BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 18 2007, pages 2391–2398
doi:10.1093/bioinformatics/btm361

Gene expression

Selection and validation of normalization methods for c-DNA microarrays using within-array replications

Jianqing Fan* and Yue Niu

Department of Operations Research and Financial Engineering Princeton University, Princeton, NJ 08544, USA

Received on February 22, 2007; revised on June 26, 2007; accepted on July 9, 2007

Advance Access publication July 27, 2007

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Normalization of microarray data is essential for multiple-array analyses. Several normalization protocols have been proposed based on different biological or statistical assumptions. A fundamental problem arises whether they have effectively normalized arrays. In addition, for a given array, the question arises how to choose a method to most effectively normalize the microarray data.

Results: We propose several techniques to compare the effectiveness of different normalization methods. We approach the problem by constructing statistics to test whether there are any systematic biases in the expression profiles among duplicated spots within an array. The test statistics involve estimating the genewise variances. This is accomplished by using several novel methods, including empirical Bayes methods for moderating the genewise variances and the smoothing methods for aggregating variance information. *P*-values are estimated based on a normal or χ approximation. With estimated *P*-values, we can choose a most appropriate method to normalize a specific array and assess the extent to which the systematic biases due to the variations of experimental conditions have been removed. The effectiveness and validity of the proposed methods are convincingly illustrated by a carefully designed simulation study. The method is further illustrated by an application to human placenta cDNAs comprising a large number of clones with replications, a customized microarray experiment carrying just a few hundred genes on the study of the molecular roles of Interferons on tumor, and the Agilent microarrays carrying tens of thousands of total RNA samples in the MAQC project on the study of reproducibility, sensitivity and specificity of the data.

Availability: Code to implement the method in the statistical package R is available from the authors.

Contact: jqfan@princeton.edu

1 INTRODUCTION

Microarray techniques have been widely used in many areas of biological research. They have substantial impact on tumor diagnostics, and classification and understanding of the molecular mechanisms of biochemical processes, tumorigenesis and tumor developments. Proper statistical analysis is vital for revealing meaningful biological results. For an overview of statistical analysis of DNA microarrays, we refer to Fan and Ren (2006) and references therein.

*To whom correspondence should be addressed.

The quality of microarray data is very important for downstream statistical analysis (Eisenstein, 2006; Marshall, 2004; Patterson *et al*., 2006; Shi *et al*., 2006). Experimental variations, such as RNA quality, probe labeling, hybridization condition, washing, signal and background detection in the scanning process, slide and block effects, pose significant challenges in the analysis of microarray data. The first step in microarray analysis is to remove the systematic biases due to the variations in experimental conditions so as to make multiple array analyses meaningful. These efforts are collectively referred to as the normalization of microarray data in the literature.

A number of useful normalization protocols have been proposed based on different assumptions. These include the global normalization (Kroll and Wölfl, 2002), rank invariant normalization (Tseng *et al*., 2001), LOWESS normalization (Dudoit *et al*., 2002), Semi-Linear In-slide Model (SLIM, Fan *et al*., 2004, 2005), Two-way Semi-Linear Models (TW-SLM, Huang *et al*., 2005), robust TW-SLM (Ma *et al*., 2006; Wang *et al*., 2005) and normalization of small diagnostic microarrays (Jaeger and Spang, 2006). Recently, two seminal normalization methods (CADS and eCADS) based on dye swap and statistical models are proposed by Dabney and Storey (2007a, b). All of the aforementioned methods are based on some statistical and biological assumptions. For example, the global normalization is adequate only when there is no print-tip block effect and no intensity effect; the LOWESS method assumes implicitly that at each given intensive level the average expression level of up- and down-regulated genes are about the same in each print-tip block; SLIM, TW-SLM, robust TW-SLIM all require that the statistical models are correct. The questions then arise naturally for a given array, which methods are most effective to normalize the data and whether the data have been properly normalized. These questions are very fundamental to the statistical analysis of microarray and have not yet been addressed.

Our study is motivated by the aforementioned fundamental concerns. Our approach is to use the duplicated spots within an array. They contain the most valuable information about possible systematic biases in the microarray experiments. The idea is that if microarray data have been properly normalized, there should be no systematic biases among duplicated spots. Therefore, when the sum of the square differences of duplicated spots, standardized by the estimated genewise variances, are aggregated among many different genes having duplications, the test statistic follows approximately a χ^2-distribution with a large degree of freedom, or more formally a normal

distribution. This provides a simple and useful diagnostic test statistic to check if an array has been properly normalized by a particular method. Regarding the test statistic as a measure of the discrepancy of replicated spots after normalization, we select a normalization method that has the smallest value of the test statistics. In addition, the associated *P*-value of the test statistic enables us to judge the degree to which the normalization has been properly carried out.

In implementing the validation tests, it involves inevitably the estimation of genewise SDs and variances. A precise estimate of SDs and variances will improve the statistical power of the validation tests. It has also important applications in selecting significant genes (Cui *et al*., 2005; Dudoit *et al*., 2003; Reiner *et al*., 2003; Storey and Tibshirani, 2003; Symth *et al*., 2005; Tusher *et al*., 2001;). The precision of estimating genewise SDs and variances depends on the number of replications that are available. Several innovative strategies will be introduced to enhance the precision of the estimate.

Duplicated spots play very important roles in the analysis of microarray data. They are not only powerful for normalization (Fan *et al*., 2004), but also useful for genewise variance estimation (Smyth *et al*., 2005) in selecting statistically differently expressed genes. Furthermore, they are fundamental to our proposed validation tests. The availability of such duplicated genes can be accomplished by the designs of c-DNA microarrays. For example, in the microarrays analyzed by Fan *et al*. (2004), 111 out of 19 968 clones of genes were printed twice randomly on the 32 print-tip blocks. This enables them to untangle the block effects and intensity effects from these 111 duplicated spots. With the increased popularity of customized microarrays, which enables researchers to focus only on hundreds of genes of their primary interest with more reliable measurements, within-array replications can easily be obtained. The gene selection biases in customized arrays require more sophisticated normalization techniques. The validation tests are essential for controlling the quality of downstream statistical data analysis of customized arrays.

2 METHODS OF NORMALIZATION

There are several useful normalization methods, which are based on different biological or statistical assumptions. We briefly review several of them that will be used in our numerical studies.

Suppose that we have J replications of a c-DNA microarray experiment. For each given array, there are N different genes. Among them, G genes are replicated I times. Let I_g denote the number of replications for gene g, with $I_g = 1$ indicating no duplication. For customized arrays, usually $G = N$ (all genes have within-array replications) and $I_g = 2$ or 3 or 4. However, this can also be designed differently. For the microarrays analyzed in Fan *et al*. (2004), $G = 111$ and $N = 19\,968 - G$ so that $I_g = 2$ only for 111 clones that are duplicated and randomly placed on 32 print-tip blocks.

Let R_{gij} and G_{gij} be respectively the intensities of red (Cy3) and green (Cy5) channels for the ith duplication of the gth gene in the jth array, and b_{gi} be the print-tip block where the ith duplication of the gth gene resides. Then, we can compute the log-ratios and log-intensities as

$$Y_{gij} = \log_2(G_{gij}/R_{gij}), \quad \textit{and} \quad X_{gij} = 0.5\log_2(G_{gij} * R_{gij}).$$

The global normalization is to compute the median $\hat{m}_j$ of log-ratios $\{Y_{gij}\}$ of the jth array and to normalize the data for the jth array as

$$\hat{Y}_{gij} = Y_{gij} - \hat{m}_j. \tag{1}$$

This basically assumes that the up-regulated and down-regulated genes are about the same, which does not usually hold for customized arrays due to gene selection biases. In addition, the global normalization assumes no block or intensity effects.

To address the above two concerns, Dudoit *et al*. (2002) apply the global normalization technique more locally to each block and each intensity level, resulting in the LOWESS normalization. For the data $\{(X_{gij}, Y_{gij}) : b_{gi} = b\}$ in a given print-tip block b of the jth array, they applied the LOWESS smoother to estimate the conditional mean function $\hat{m}_{bj}(x)$ of the log-ratios Y given the intensity level $X = x$, and computed the normalized log-ratios in the bth block in jth array as follows:

$$\hat{Y}_{gij} = Y_{gij} - \hat{m}_{bj}(X_{gij}), \qquad \textit{with } b_{gi} = b. \tag{2}$$

This significantly relaxes the restrictions of the global normalization, but still assumes that up-regulated and down-regulated genes are about the same at each given intensity level. Again, this might not be appropriate when the cells or tissues are treated by cytokines. It might not be valid for customized arrays due to gene selection biases.

To address these issues, Fan *et al*. (2004) introduced the SLIM technique based on the model assumption:

$$Y_{gij} = \alpha_g + \beta_{j,b_{gi}} + m_j(X_{gij}) + \varepsilon_{gij}, \tag{3}$$

in which α_g is the treatment effect on gene g, $\boldsymbol{\beta}_j$ and m_j represent respectively the array-specific block and intensity effect and ε is the stochastic noise. For each given array j, Fan *et al*. (2004) estimated the model parameters using the data from duplicated spots and computed the normalized data

$$\hat{Y}_{gij} = Y_{gij} - \hat{\beta}_{j,b_{gi}} - \hat{m}_j(X_{gij}). \tag{4}$$

In Fan *et al*. (2005), the efficiency of parameters in (4) is improved by aggregating the data from all arrays. Namely, the parameters in (3) are jointly estimated using all arrays.

TW-SLM (Huang *et al*., 2005) and robust TW-SLM (Ma *et al*., 2006; Wang *et al*., 2005) are in the same spirit as SLIM. It does not require duplications of spots but relies heavily on the assumption that the treatment effect α_g is independent of the array. Statistically, it is a specific model of (3) with $\beta = 0$. With estimated parameters, the normalized data are computed as in (4) with $\hat{\beta}_{j,b_{gi}} = 0$. TW-SLM applies the technique for each given block so that the block effects in each array have been properly taken care of.

The effectiveness of SLIM, TW-SLM and robust TW-SLM depends critically on the statistical model assumption. With so many normalization methods, the questions arise naturally which method is the most appropriate for a specific array and whether the expression profiles have been properly normalized.

The CADS and eCADS (Dabney and Storey, 2007a, b) are the normalization methods based on dye-swap and statistical models. They aim at preserving time differential expression relationships among the treatment and control arrays after normalization.

3 VALIDATION TESTS

Within array replications not only provide useful information about the possible systematic biases: block effect and intensity effect, but also play an important role in validation tests of the necessity and effectiveness of the normalization methods. When the log-ratios have been properly normalized, the systematic biases should be negligible and hence the following model holds approximately:

$$\hat{Y}_{gij} = \alpha_g + \varepsilon_{gij}, \tag{5}$$

$$g = 1, \ldots, N, \quad i = 1, \ldots, I_g, \quad j = 1, \ldots, J,$$

where I_g is the number of duplication for gene g. See also (4).

Duplicated spots provide valuable information about the validity of model (5) for each given array. We will use only G genes with I replications for each given array to develop the validation tests, and drop the subscript j to facilitate the notation. This leads to the simplified notation:

$$\hat{Y}_{gi} = \alpha_g + \varepsilon_{gi}, \qquad g = 1, \ldots, G, \quad i = 1, \ldots, I. \tag{6}$$

Hence, the difference $\hat{Y}_{gi} - \bar{Y}_g$ should have mean zero, where $\bar{Y}_g = \sum_{i=1}^{I} \hat{Y}_{gi}/I$. We will assume that the first four moments of the noise behave like those of a normal distribution:

$$E\varepsilon_{gi} = 0, \quad E\varepsilon_{gi}^2 = \sigma_g^2, \quad E\varepsilon_{gi}^3 = 0, \quad E\varepsilon_{gi}^4 = 3\sigma_g^4.$$

3.1 Genewise standardization

Two natural test statistics for testing model (6) are based on the genewise standardized L_2- and L_1-norm, aggregated over G genes having duplications. Specifically, we let

$$T_1 = \sum_{g=1}^{G} \Big\{ \sum_{i=1}^{I} (\hat{Y}_{gi} - \bar{Y}_g)^2 / \sigma_g^2 \Big\}. \tag{7}$$

Under the normality assumption $\varepsilon_{gi} \sim N(0, \sigma_g^2)$, if the data have been properly normalized, the test statistic T_1 follows the χ^2-distribution with degree of freedom $(I-1)G$. In particular, when $I = 2$,

$$T_1 = \sum_{g=1}^{G} (\hat{Y}_{g1} - \hat{Y}_{g2})^2 / (2\sigma_g^2) \sim \chi_G^2.$$

Note that the test statistic above is reasonably robust to the normality assumption. In the validation test, the number of genes with duplications G is reasonably large and by the Central Limit Theorem, T_1 follows approximately a normal distribution. In this case, $\chi^2_{(I-1)G}$ is also approximately a normal distribution. Hence, the null distribution $\chi^2_{(I-1)G}$ is approximately valid.

Instead of using the L_2-norm, one can use a more robustified norm L_1-norm. This leads naturally to consider the test based on the genewise standardized L_1-norm:

$$T_2 = \sum_{g=1}^{G} \Big\{ \sum_{i=1}^{I} \Big| \hat{Y}_{gi} - \bar{Y}_g \Big| / \sigma_g \Big\}. \tag{8}$$

In particular, when $I = 2$, the test statistic reduces to

$$T_2 = \sum_{g=1}^{G} \Big| \hat{Y}_{g1} - \hat{Y}_{g2} \Big| / \sigma_g.$$

By the Central Limit Theorem,

$$T_2 \overset{a}{\sim} N(\xi_I G, \eta_I^2 G), \tag{9}$$

where $\xi_I = \sqrt{2I(I-1)/\pi}$ and $\eta_I^2 = \mathrm{var}\Big(\sum_{i=1}^{I} |\hat{Y}_{gi} - \bar{Y}_g| / \sigma_g \Big)$. Specifically,

$$\eta_I^2 = \begin{cases} 2(1 - 2/\pi) = 0.7268, & \textit{when } I = 2 \\ (4\sqrt{3} - 12)/\pi + 8/3 = 1.0523, & \textit{when } I = 3. \end{cases}$$

3.2 Aggregated standardization

Accurate estimates of the genewise SD σ_g are challenging. The process itself may depend on the selection of a method of normalization. The test statistics T_1 and T_2 may not be well approximated by the χ^2-distribution or the normal distribution, when the genewise variances are not estimated accurately. In addition, the standardized square differences in T_1 and T_2 are sensitive to the estimation error in σ_g. On the other hand, the aggregated variance $G^{-1} \sum_{g=1}^{G} \sigma_g^2$ or aggregated SD $G^{-1} \sum_{g=1}^{G} \sigma_g$ can be more accurately estimated. These considerations lead us the unweighted differences among the expression profiles of replicated spots.

The aggregated differences among replicated spots when standardized, are given by

$$T_3 = \frac{\sum_{g=1}^{G} \sum_{i=1}^{I} (\hat{Y}_{gi} - \bar{Y}_g)^2 - (I-1) \sum_{g=1}^{G} \sigma_g^2}{\sqrt{2(I-1) \sum_{g=1}^{G} \sigma_g^4}}. \tag{10}$$

The test statistic T_3 is obtained by aggregation first followed by standardization. On the other hand, the test statistic T_1 is obtained by standardization first and followed by aggregation. The null distribution of T_3 follows approximately $N(0, 1)$ when G is large.

Following the same spirit, a more robustified counterpart of T_2 is

$$T_4 = \Big\{ \sum_{g=1}^{G} \sum_{i=1}^{I} \Big| \hat{Y}_{gi} - \bar{Y}_g \Big| - \xi_I \sum_{g=1}^{G} \sigma_g \Big\} / \Big\{ \eta_I \Big(\sum_{g=1}^{G} \sigma_g^2 \Big)^{1/2} \Big\}. \tag{11}$$

where ξ_I and η_I are two constants defined in (10). The null distribution of T_4 follows approximately $N(0, 1)$, when the data are properly normalized.

Note that the test statistic T_4 involves the estimation of aggregated SD $\sum_{g=1}^{G} \sigma_g$. If this is estimated with bias, then the null distribution will be shifted. The genewise variance is usually estimated by the sample variance or aggregated sample variance S_g^2. Suppose that S_g^2 has the distribution $KS_g^2/\sigma_g^2 \sim \chi_K^2$ with a given degree of freedom K, then S_g^2 is an unbiased estimator of σ_g^2, but S_g is not an unbiased estimator of σ_g. An unbiased estimator of σ_g is given by (Gurland and Tripathi, 1971)

$$\{(K/2)^{1/2} \Gamma(K/2) / \Gamma((K+1)/2)\}\, S_g. \tag{12}$$

In our numerical studies, this is implemented in T_2 and T_4. Our experiences show that the correction factor in (12) is necessary in order to obtain a more accurate approximation of the null distribution.

3.3 Choosing a method of normalization

The test statistics $T_1, \ldots, T_4$ can be regarded as measures of effectiveness of normalization. The smaller, the less discrepancy among repeated measurements, and the more effectiveness of a normalization method. For a given array, among several normalization methods, we would choose the one that has the smallest test statistic. The associated P-value gives us an idea on the extent to which the expression profiles have been normalized. The power of the validation test depends on the number of data points G in the training set. Excessively, large G will result in overpower of the tests to reject even a tiny systematic bias.

As to be discussed in Section 4, the implementation of the validation tests might require the choice of an appropriate normalization method first in order to estimate the genewise variances. The aggregated standardization tests T_3 and T_4 can be used for this purpose, since the normalization constants can be ignored.

3.4 Training and testing sets

In many situations, there are many genes that have duplications. This is particularly the case for the customized arrays. In these situations, we can randomly select between 50 and 100 different genes as the testing set and use the remaining genes as the training set. The training set is used to estimate the parameters in the normalization, while the testing set is applied to the validation tests.

In other situations, there are limited genes that have duplicated spots. In this case, the multi-fold cross-validation ideas can be employed to choose the training and testing sets.

4 ESTIMATION OF GENEWISE VARIANCE

Accurate estimation of genewise variance σ_g^2 is important for assessing the effectiveness of normalization. It is also critically important for selecting the genes that are statistically differently expressed among treatments and controls (Cui *et al*., 2005; Fan and Ren 2006; Fan *et al*., 2004; Storey, and Tibshirani, 2003; Tusher, *et al*., 2001). In particular, Cui *et al*. (2005) demonstrates that genewise variance estimation has direct impact on the sensitivity and specificity of selecting differently expressed genes.

4.1 Use of within array replications

A natural estimate of the genewise variance is the sample variance of the duplicated expressions. These log-ratios are computed after the data are normalized. If there are several normalization methods available, one can use T_3 and T_4 without standardization to help select a method of normalization.

Suppose that we have J replicated arrays. If we assume that $\text{var}(\hat{Y}_{gij}) = \sigma_g^2$, which are the same across J arrays, then we would pool the variability information from other arrays. This leads to a pooled estimator of σ_g^2 by

$$s_{W,g}^2 = \frac{1}{J(I-1)} \sum_{j=1}^{J} \sum_{i=1}^{I} (\hat{Y}_{gij} - \bar{Y}_{gj})^2. \tag{13}$$

The above estimate ignores the correlation among duplicated genes. If within-array replications have a common correlation $\rho_g = \rho$ and observations across arrays are independent, Smyth *et al*. (2005) introduced the residual maximum likelihood (REML) estimator of σ_g^2 as follows:

$$s_g^2 = \frac{1}{IJ-1} \left\{ \frac{(J-1)s_{B,g}^2}{1+(I-1)\hat{\rho}} + \frac{J(I-1)s_{W,g}^2}{1-\hat{\rho}} \right\}, \tag{14}$$

where $s_{B,g}^2 = I \sum_{j=1}^{J} (\bar{Y}_{gj} - \bar{Y}_g)^2/(J-1)$ with $\bar{Y}_g = \sum_{j=1}^{J} \bar{Y}_{gj}/J$ is the between-arrays variance and $\hat{\rho}$ is an estimate of ρ. They also proposed the REML estimation of $\hat{\rho}$ by

$$\hat{\rho} = \frac{\sum_{g=1}^{G} s_{B,g}^2 - \sum_{g=1}^{G} s_{W,g}^2}{\sum_{g=1}^{G} s_{B,g}^2 + (I-1)\sum_{g=1}^{G} s_{W,g}^2}.$$

They argued further that the degree of freedom of s_g^2 is $(IJ-1)$.

4.2 Smoothing estimator

The degree of freedom for the within-array estimate $s_{W,g}^2$ of the variance is still limited. As the variability of c-DNA microarray measurements is related to the intensity level (Fan *et al*., 2004; Tseng *et al*., 2001), one can pool the information of variability from expression profiles with similar intensity levels (Fan *et al*., 2004). This results in a non-parametric estimate of the intensity-specific variance function $\sigma^2(\cdot)$ from the non-parametric regression model:

$$\hat{Y}_{gi} = \alpha_g + \sigma(X_{gi})\eta_{gi}, \quad i = 1, \ldots, I_g, \quad g = 1, \ldots, N. \tag{15}$$

See also (5) with the array index j suppressed. The procedure of Fan *et al*. (2004) is to first smooth $\{\hat{Y}_{gi}\}$ on $\{X_{gi}\}$ by using a local linear estimate, which is really a smoothing estimator of the scatter plot $\{(X_{gi}, \hat{Y}_{gi})\}$, to obtain an estimate of α_g by regarding it as a smooth function of the log-intensity X_{gi}, and then smooth the squared residuals $\{(\hat{Y}_{gi} - \hat{\alpha}_g)^2\}$ on $\{X_{gi}\}$ to obtain an estimate of the intensity-specific variance function $\sigma^2(\cdot)$. The fundamental assumption in Fan *et al*. (2004) is that α_g is basically a smooth function of X_{gi}. When this assumption is not valid, the estimator will be biased.

The bias issue in Fan *et al*. (2004) can be significantly reduced when the within-array replications are available as in (15). Consider those genes with I replications. Replacing α_g by its unbiased estimate $\bar{Y}_g = \sum_{i=1}^{I} \hat{Y}_{gi}/I$, we have the variance of the residual

$$E(\hat{Y}_{gi} - \bar{Y}_g)^2 = (I-1)^2\sigma^2(X_{gi})/I^2 + \sum_{j \neq i} \sigma^2(X_{gj})/I^2.$$

When the variability of log-intensities is not large, we can approximate X_{gj} by its average $\bar{X}_g$. Owing to the smoothness assumption, we have

$$\text{var}(\hat{Y}_{gi} - \bar{Y}_g) \approx (I-1)\sigma^2(\bar{X}_g)/I. \tag{16}$$

Letting $r_g^2 = \sum_{i=1}^{I} (\hat{Y}_{gi} - \bar{Y}_g)^2/(I-1)$ be the sample SD, we have

$$Er_g^2 \approx \sigma^2(\bar{X}_g).$$

Therefore, the intensity-specific variance $\sigma^2(\cdot)$ can be estimated by smoothing the pairs $\{(\bar{X}_g, r_g^2) : g = 1, \ldots, G\}$, resulting in an estimated function $\hat{\sigma}^2(\cdot)$. Hence, our estimate of σ_g^2 is

$$\hat{s}_g^2 = \hat{\sigma}^2(\bar{X}_g).$$

The approach is particularly appealing to the customized arrays, in which the variability of intensities is not large and the replicated spots are available. See Section 5.3.

4.3 Empirical Bayes estimator

With within-array replications, we have two estimators: REML s_g^2 and the intensity-specific estimator $\hat{s}_g^2$. One naturally uses the intensity-specific estimator $\hat{s}_g^2$ to augment the REML estimator s_g^2. One simple way to do this is the following empirical Bayes shrinkage estimator:

$$\tilde{s}_g^2 = \frac{(IJ-1)s_g^2 + d\hat{s}_g^2}{IJ-1+d}, \tag{17}$$

where d is the degree of freedom of $\hat{s}_g^2$. The estimator (17) is the Bayes estimator with the inverse Gamma prior with the prior mean $\hat{s}_g^2$ and prior degree of freedom d. Since the prior parameters are estimated from the data, (17) is indeed an empirical Bayes estimator.

The degree of freedom of the non-parametric estimator $\hat{s}_g^2$ is often estimated by the effective sample size, which is the inverse of the asymptotic variance of the kernel local linear regression estimator. For example, if the kernel local regression estimator is used, then, we have

$$d = \text{the number of local data}/(2\|K\|^2),$$

where K is the kernel function used in the smoothing . For example, in the LOWESS smoother, the kernel function is $\frac{70}{81}(1-|x|^3)^3 I(|x| \le 1)$.

Note that when the within-array variance $s^2_{W,g}$ in (13) is used in (17), the factor $(IJ-1)$ should be replaced by $I(J-1)$. In addition, the intensity-specific estimate of variance is usually not as reliable as the within-array variance $s^2_{W,g}$, the constant d can be replaced by some smaller numbers to reduce somewhat of its influence.

5 SIMULATIONS AND APPLICATIONS

In this section, we first use the simulated data set to illustrate the validity and the power of our proposed validation tests. In particular, the advantages of the aggregated standardization tests T_3 and T_4 are demonstrated. The three real data analyses illustrate the methodological power of our approaches.

5.1 Simulation

In each simulation, we generate $J=4$ arrays from model (3) with $G = N = 2000$ genes, each having $I=3$ replications randomly assigned over the 48 blocks. The details of simulation scheme for this example are summarized as follows:

$\boldsymbol{\alpha}_g$: The expression levels of the first 150 genes are generated from the standard double exponential distribution. The rest are 0's. These expression levels are the same over four arrays in each simulation, but may vary over simulations.

$\boldsymbol{\beta}$: The 48 parameters for the block effects $\boldsymbol{\beta}_j$ in array j are all set to $\boldsymbol{\beta}$, which is given by

$$\begin{aligned}\boldsymbol{\beta} =(&-0.15\ -0.3\ -0.15\ 0.01\ -0.01\ 0.04\ 0.07\ 0.08\ 0.08\ 0.23\\ &0.22\ 0.09\ 0.11\ 0.24\ 0.28\ 0.24\ 0.07\ 0.22\ 0.19\ 0.15\ -0.03\\ &-0.05\ -0.02\ 0.03\ -0.14\ -0.04\ -0.18\ 0.06\ -0.05\\ &-0.02\ -0.17\ -0.02\ 0.08\ 0.08\ -0.04\ -0.16\ 0.04\\ &0.01\ -0.05\ -0.12\ 0.07\ -0.19\ -0.03\ -0.15\\ &-0.07\ -0.22\ -0.11\ -0.22)^T.\end{aligned}$$

These parameter values are taken from the estimates in Ma *et al*. (2006).

X: The intensity is generated from a mixture distribution: with probability 0.8 from probability distribution $0.0004(16-x)^3 I(6<x<16)$ and 0.2 from the uniform distribution over [4, 16].

$m_j(\cdot)$: Set the function $m_j(X) = 5(0.3592 - \sqrt{(X-4)/32})$, whose expectation with respect to X is approximately zero.

b_g: For each given gene, its associated block is assigned at random at one of 48 print-tip blocks.

ε: ε_{gij} is generated from the standard normal distribution with mean zero and variance $\sigma^2(X_{gi}) = 0.15 + 0.015(X_{gi}-9)^2 I\{X_{gi}>9\}$.

This is a heteroscedastic model with small block effect and intensity effect. Three normalization methods are used: Global normalization, LOWESS normalization using all blocks and the SLIM normalization (Fan *et al*., 2004). They are applied to 200 pseudo-data sets, each having $J=4$ arrays, giving a total of 800 arrays.We drew 200 genes at random from the 2 000 different genes as the testing set, and the remaining genes as the learning set.

We first apply the validation test statistic T_3 with genewise variances estimated by the REML estimators (14) to check the effectiveness of the three normalization methods over 200 pseudo-data sets, which comprise of 800 pseudo-arrays. The distributions of the test statistic $|T_3|$ based on the three normalization methods are presented in Figure 1. They are well separated (Fig. 1a). First of all, T_3 for the global normalization is the same that with no normalization, which by far the largest. This shows the global normalization is the worst method for the data sets. Indeed, the corresponding 800 P-values are all zero, which shows the power of validation test is 100%. The LOWESS normalization using all blocks ignores the small block effects $\boldsymbol{\beta}$. The resulting statistics T_3 from 200 simulations are smaller than those based on the global normalization, but are stochastically larger than those based on the SLIM normalization, which accounts for these small block effects. This shows that the LOWESS normalization ignoring block effect is not as effective as the SLIM, but is more effective than the global normalization. Figure 1 also depicts the distribution of the 800 P-values of the LOWESS and SLIM method. The P-values of SLIM follow nearly a uniform distribution, which indicates that systematically biases have been effectively removed. On the other hand, for a portion of arrays, the LOWESS normalization is inadequate. This demonstrates the power of the validation: even ignoring small block effects, the test statistic T_3 is able to detect these small systematic biases.

After demonstrating the power of the test statistic T_3, we now examine the accuracy of the null distributions $T_1, \ldots, T_4$ using SLIM as the method of normalization. Since the correct method is used in the normalization for the pseudo-data, the estimation errors come from two sources: estimation of block and intensity effect and estimation of genewise variance by (14). First of all, without normalization, all validation test statistics $T_1, \ldots, T_4$ report zeros P-values for all realizations. This indicates the sensitivity of these tests. When the SLIM is used to normalize the data, if the normalization method is effective and the null distributions are accurate, the P-values follow the uniform distribution on the interval [0, 1]. Figure 2 depicts the P-values based on the SLIM normalization. The distributions of P-values based on T_3 are reasonably uniform. This shows that both the normalization method is effective and the null distribution is accurate for the test statistic T_3. However, the distributions of T_2 and T_4 have large deviations from the uniform distribution, which indicates the estimation errors from the REML estimators (14) of genewise SDs. These deviations have greatly been mitigated by correcting the REML estimators to the unbiased estimators (12).

5.2 Application to human placenta data

A collection of human placenta cDNAs comprising 7042 clones was identified and used as the probe set for cDNA microarray

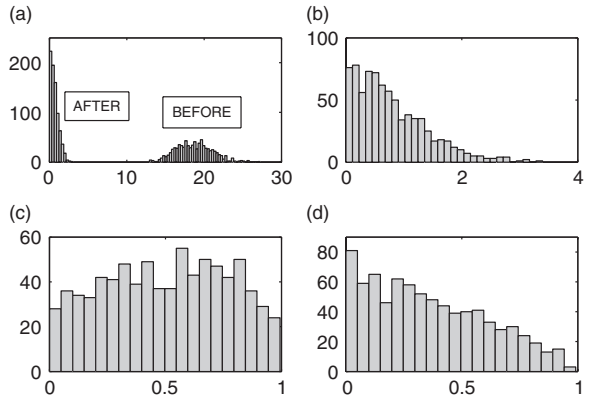

Fig. 1. (**a**) Distributions of $|T_3|$ before normalization (right) and after normalization using SLIM (left). (**b**) Distribution of $|T_3|$ after the LOWESS normalization without accounting small block effects. (**c**) and (**d**) Distributions of *P*-values of the validation test T_3 after normalization using SLIM (left panel) and LOWESS (right panel) methods. *P*-values based on test T_3 before normalization are all zero.

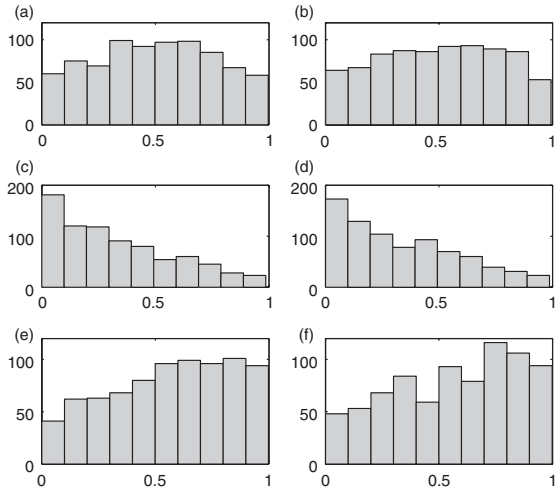

Fig. 2. Distributions of *P-value*s for test statistics T_1, T_2, T_3 and T_4 after the SLIM normalization. (**a**)–(**d**) Distributions of *P*-values for test statistics T_1, T_3, T_2 and T_4 after the SLIM normalization. (**e**) and (**f**) Distributions of *P*-values for test statistics T_2 and T_4 calculated by using unbiased estimators of genewise SD (13). The distribution of T_3 is somewhat more uniform than that based on T_1. The same conclusion applies to T_4 and T_2, both using the REML estimators and unbiased REML estimators.

fabrication in this study (Ma *et al*., 2005). Three kinds of RNA samples were used. These include the common reference RNA derived from the probe set (PS) in equal amounts representing artificial RNA produced by *in vitro* transcription, the 'Universal Human Reference RNA' from Stratagene, which is comprised of 10 different cell lines and human full-term placenta RNA. The original goal of the study was to evaluate the performance of the PS RNA as a reference RNA in comparison with that of Stratagene's universal reference RNA.

For the sake of studying the normalization and validation tests, we only compare the 'Universal Human Reference' RNA with human placenta RNA in this study. Gene expression values were obtained through direct hybridizations between these two kinds of RNAs. There are four slides, including two dye-swapped slides. Each clone was printed three times on different blocks in each slide. There are 48 blocks on each array. After preprocessing that filters low quality spots, there remained 2149 genes that have three replications. Our analysis focuses on this subset. We compared the effectiveness of four normalization methods: Global, LOWESS, SLIM and aggregated SLIM by using test statistics T_3 and T_4. One hundred different genes were selected as the validation set, while the remaining 2049 genes were used to estimate the parameters.

The key assumption of the LOWESS method is that up- and down-regulated genes' expressions are symmetrically distributed around 0, which is usually not the case for genes investigated in placenta tissue (Ma *et al*., 2005). Figure 3 (a)–(d) compares the effectiveness of the four normalization methods using the validation test statistics T_3 and T_4. The genewise variances are estimated by (14). The results show clearly that normalizations are needed for each array. In addition, the blockwise LOWESS normalization is inadequate except the fourth array at significant level 5%. The outcomes of validation tests show that we can choose either SLIM or aggregated SLIM to normalize the log-ratios for each array.

5.3 Application to interferons data using customized arrays

An important property of interferons (IFNs), a cytokine, is their anti-tumor activity. IFNs have efficacy in the treatment of several types of solid tumors carcinomas. Interestingly, it has been reported that IFN-β has greater anti-tumor effects than IFN-α on melanoma. One probable mechanism for the different effects may be the different affinity of IFN-α and IFN-β binding to the IFN receptors. Another possibility may be the differences in intracellular signaling. To address whether different signal pathways are involved in IFN-α and IFN-β-mediated anti-tumor activity, customized c-DNA chips are designed to include IFN stimulated genes and genes involved in multiple pathways. Gene expression changes induced by IFN-α and IFN-β were investigated and compared by using the customized c-DNA microarrays.

The customized c-DNA microarrays provides an ideal platform for our validation tests. Usually, only several hundreds of genes of primary interest are monitored for the changes of expression profiles and these genes are often duplicated several times. For our particular applications, 768 genes that might be induced by IFN are printed on the 16 blocks with 12 rows and 12 columns of spots in each block. These 768 genes are duplicated 3 times. The three replications of each gene reside in the same row in the same block but adjacent columns. The expression profiles of these 768 genes with IFN-α or IFN-β stimulations are compared with those without stimulations. This results in two customized microarrays, one with INF-α stimulation and the other with INF-β stimulation, each having with 3×768 spots. After some preprocessing that filtered the data with low quality, there were 572 genes left for further analysis.

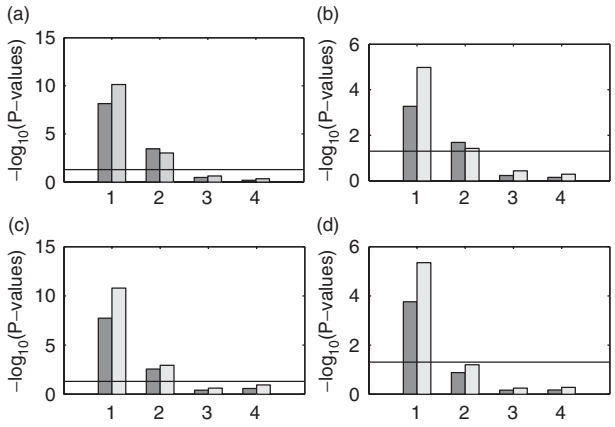

Fig. 3. The *P*-values for validation tests T_3 and T_4, based on the human placenta data, under four different normalization approaches: Global (labeled '1'), LOWESS (labeled '2'), SLIM (labeled '3'), and aggregated SLIM (labeled '4') methods. The *y*-axis is $-\log_{10}(P\text{-value})$. The dark gray columns are *P*-values for T_3, while the light gray ones are for T_4. The results for four arrays are plotted in (**a**) – (**d**). The lines correspond to 5% significance level. The *P*-values for Global normalization are highly statistically significant at level 5% for all four arrays, indicating the ineffectiveness of the method. The LOWESS normalization method improves the *P*-values a lot but still are statistically significant at level 5% except the fourth array. The *P*-values based on SLIM and aggregated SLIM methods show that the systematic biases due to the variation of experimental conditions have been effectively removed by either of these two methods. Therefore, the resulting log-ratios can be used for the downstream statistical analysis.

To examine the necessity and effectiveness of normalization, we randomly selected 50 genes as the test set; the remaining 522 genes were used as the learning set. Three normalization methods are employed: Global, LOWESS and SLIM normalizations. Due to the specific designs of the replications, the block effect is not estimable. The genewise variance are estimated by using the empirical Bayes method in Section 4.3—the variance estimates from two arrays are not aggregated. Table 1 depicts the *P*-values using the test statistics $T_1, \ldots, T_4$. From the table, we conclude that there is no need for normalization since all of the *P*-values for test statistics are high for both slides and for all methods. Note that the *P*-values by using the SLIM normalization are a little bit higher than those using the Global normalization, and they are larger than the *P*-values using the LOWESS method. That is because the fundamental assumption for LOWESS method—up- and down- regulated genes are about the same—are not substantiated here.

5.4 Application to human total RNA samples using Agilent arrays

Our third example comes from the Microarray Quality Control (MAQC) project (Patterson *et al.*, 2006). The MAQC project studies the reproducibility, sensitivity and specificity of the microarray data across different platforms and sites. It compared two RNA samples, Stratagene Universal Human Reference total RNA and Ambion Human Brain Reference total RNA using different microarray technology. Our study focuses only on the RNA samples generated at three test sites using Agilent platform. At each site, 10 Agilent two-color microarrays were processed with five arrays for each dye configuration, which assayed a total of 30 microarrays. Following Patterson *et al.* (2006), we excluded two outlier microarrays based on single microarray quality metrics, resulting in 28 microarrays. After preprocessing, we obtained 21 767 genes from a total of 43 931 and found four genes each having 10 replications, randomly located on the microarray.

Table 1. *P*-values for $T_1, \ldots, T_4$ based on human placenta data

	Statistics	*P*-values		
		Global	LOWESS	SLIM
Slide 1	T_1	0.3294	0.2857	0.3333
	T_2	0.3979	0.3139	0.4044
	T_3	0.5431	0.5155	0.5450
	T_4	0.3947	0.3097	0.4004
Slide 2	T_1	0.4704	0.4966	0.4730
	T_2	0.6952	0.6615	0.6964
	T_3	0.3012	0.2261	0.3009
	T_4	0.4322	0.3913	0.4318

For our validation test, gProcessedSignal and rProcessedSignal values from Agilent's Feature Extraction software were used as input to calculate the test statistics. If the genewise variance is estimated by using (14), all *P*-values are at least 99.99%, indicating the proper normalization of all Agilent arrays. Even using the most stringent estimation of genewise variance (13), almost all the microarrays pass the validation test at significant level 1% except for one array AGL-3-D3 at the test site 3. See Table 2. The results show the data processed by Agilent software are properly normalized and reliable. These are in line with the conclusion of the MAQC project.

6 CONCLUSION

Motivated by the urge of measures for comparing different normalization methods, we proposed four validation tests to evaluate the necessity and effectiveness of normalization methods, relying on the replicated clones. They are based on the standardized differences of expression profiles among replicated clones aggregated over different genes, resulting in the test statistics T_1 and T_2. These tests depend on genewise variances and SDs and hence cannot be used to compare the effectiveness of the normalization without estimation of these genewise variances. This leads us to consider the (unweighted) differences of expression profiles among replicated clones aggregated over different genes, standardized after aggregation, resulting in the test statistics T_3 and T_4. The unscaled test statistics T_3 and T_4 can be used to compare the effectiveness of normalization without estimating genewise variance (which itself depends on selecting a method of normalization). The aggregated standardization tests T_3 and T_4 depend on the aggregated SD and variance, which can be more precisely

Table 2. *P*-values based on $T_1, \ldots, T_4$ for MAQC project data

Slide name	*P*-values			
	T_1	T_2	T_3	T_4
AGL-1-C1	1.0000	1.0000	0.9993	0.9999
AGL-1-C2	1.0000	1.0000	0.9989	0.9996
AGL-1-C3	1.0000	1.0000	0.9982	0.9997
AGL-1-C4	1.0000	1.0000	0.9992	1.0000
AGL-1-C5	0.9998	1.0000	0.9962	0.9957
AGL-1-D2	0.2706	0.9948	0.2965	0.3792
AGL-1-D3	0.6940	0.9993	0.6931	0.7336
AGL-1-D4	0.5144	0.9981	0.5334	0.5785
AGL-1-D5	0.9894	1.0000	0.9659	0.9993
AGL-2-C1	0.7041	0.9977	0.6903	0.5178
AGL-2-C2	0.9479	0.9999	0.9247	0.9396
AGL-2-C3	0.9794	1.0000	0.9632	0.9919
AGL-2-C4	0.4056	0.9932	0.3823	0.2779
AGL-2-D1	0.0848	0.9502	0.0821	0.0562
AGL-2-D2	0.2218	0.9688	0.2497	0.1067
AGL-2-D3	0.3557	0.9889	0.3980	0.2688
AGL-2-D4	0.0177	0.6857	0.0109	0.0004
AGL-2-D5	0.1242	0.9551	0.1272	0.0662
AGL-3-C1	0.0570	0.8793	0.0451	0.0085
AGL-3-C2	0.0389	0.7861	0.0291	0.0019
AGL-3-C3	0.1024	0.9474	0.1033	0.0505
AGL-3-C4	0.1203	0.9092	0.1276	0.0215
AGL-3-C5	0.1328	0.9393	0.1337	0.0366
AGL-3-D1	0.2748	0.9645	0.3138	0.0888
AGL-3-D2	0.1372	0.8813	0.1371	0.0094
AGL-3-D3	**0.0034**	0.5220	**0.0007**	**0.0000**
AGL-3-D4	0.1907	0.9694	0.2071	0.1004
AGL-3-D5	0.0388	0.8792	0.0348	0.0115

estimated than the genewise SD and variance. As a result, the null distribution of the test statistics T_3 and T_4 can be more accurately approximated.

We have also demonstrated that the within-array replications are essential for estimating the genewise variance. A novel non-parametric approach is proposed, which aggregates variance information from genes with similar intensity level. Several innovative approaches are proposed to enhance the accuracy of the estimation of the genewise variance. These new methods can also be used to improve the power of selecting significantly differently expressed genes (Cui *et al*., 2005; Smyth *et al*., 2005).

Our simulation studies show convincingly the power of the validation tests and their validity. The applications to three real data sets demonstrate the methodological power of our proprietary methods in choosing a normalization method and in assessing whether the systematic biases due to variations in the experimental conditions have been properly removed. The customized array provides an ideal platform for the applications of our proposed approaches. Furthermore, our idea does not restrain to the two-color arrays. As long as replicated genes exist, we could apply our validation test to check the necessity and effectiveness of normalization. The validation tests could have been reliably applied the one-color Affymetrix GeneChip arrays, if they were within-array replications. Without such replications, we need to aggregate information across arrays. Assuming the absence of the gene and array interactions, our validation tests can be applied to the Affymetrix array to check the necessity and effectiveness of normalization, by treating each array as a block in a synthetic super-array.

ACKNOWLEDGEMENTS

The authors thank Dr Yi Ren of Rutgers University for printing the customized arrays used in the study described in the article. This research was supported by NIH Grant R01-GM072611, NSF Grant DMS-0354223 and DMS-0714554.

Conflict of Interest: none declared.

REFERENCES

Cui,X. *et al*. (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59–75.

Dabney,A.R. and Storey,J. (2007a) A new approach to intensity-dependent normalization of two-channel microarrays. *Biostatistics*, **8**, 128–139.

Dabney,A.R. and Storey,J. (2007b) Normalization of two-channel microarrays accounting for experimental design and intensity-dependent relationships. *Genome Biol*., 8, R44.

Dudoit,S. *et al*. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc*., **97**, 77–87.

Dudoit,S. *et al*. (2003) Multiple hypothesis testing in microarray experiments. *Stat. Sci*., **18**, 71–103.

Eisenstein,M. (2006) Microarrays: quality control. *Nature*, **442**, 1067–1070.

Fan,J. and Ren,Y. (2006) Statistical analysis of DNA microarray data in cancer research. *Clin. Cancer Res*., **12**, 4469–4473.

Fan,J. *et al*. (2005) Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency. *J. Am. Stat*., (with discussion), **100**, 781–813.

Fan,J. *et al*. (2004) Normalization and analysis of cDNA micro-arrays using within-array replications applied to neuroblastoma cell response to a cytokine. *Proc. Natl Acad. Sci. USA*, **101**, 1135–1140.

Gurland,J. and Tripathi,R.C. (1971) A simple approximation for unbiased estimation of the standard deviation. *Am. Stat*., **25**, 30–32.

Huang,J. *et al*. (2005) A two-way semi-linear model for normalization and significant analysis of cDNA microarray data. *J. Am. Stat. Assoc*., **100**, 814–829.

Jaeger,J. and Spang,R. (2006) Selecting normalization genes for small diagnostic microarrays. *BMC Bioinformatics*, **7**,388.

Kroll,T.C. and Wölfl,S. (2002) Ranking: a closer look on globalisation methods for normalisation of gene expression arrays. *Nucleic Acids Res*., **30**, 1–6.

Ma,S. *et al*. (2006) Robust semiparametric microarray normalization and significance analysis. *Biometrics*, **62**, 555–561.

Marshall,E. (2004) Getting the noise out of gene arrays. *Science*, **306**, 630–631.

Patterson,T. *et al*. (2006) Performance comparison of one-color and two-color platforms within the MicroArray Qualtiy Control (MAQC) project. *Nat. Biotechnol*., **24**, 1140–1150.

Reiner,A. *et al*. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.

Shi,L. *et al*. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol*., **24**, 1151–1161.

Smyth,G. *et al*. (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, **21**, 2067–2075.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Tseng,G.C. *et al*. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res*., **29**, 2549–2557.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Wang,D. *et al*. (2005) A robust two-way semi-linear model for normalization of cDNA microarray data. *BMC Bioinformatics*, **6**,14.