log_{2}(\hat{\mu}_{B, blood}) - \log_{2}(\hat{\mu}_{A, blood})$, otherwise, define FC as $\log_{2}(\hat{\mu}_{B, blood}) - \log_{2}(\hat{\mu}_{A, placenta})$ for non-log2 data.

2.2.3 The estimated fold change
Now we can get the estimators of AP $\hat{p}_1 \cdots \hat{p}_{237729}$ from the solution of Equation (9), and substitute $\hat{p}_1 \cdots \hat{p}_{237729}$ with $\hat{p}_1 \cdots \hat{p}_{237729}$ to calculate the design matrix $X_g$ in Equation (5). The true intensity vector $\hat{\beta}$ will be derived from Equation (7).

The estimated fold change (FC$_g$) derived from $\hat{\mu}_{B, blood}$ and $\hat{\mu}_{B, placenta}$, considers each of the dilution series. It is more reasonable than the traditional fold change derived from the pure samples, and some discussions were shown in Figure S1 of Supplementary Material. If the intensity data have been log2 transformed, define FC$_g = \log_{2}(\hat{\mu}_{B, blood}) - \log_{2}(\hat{\mu}_{A, blood})$, otherwise, define FC$_g = \log_{2}(\hat{\mu}_{B, blood}) - \log_{2}(\hat{\mu}_{A, placenta})$ for non-log2 data.

2.3 Effect of log2-transform on beta ratio and adjusted fold change

It is a normal practice that the microarray data is log2 transformed before data analysis. BR of log2-transformed data is different from that of non-log2 transformed data. We tried a method to adjust log ratio (i.e. FC), based on the relationship between BR and log2-transformed data.

In the perfect situation, the measurement function is $f(y) = cy$ where $c$ is a positive constant, and AP will be accurately equal to DP. If the data has been log2 transformed, the measurement function will become $f(y) = \log_{2}(f(y)) = \log_{2}(y) + \log_{2}(c)$, and FC satisfies $FC_g = \log_{2}(\mu_{B, placenta}/\mu_{A, blood}) = \log_{2}(\hat{\mu}_{B, placenta}/\hat{\mu}_{A, blood})$. The formula of AP according to the new measurement function $f$ will be:

$$AP_g(\theta) = \frac{\log_{2}[(1 - \theta) + \theta \hat{\mu}_{B, placenta}/\hat{\mu}_{A, blood}]}{\log_{2}(\hat{\mu}_{B, placenta}/\hat{\mu}_{A, blood})} = \frac{\log_{2}[1 - \theta + \theta \cdot 2^{FC_g}]}{2^{FC_g}}$$

for $FC_g \neq 0$, where $\theta$ is DP. The DP-vs-AP curve family was shown in Figure 4a.

To derive the relationship between BR and FC$_g$ under log2 transform, for each FC$_g$, we find a regularized beta function $L_{a,b}(\theta)$, which is closest to $AP_g(\theta)$, defined in $L^2$ space. It is an optimization problem with the parameters $a, b$ as:

$$\min_{a,b} \left[ A_{a,b}^2 = \min_{a,b} \int_{0}^{1} \left( L_{a,b}(\theta) - AP_g(\theta) \right)^2 d\theta \right]$$

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/23/11/1339/199605/Modeling-nonlinearity-in-dilution-design by guest on 03 October 2017
3 RESULTS AND DISCUSSION

A key assumption in the microarray data analysis is that the quantified signal intensities are linearly related to the expression of the corresponding genes in the target sample (Ramdas et al., 2001). However, as pointed out by (Ramdas et al., 2001), there are sources of nonlinearity in microarray expression measurements. Dilution design (mixed tissue RNA design) has been utilized by some researchers to evaluate and assess the performance of multiple microarray platforms (Barnes et al., 2005; MAQC Consortium, 2006; Thompson et al., 2005). RNA mixtures can be used to measure the gene transcription levels in different dilution profiles and measure the linear range as a function of both signal intensity and fold change; deviations from linearity can result from signal noise and assay imprecision as well as from signal plateauing at upper or lower limits of linearity out of dynamic range (Thompson et al., 2005).

We proposed proportion model to describe the nonlinearity phenomenon in dilution design of microarray data. A calibration value (BR) and adj-FC were calculated to approximately quantify and correct the nonlinearity respectively, based on log2-transformed data. We applied proportion model to the normalized log2-transformed data from AFX and ILM prepared by Barnes et al. (2005) respectively. By the definition of AP from Equation (2), we determined a DEG list via ANOVA (Zar, 1999) of which each gene had statistical significance according to the null hypothesis $\mu_{g, blood} = \mu_{g, placenta}$. As suggested by MAQC Consortium (2006) 0.001 was selected as $P$-value threshold.

3.1 Nonlinear characterization and adjusted FC

Figure 5 shows the relationship between BR and FC when the log2 transformed data was analyzed. The red lines denote the perfect BR-vs-FC curve, around which the genes should be plotted if the gene intensities in different samples follow linear relationship. The deviation of the genes’ data points from the red curves can be observed clearly. Most of the data points have higher BR values than they should have in the linear assumption (i.e. the corresponding values on the red curve), which indicates that there are global biases from the linear assumption to some extent. The APs for non-log transform according to 5% DP also indicate global biases (shown in Fig. S2 of Supplementary Material); and the average APs for all DEGs were greater than the corresponding DPs, which were shown in Figure S3 of Supplementary Material. The similar comparison and analysis have been reported by Shippy et al. (2006). Furthermore, such global biases imply that placenta RNA contributes more to the mixed intensity than blood RNA.

From Figure 5, we can also observe that the genes with small FC (e.g. close to 0) distributed randomly without any pattern. It is widely accepted that the genes with small FC are usually not DEGs. When corresponding BRs are calculated for such genes, random pattern will dramatically influence the effectiveness and robustness of the adj-FCs.

Table 1 and Table 2 show the top 100 FC-sorted genes from AFX and ILM platforms respectively, and those genes were ranked by their adj-FCs shown in the tables. ‘Symbol’ and ‘Name’ columns were obtained from the ‘WebGestalt’ web site (http://bioinfo.vanderbilt.edu/webgestalt/) in Table 1. Dudley et al. (2002) and Yuen et al. (2002) has pointed out that...
intensity ratio underestimation is a well-known phenomenon of microarray data. Notably, there is a significant bias in log ratios calculated from intensities: absolute log ratios are dramatically underestimated compared to truth, in particular for large fold change (Shi et al., 2004). Hence, it is not surprising that the adj-FCs are significantly greater than FCs shown in Tables 1 and 2.

3.2 ‘Poison’ Effect

Further analysis of those highly expressed genes with large absolute value of BRs shown in Figure 5 reveals that some of the genes with large FC exhibit highly significant nonlinearity. The adj-FCs of the top-rank genes shown in Table 1 and Table 2 are too large comparing with the other genes, which is what we call ‘poison’ effect.

The twelve target genes, identified by Affymetrix are shown to have the ‘poison’ effect (i.e. the genes with red points on Fig. 5a and listed in Table 1), were submitted to the ‘WebGestalt’ web site for further analyses. It turned out that these twelve target genes map to three major gene transcripts (i.e. CSH1, CSH2 and CSHL1 in Table 1). The two target genes identified by Illumina in Table 2 were also found to map to CSH1 and CSH2 via WebGestalt. CSH1 encodes for chorionic somatomammotropin hormone 1, CSH2 encodes for chorionic somatomammotropin hormone 2, and CSHL1 encodes for chorionic somatomammotropin hormone-like 1 proteins; these encoded protein isoforms belong to the same protein family called ‘placental lactogen’, which is a placental form of growth hormone variant. Chorionic somatomammotropin hormones 1 and 2 are upregulated and expressed to the identical mature proteins during development (Ng et al., 2003; Selvanayagam et al., 1984). Human placental lactogen (hPL), a potent glycoprotein, is made throughout gestation, increasing progressively until the 36th week, where it can be found in the maternal serum at a concentration of 5–15 μg/ml, the highest concentration of any known protein hormone (Kliman et al., 1988). This particular family member of genes (CSH1, CSH2 and CSHL1) is expressed mainly in the placenta and utilizes multiple transcription initiation sites.

This feature of ‘multiple transcription initiation’ can be the major reason for the ‘highest concentration seen for human protein hormones of hPL’ and ‘poison’ effect seen on these expressed target genes in our analyses. Due to ‘multiple initiation sites of transcription’, CSH1 or CSH2 can be found highly expressed even in a sample mixture with little proportion of placenta RNA (e.g. 5% placenta RNA mixture); also, because of high concentrations and the digital saturation of the scanner, the intensity no longer increases linearly with concentration (Shi et al., 2004). Furthermore, the similarity of the sequences of the target genes (CSH1, CSH2 and CSHL1) in the RNA mixture may also enhance the intensity by cross hybridizations to their complementary probe genes in the microarray plate. Nonetheless, in either case, it confirms ‘poison’ effect of the gene expression for some special genes.

The adj-FC of those genes with ‘poison’ effect provides a way to estimate the approximate true FC in quantity without nonlinearity. However, since we do not know the true values, it is difficult to verify the estimates directly. Some studies correcting underestimation by other specific ways have been reported. Hekstra et al. (2003), Naef et al. (2003) and Held et al. (2003) addressed the problem of sequence-specific response of fluorescent signal as a function of concentration, and proposed ways to correct ratio underestimation for genes with high-fold changes observed in Affymetrix chips based on Langmuir adsorption and free-energy calculations, respectively (Shi et al., 2004).
Modeling nonlinearity in dilution design

Table 1. Ranking of the top 100 FC-sorted genes by their adj-FCs for AFX platform, under 0.001 P-value threshold (only show the top 18 genes, the complete table is available in Table S1 of Supplementary Material.)

| Rank | AFX- Probe | Adj-FC | FC   | Beta Ratio | P-value | Symbol | Name | 1
|------|------------|--------|------|------------|---------|--------|------|------
| 1    | 211739_x_at | 570.84 | 8.25 | 0.99495   | 8.71E-05 | CSH1   | chorionic somatomammotropin hormone 1 (placental lactogen) |
| 2    | 203807_x_at | 463.22 | 8.36 | 0.99377   | 7.16E-05 | CSH2   | chorionic somatomammotropin hormone 2 |
| 3    | 202493_x_at | 397.97 | 8.48 | 0.99275   | 0.000228 | CSH1   | chorionic somatomammotropin hormone 1 (placental lactogen) |
| 4    | 208342_x_at | 273.89 | 8.21 | 0.98947   | 4.22E-06 | CSH2   | chorionic somatomammotropin hormone 2 |
| 5    | 208357_x_at | 207.99 | 8.40 | 0.98614   | 5.58E-05 | CSH1   | chorionic somatomammotropin hormone 1 (placental lactogen) |
| 6    | 208295_x_at | 192.39 | 8.74 | 0.98502   | 1.94E-05 | CSHL1  | chorionic somatomammotropin hormone-like 1 |
| 7    | 207285_x_at | 163.44 | 8.10 | 0.98236   | 1.34E-05 | CSHL1  | chorionic somatomammotropin hormone-like |
| 8    | 208293_x_at | 158.18 | 8.31 | 0.98178   | 5.33E-06 | CSH1   | chorionic somatomammotropin hormone 1 (placental lactogen) |
| 9    | 208356_x_at | 156.47 | 8.51 | 0.98158   | 1.50E-05 | CSH1   | chorionic somatomammotropin hormone 1 (placental lactogen) |
| 10   | 205958_x_at | 134.69 | 8.49 | 0.97860   | 1.97E-06 | CSHL1  | chorionic somatomammotropin hormone-like 1 |
| 11   | 208341_x_at | 130.41 | 8.61 | 0.97790   | 4.38E-05 | CSH2   | chorionic somatomammotropin hormone 2 |
| 12   | 208294_x_at | 124.63 | 7.86 | 0.97688   | 1.85E-06 | CSHL1  | chorionic somatomammotropin hormone-like 1 |
| 13   | 207770_x_at | 61.44  | 7.88 | 0.95317   | 5.65E-06 | CSH2   | chorionic somatomammotropin hormone 2 |
| 14   | 192978_x_at | 59.36  | 10.19| 0.95152   | 2.27E-10 | TFPI2  | tissue factor pathway inhibitor 2 |
| 15   | 208069_x_at | 56.90  | 8.55 | 0.94944   | 1.04E-07 | NA     |  |
| 16   | 206885_x_at | 52.59  | 7.99 | 0.94531   | 3.99E-09 | GH1    | growth hormone 1 |
| 17   | 206475_x_at | 50.99  | 8.31 | 0.94360   | 3.66E-06 | CSH2   | chorionic somatomammotropin hormone 2 |
| 18   | 208294_x_at | 211151_x_at | 45.83 | 8.78 | 0.93728   | 2.94E-06 | GH1    | growth hormone 1 |

1Obtained from the ‘WebGestalt’ web site (http://bioinfo.vanderbilt.edu/webgestalt/)
2The genes with ‘poison’ effect, which belong to the same protein family called ‘placental lactogen’; they are translated to become human placental lactogen which has the highest concentration known in pregnancy and behave with the feature of multiple initiation sites of transcription.
3The gene with the largest FC.

Table 2. Ranking of the top 100 FC-sorted genes by their adj-FCs for ILM platform, under 0.001 P-value threshold (only show the top eight genes, the complete table is available in Table S2 of Supplementary Material.)

<table>
<thead>
<tr>
<th>Rank</th>
<th>ILM-Probe</th>
<th>Adj-FC</th>
<th>FC</th>
<th>Beta ratio</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GI_20819996-A</td>
<td>304.99</td>
<td>7.05</td>
<td>0.99054</td>
<td>9.81E-05</td>
</tr>
<tr>
<td>2</td>
<td>GI_20819972-A</td>
<td>93.76</td>
<td>6.66</td>
<td>0.96928</td>
<td>0.000443</td>
</tr>
<tr>
<td>3</td>
<td>GI_20809252-A</td>
<td>26.43</td>
<td>8.10</td>
<td>0.89158</td>
<td>1.17E-06</td>
</tr>
<tr>
<td>4</td>
<td>GI_31543803-S</td>
<td>24.93</td>
<td>8.35</td>
<td>0.88509</td>
<td>9.45E-08</td>
</tr>
<tr>
<td>5</td>
<td>GI_10800407-S</td>
<td>18.30</td>
<td>8.14</td>
<td>0.84401</td>
<td>1.64E-08</td>
</tr>
<tr>
<td>6</td>
<td>GI_13702710-S</td>
<td>18.23</td>
<td>8.44</td>
<td>0.84338</td>
<td>1.97E-08</td>
</tr>
<tr>
<td>7</td>
<td>GI_21361295-S</td>
<td>17.02</td>
<td>7.90</td>
<td>0.83239</td>
<td>1.70E-08</td>
</tr>
<tr>
<td>8</td>
<td>GI_38045914-S</td>
<td>17.01</td>
<td>8.07</td>
<td>0.83227</td>
<td>1.01E-07</td>
</tr>
</tbody>
</table>

1The genes with ‘poison’ effect, map to CSH1 and CSH2 which belong to the same protein family called ‘placental lactogen’.
2The gene with the largest FC.

3.3 Consistency of adj-FC

Multiple commercial microarray platforms are widely used for genome-scale gene expression analysis (Shippy et al., 2004). Microarray manufacturers provide different underlying technologies and protocols for data analysis (Barnes et al., 2005; Draghici et al., 2006; Shippy et al., 2004; Thompson et al., 2005). Conflicting results have been reported on cross-platform reproducibility and consistency (Bammler et al., 2005; Draghici et al., 2006; Irizarry et al., 2005; Miller et al., 2004; Barnes et al., 2005) experimentally compared and validated Affymetrix and Illumina platforms. The results of platform comparison indicated high agreement between these two platforms, especially for the DEGs. Our results confirm this conclusion. The shapes of the BR-vs-FC figures derived from the two platforms are very similar (Fig. 5).

Percentage of overlapping genes (POG) has been reported to be an effective metric for cross-platform consistency analysis (Shi et al., 2005). POG represents the number of genes common in the two ‘significant’ gene lists (derived from the two platforms, AFX and ILM), divided by L, the number of genes in the selected gene list. The higher the POG values, the more consistent the two platforms are. Shi et al. (2005) found that ranking genes by their FC yields a better POG curve than ranking genes by their adj-FCs. The results from MAQC Consortium (2006) had also indicated that a straightforward approach of FC ranking plus a non-stringent P-value cutoff can be successful in identifying reproducible gene lists. In this study, we intend to test if adj-FC can improve the POG curve and the cross-platform consistency via 0.001 P-value threshold.

In microarray experiments, many probes are not specific for a single transcript of a gene. Probes that did not map to a ‘known’ gene, or to genes represented on only one platform (Barnes et al., 2005). In this article, we only consider the genes in the common gene list in POG analysis. The significant (P-value < 0.001) genes were sorted by their fold changes, and top 1500 genes from each platform were selected. In Figure 6, two POG curves were...
plotted: one was directly derived from the 1500 FC-sorted significant genes; the other was generated from these 1500 genes after they were sorted by adj-FC. L was set as 1, 2, 3, . . . , 1500. Figure 6 indicates that adj-FC is more consistent than simple fold change to rank significant genes between two platforms, especially for the top 500 genes. When the genes after 1500 ranking were considered, these two POG curves are similar. As more candidate genes were selected, more noise and errors are also introduced because more genes with small FC were included, which affected BRs and finally influence adj-FCs.

3.4 Robustness and Application

When the log2 unnormalized data from AFX and ILM platforms were analyzed, their BR-vs-FC figures were very similar to Figure 5. We also identified 12 and 2 genes with ‘poison’ effect respectively from AFX platform and ILM platform. It should be noted that the different data processing techniques, such as PLIER, RMA, GCRMA, yield different numbers of genes showing significant deviations in expression values between samples A and B but normalization methods are not the primary causes for the deviation from the linear assumption (Shippy et al., 2006). Our results indicated that BRs were consistent between normalized data and unnormalized data especially for those genes with large FCs. Furthermore, the number of dilution series can be one of the important factors influencing robustness of BR. We suggest more dilution series and symmetrical dilution design in the microarray experiment, which can improve the analysis of proportion model.

The adj-FC from proportion model predicted the real level of gene expression outside the dynamic range of the current array technology, and also enhanced the reproducibility measure of differentially expressed genes (DEGs) among different platforms, especially for the top and highly significant DEGs that are of biologist’s interests. Furthermore, the BR-versus-FC figure can be a calibration metric to test whether the quantified signal intensities are linearly related to the gene expressions.

4 CONCLUSION AND FUTURE WORK

Our reanalysis of the dataset of Barnes et al. (2005) indicated that there are global biases comparing with the linear assumption to some extent. BR-vs-FC figure can clearly illustrate the distribution of BR and FC, which may deviate the perfect curve of linear assumption. Genes with ‘poison’ effect, those large absolute values of FC with BR close to 1 or −1, also can be clearly shown in BR-vs-FC figure. Moreover, their adj-FCs imply some inherited feature of the genes, such as ‘multiple initiation sites of transcription’. The comparison of different POGs between adj-FC and simple FC indicates that adj-FC outperformed simple FC to show that AFX and ILM platforms are consistent, via 0.001 P-value threshold.

Figure 5 showed that BRs distributed randomly without any pattern for those genes with small FC, although their P-values were lower than 0.001. The variance of BR or AP can be considered as one of the important factors, when we develop a specific method to estimate P-value for the proportion model in future. Moreover, BR from the regularized beta function is an approximate metric, and may be not enough for the future work.

ACKNOWLEDGEMENTS

We thank Michael Barnes and Paul Pavlidis for sharing the dilution design microarray datasets. We also would like to thank Michael Baron from Department of Mathematical Sciences in University of Texas at Dallas for critical reading of the manuscript.

Conflict of Interest: none declared.

REFERENCES


Shi,L. et al. (2005) Cross-platform comparability of microarray technology: intraplatform consistency and appropriate data analysis procedures are essential. BMC Bioinformatics, 6 (Suppl. 2), S12.