Journal of Biopharmaceutical Statistics, 18: 808–826, 2008

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400802277785

Taylor & Francis
Taylor & Francis Group

A PERMUTATION TWO ONE-SIDED TESTS PROCEDURE TO DETECT MINIMAL FOLD CHANGES OF GENE EXPRESSION LEVELS

Jen-pei Liu[1], Chen-Tuo Liao[2], Shih-Ting Chiu[2], and Jia-yan Dai[2]
[1]*Division of Biometry, Graduate Institute of Agronomy, National Taiwan University, Taipei, Taiwan and Division of Biostatistics and Bioinformatics, National Health Research Institutes, Taipei, Taiwan*
[2]*Division of Biometry, Graduate Institute of Agronomy, National Taiwan University, Taipei, Taiwan*

Current approaches to identifying differentially expressed genes are based either on the fold changes or on the traditional hypotheses of equality. However, the fold changes do not take into consideration the variation in estimation of the average expression. In addition, the use of fold changes is not in the frame of hypothesis testing, and hence the probability associated with errors for decision making of identification of differentially expressed genes cannot be quantified and evaluated. On the other hand, the traditional hypothesis of equality fails to take into consideration the magnitudes of the biologically meaningful fold changes that truly differentiate the genes between populations. Because of the large number of genes tested and small number of samples available for microarray experiments, the false positive rate for differentially expressed genes is quite high and requires further multiplicity adjustments, or use of an arbitrary cutoff for the p-values. However, all these adjustments do not have any biological justification. Hence, based on the interval hypothesis, Liu and Chow proposed a two one-sided tests procedure by consideration of both the minimal biologically meaningful fold changes and statistical significance simultaneously. To incorporate the correlation structure of expression levels among different genes and possible violation of normality assumption, we propose to apply a permutation method to the two one-sided tests procedure. A simulation study is conducted to empirically compare the type I error rate and power of the procedures based on the traditional hypothesis and the proposed permutation two one-sided tests procedure based on the interval hypothesis under various combinations of fold changes, variability, and sample sizes. Simulation results show that the proposed permutation two one-sided tests procedure based on the interval hypothesis not only can control the type I error rate at the nominal level but also provide adequate power to detect differentially expressed genes. Numerical data from public domains illustrate the proposed methods.

Key Words: Fold change; Interval hypothesis; Multivariate permutation methods; Power; Two one-sided tests procedure; Type I error.

Received September 9, 2007; Accepted March 5, 2008
Address correspondence to Jen-pei Liu, Division of Biometry, Graduate Institute of Agronomy, National Taiwan University, Taipei, Taiwan; E-mail: jpliu@ntu.edu.tw

INTRODUCTION

After completion of the Human Genome Project (HCP), the disease targets at the molecular level can be identified. Diagnostic accuracy for identification of molecular targets become increasingly important as more targeted clinical trials are conducted in the foreseeable future (Casciano and Woodcock, 2006; Dalton and Friend, 2006; Maitournam and Simon, 2005; Simon and Maitournam, 2004; Varmus, 2006). For example, the Trial Assigning IndividuaLized Options for tReatment (TAILORx) sponsored the U.S. National Cancer Institute (NCI) compares a combined regimen of adjuvant chemotherapy and hormonal therapy to the adjuvant hormonal therapy alone for the patients with a recurrence score of 11 to 25 as determined by a 21-gene Oncotype DX® breast cancer assay (Sprarano et al., 2006). On the other hand, using a 70-gene molecular signature, the patients with a low-risk molecular prognosis and a high-risk clinical prognosis were randomized to either the use of clinicopathologic criteria or gene signature in treatment decisions for the possible avoidance of chemotherapy in the MIcroarray In Node-negative Disease may Avoid ChemoTherapy Trial (MINDACT) (EORTC, 2006). Swain (2006) pointed out that these two trials have profound impact on the future personalized treatments for thousands of breast cancer patients. The MINDACT trials employed MammaPrint®, a 70-gene molecular signature derived from the microarray technology, which was approved by the U.S. Food and Drug Administration (FDA) in June 2007 (Van't Veer, 2002; Van de Vijver et al., 2002). On the other hand, the Oncotype DX® used in the TAILORx trial is a reverse transcriptase–polymerase chain reaction (RT-PCR) assay based on 21 genes (Paik et al., 2004, 2006). Both Oncotype DX® and MammaPrint® measure expression levels of the selected genes for prediction of recurrence of breast cancer. Therefore, gene expression levels can be thought as surrogates for phenotypes or indicators of clinical outcomes such as recurrence of breast cancer. Furthermore, they can also be employed for decision making on selection of treatments for patients or evaluation of effectiveness of the therapeutic regimens.

Although different technical platforms were employed in the diagnostic devices for molecular targets used in two trials, both assays are in fact the in vitro diagnostic multivariate index assays (IVDMIA; U.S. Food and Drug Administration, 2006) using the selected differentially expressed genes for identification of the patients with the molecular signatures. In general, for reduction of variability and assurance of quality of the expression levels, the IVDMIAs do not usually use all genes during the development stage. As a result, methods for selection of the differentially expressed genes to differentiate the patients with the molecular targets to those without the targets play a crucial role on the accuracy and reliability of the devices for diagnosis of the patients with the molecular targets.

Currently, the observed fold changes and the statistical procedures based on hypothesis of equality are the most frequently employed approaches to identifying the differentially expressed genes. For gene i, the observed fold change is the ratio of observed average expression levels of the gene under one condition, say tested or patients with the molecular targets, to that under another condition, say controlled or normal subjects without the targets, $i = 1, \ldots, G$. Gene i is identified as differentially expressed if the observed fold change either exceeds a prespecified upper threshold for overexpression, say, C_i or is below a predetermined

lower threshold for underexpression, say C_i', $i = 1, \ldots, G$. We call this method "the fixed fold-change rule." The fixed fold-change rule does not consider variability associated with the observed average expression level. In addition, it is not based on statistical hypotheses. Therefore the type I error rate and power for identifying the differentially expressed genes cannot be quantified. On the other hand, most current available statistical methods, including the t-test, permutation t-test, or significance analysis of microarray (SAM), are in fact based on the traditional hypothesis of equality (Dudoit et al., 2002; Simon et al., 2003; Tusher et al., 2001; Wang and Ethier, 2004):

$$H_o : \mu_{iT} - \mu_{iC} = 0 \quad \text{vs.} \quad H_a : \mu_{iT} - \mu_{iC} \neq 0, \;\; i = 1, \ldots, G; \tag{1}$$

where μ_{iT} and μ_{iC}, respectively, are the true unknown average expression levels on the log-scale (base 2) of gene i of the patients with the molecular targets and the normal subjects without the molecular targets; $i = 1, \ldots, G$; and G is the number of total genes under investigation.

However, the objective of the hypothesis of equality is only to detect whether the difference in the average expression levels between the tested and controlled conditions is not 0. Hence, hypothesis of equality does not take the magnitudes of the biologically meaningful fold changes into account. In addition, under the hypothesis of equality, the false positive rate for identifying differentially expressed genes is extremely high because of simultaneously testing tens of thousand of genes at the same time with a small number of samples. Therefore numerous methods were attempted to resolve this issue. They are applications of multiple comparison procedures to use some arbitrarily selected stringent cutoff of p-values to control false positive or discovery rate (Benjamini and Hochberg, 1995; Hochberg and Tamhane, 1987). However, all these methods do not simultaneously consider both magnitudes of biologically meaningful fold changes and statistical significance.

To resolve this dilemma, Liu and Chow (2008) proposed to formulate the hypothesis for identifying the differentially expressed genes as the interval hypothesis by simultaneously taking both the minimal biologically meaningful fold changes and the statistical significance into considerations. Based on the interval hypothesis, a two one-sided tests (TOST) procedure is also suggested by Liu and Chow (2008). The average expression levels of the differentially expressed genes between the tested and controlled groups identified by the TOST procedure are not only statistically significant at the prespecified nominal significance level but also exceed the minimal biologically meaningful fold change. However, the performance of the size and power of their proposed TOST procedure was not investigated. In addition, the structures of correlations of expression levels among different genes were not incorporated into their proposed TOST procedure. Therefore we proposed to apply the permutation method proposed by Simon et al. (2003) to the TOST procedure. In addition, a simulation is conducted to empirically investigate the size and power of the current approaches and our proposed permutation two one-sided tests (PTOST) procedure based on the interval hypothesis. The formulation of minimal biologically meaningful fold changes into the interval hypothesis for identifying the differentially expression genes proposed by Liu and Chow (2008) is reviewed in the Interval Hypothesis section. A PTOST procedure is proposed for testing the interval hypothesis is presented in the Simulation Study section.

Simulation results are presented in the Simulation Study section. In the Example section, a numerical example using a published data set illustrates the proposed method. In the last section are discussion and final remarks.

INTERVAL HYPOTHESIS

Let Y_{ijk} be the normalized log-transformed (base 2) intensity for gene i on array j receiving treatment k, $i = 1, \ldots, G$; $j = 1, \ldots, n_{ik}$; $k = T, C$. Define the gene-specific sample mean and sample variance for treatment k, respectively, as

$$\overline{Y}_{ik} = \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} Y_{ijk} \quad \text{and} \quad s_{ik}^2 = \frac{1}{n_{ik}} \sum_{j=1}^{n_{ik}} \left(Y_{ijk} - \overline{Y}_{ik}\right)^2$$

where n_{ik} is the number of arrays for treatment k; $i = 1, \ldots, G$; $k = T, C$.

In microarray experiments, tens of thousands of genes are compared simultaneously. If each gene is tested at the same significance level, it could lead to a very high probability of declaring a large number of unexpressed genes differentially expressed. This is a typical example of the multiple comparison problems. One simple common approach to resolving this issue is to use the stringent p-value such as the Bonferroni adjustment. The Bonferroni adjusted p-value is given by $\tilde{p}_i = \min(gp_i, 1)$, $i = 1, \ldots, G$, where p_i is the two-sided p-value based on the unpaired t-statistic for gene i. The unpaired two-sample t-test with the Bonferroni adjustment based on the traditional hypothesis of equality declares that gene i is differentially expressed at the α significance level if $\tilde{p}_i < \alpha$ or $pi < \alpha$. The Bonferroni approach controls the familywise experimental error rate but is conservative (Hochberg and Tamhane, 1987).

Another commonly employed method for identifying differentially expressed genes is the fixed fold-change rule. Gene i is claimed to be differentially expressed between test and control samples if $\overline{Y}_{iT} - \overline{Y}_{iC} > C_i$ or $\overline{Y}_{iT} - \overline{Y}_{iC} < -C_i'$, where C_i is the minimal biologically meaningful limit for overexpression, $-C_i'$ is the maximal biological meaningful limit for underexpression, and C_i and C_i' are positive. One frequently used requirement is the two-fold change when $C_i = C_i' = 1$; $i = 1, \ldots, G$. Recently, the MAQC (Microarray QC Metrics and Thresholds Project) Consortium (2006) suggests a combination of less stringent p-value for the traditional hypothesis of equality and fixed fold-change rule. In other words, gene i is declared to be differentially expressed if the null hypothesis of equality for gene i is rejected at the α significance level by the unpaired two-sample t-test and either $\overline{Y}_{iT} - \overline{Y}_{iC} > C_i$ or $\overline{Y}_{iT} - \overline{Y}_{iC} < -C_i'$, $i = 1, \ldots, G$. We refer it to as the combined fold-change rule. Note that the significance level used in the combined fold-change rule is referred to the comparisonwise error rate.

Current methods described above do not have any biological justification and fail to address both biological and statistical significance at the same time. Therefore, based on the normal assumption, Liu and Chow (2008) proposed to use the interval hypothesis by directly taking both the minimal biologically meaningful expression levels and statistical significance into the formulation of hypothesis. The objective is to identify the differentially expressed genes; therefore, the hypothesis for identifying differentially expressed genes should be formulated as the alternative

hypothesis. On the other hand, gene i is said to be differentially expressed if the differences in average expression levels between the test and control samples are greater (less) than some prespecified minimal (maximal) biologically meaningful expression levels. As a result, the hypothesis for identifying differential expressed genes between the test and control samples can be formulated as the interval hypothesis as follows:

$$\begin{aligned} H_0 &: -C_i' \le \mu_{iT} - \mu_{iC} \le C_i \quad \text{vs.} \\ H_1 &: \mu_{iT} - \mu_{iC} < -C_i' \quad \text{or} \quad \mu_{iT} - \mu_{iC} > C_i, \quad i = 1, \ldots, G. \end{aligned} \tag{2}$$

If $C_i = -C_i' = C$, the interval hypothesis for differentially expressed gene can be then formulated as the following two one-sided hypotheses:

$$\begin{gathered} H_{0U} : \mu_{iT} - \mu_{iC} \le C \quad \text{vs.} \quad H_{1U} : \mu_{iT} - \mu_{iC} > C \\ \text{or} \\ H_{0L} : \mu_{iT} - \mu_{iC} \ge -C \quad \text{vs.} \quad H_{1L} : \mu_{iT} - \mu_{iC} < -C, \quad i = 1, \ldots, G. \end{gathered} \tag{3}$$

Under the normal assumption, the two one-sided tests procedure proposed by Liu and Chow (2008) rejects the null hypothesis in (3) and gene i is declared as differentially expressed between the test and control samples at the α significance level if

$$t_{Ui} = \frac{\overline{Y}_{iT} - \overline{Y}_{iC} - C}{s_{pi} m_i} > t_{(\alpha/2, n_{iT}+n_{iC}-2)} \quad \text{or} \quad t_{Li} = \frac{\overline{Y}_{iT} - \overline{Y}_{iC} + C}{s_{pi} m_i} < -t_{(\alpha/2, n_{iT}+n_{iC}-2)}, \tag{4}$$

where $m_i = \sqrt{\frac{1}{n_{iT}} + \frac{1}{n_{iC}}}$, $t_{(\alpha/2, n_{iT}+n_{iC}-2)}$ is the $\alpha/2$ upper percentile of a central t-distribution with $n_{iT} + n_{iC} - 2$ degrees of freedom, s_{pi}^2 estimates the pooled within-sample variation of the gene expressions and is given as

$$s_{pi}^2 = \frac{(n_{iT} - 1)s_{iT}^2 + (n_{iC} - 1)s_{iC}^2}{n_{iT} + n_{iC} - 2}, \quad i = 1, \ldots, G.$$

Note that the significance level for the TOST procedure is referred to the comparisonwise error rate. On the other hand, the parameter space for the null hypothesis is an interval from $-C$ to C.

For the combined fold-change rule, gene i is declared to be differentially expressed at the α significance level if $|\bar{d}_i| > \max\{C, t_{(\alpha/2, n_{iT}+n_{iC}-2)} \times m_i \times s_{pi}\}$, where $\bar{d}_i = \overline{Y}_{iT} - \overline{Y}_{iC}$. From (4), the TOST procedure identifies gene i as differentially expressed if $|\bar{d}_i| > C + t_{(\alpha/2, n_{iT}+n_{iC}-2)} \times m_i \times s_{pi}$. It follows that at the same significance level the TOST procedure identifies fewer differentially expressed genes than the combined fold-change rule. However, the empirical size and power of fixed and combined fold-change rule, and their proposed TOST procedure, were not thoroughly investigated.

PERMUTATION TWO ONE-SIDED TESTS PROCEDURE

Although the TOST procedure proposed by Liu and Chow (2008) takes into consideration the minimal biological meaningful fold changes and statistical significance, it may be very conservative in identification of differentially expressed genes. On the other hand, the two t-statistics for the TOST procedure were based on the normality assumption, which is almost impossible to verify for the expression levels obtained from microarray experiments due to small number of arrays. In addition, the TOST procedure fails to take the correlation structure of expression levels among different genes into consideration.

For gene i, define

$$Y_{ijk}^{U} = \begin{cases} Y_{ijT}, & \text{if } k = T \\ Y_{ijC} + C, & \text{if } k = C;\ \ i = 1, \ldots, G,\ \ j = 1, \ldots, n_{ik}, \end{cases} \quad \text{and}$$
$$Y_{ijk}^{L} = \begin{cases} Y_{ijT}, & \text{if } k = T \\ Y_{ijC} - C, & \text{if } k = C;\ \ i = 1, \ldots, G,\ \ j = 1, \ldots, n_{ik}. \end{cases} \tag{5}$$

Then

$$E(Y_{ijk}^{U}) = \begin{cases} \mu_{ijT}, & \text{if } k = T \\ \mu_{ijC} + C, & \text{if } k = C;\ \ i = 1, \ldots, G,\ \ j = 1, \ldots, n_{ik}, \end{cases} \quad \text{and}$$
$$E(Y_{ijk}^{L}) = \begin{cases} \mu_{ijT}, & \text{if } k = T \\ \mu_{ijC} - C, & \text{if } k = C;\ \ i = 1, \ldots, G,\ \ j = 1, \ldots, n_{ik}. \end{cases} \tag{6}$$

Note that the interval hypothesis in (3) is equivalent to the following hypotheses:

$$\begin{aligned} &H_{0U} : \mu_{iT} \le (\mu_{iC} + C) \quad \text{vs.} \quad H_{1U} : \mu_{iT} > (\mu_{iC} + C) \quad \text{or} \\ &H_{0L} : \mu_{iT} \ge (\mu_{iC} - C) \quad \text{vs.} \quad H_{1L} : \mu_{iT} < (\mu_{iC} - C),\ \ i = 1, \ldots, G. \end{aligned} \tag{7}$$

Under null hypothesis H_{0U} and homogeneity of covariance matrices, $\{Y_{iTk}^{U}, i = 1, \ldots, G; j = 1, \ldots, n_{ik}\}$ and $\{Y_{iCk}^{U}, i = 1, \ldots, G; j = 1, \ldots, n_{ik}\}$ [or under H_{0L}, $\{Y_{iTk}^{L}, i = 1, \ldots, G; j = 1, \ldots, n_{ik}\}$ and $\{Y_{iCk}^{L}, i = 1, \ldots, G; j = 1, \ldots, n_{ik}\}$] are exchangeable. Hence, the permutation method proposed by Simon et al. (2003) can be applied to the TOST procedure based on the t-statistics given in (4).

Let B be the total number of all possible permutations for sample sizes of n_{iT} and n_{iC}, which is given as

$$B = \frac{(n_{iT} + n_{iC})!}{n_{iT}! n_{iC}!}.$$

The algorithm for the PTOST procedure is given below:

(i) Let p_{Ui} and p_{Li} be the p-values using central t-distributions obtained from the TOST procedure based on the samples $\{Y_{ijk}^{U}, i = 1, \ldots, G; k = T, C;$

$j = 1, \ldots, n_{ik}\}$ and $\{Y^L_{ijk}, i = 1, \ldots, G; k = T, C; j = 1, \ldots, n_{ik}\}$ of gene i for the null hypotheses H_{oU} and H_{oL}; respectively, $i = 1, \ldots, G$.

(ii) For each of B permutations, compute the p-values from the central t-distribution and denote them as p^b_{ui} and p^b_{Li}, $i = 1, \ldots, G$; $b = 1, \ldots, B$.

(iii) Sort p^b_{ui} and p^b_{Li} separately in ascending order to obtain the order statistics $p^b_{U(1)} \le p^b_{U(2)} \le \cdots \le p^b_{U(G)}$, and $p^b_{L(1)} \le p^b_{L(2)} \le \cdots \le p^b_{L(G)}$, respectively, $b = 1, \ldots, B$.

(iv) Then the two adjusted p-values obtained from the PTOST procedures for gene i, $i = 1, \ldots, G$; are given as

$$p^*_{Ui} = \{\text{number of permutation where } p^b_{U(1)} \le p_{Ui}\}/B \quad \text{and} \quad p^*_{Li} = \{\text{number of permutation where } p^b_{L(1)} \le p_{Li}\}/B. \tag{8}$$

(v) The null hypothesis in (7) is rejected and gene i is declared as differentially expressed between the test and control samples at the α significance level if

$$p^*_{Ui} \le \alpha/2 \quad \text{or} \quad p^*_{Li} \le \alpha/2, \quad i = 1, \ldots, G.$$

Note that since p-values and t-values have a one-to-one corresponding relationship, the above-mentioned PTOST procedure can also be conducted in terms of t-values. According to Simon et al. (2003), in addition to the advantages of consideration of correlation structure of expression levels among genes and less conservatism, the multivariate permutation method described above provides a reference distribution of $p^b_{U(1)}(p^b_{U(1)})$ for the observed p-values of the TOST procedure. Hence, the proposed PTOST procedure does not depend upon the normality assumption.

SIMULATION STUDY

A simulation study was conducted to investigate and compare performance of the empirical size and power between the four current approaches based on the hypothesis of equality, the TOST procedure and PTOST procedure based on the interval hypothesis for identification of differentially genes. Fortran 90 and IMSL STAT/LIBRARY Fortran subroutines are used in the simulation study. The data are generated according to the models suggested by Tsai et al. (2003). The first model (Model I) is the model for the background-subtracted intensities (without normalization), X_{ijk}; $i = 1, \ldots, G$; $j = 1, \ldots, n_{ik}$; $k = T, C$. In other words,

$$X_{ijk} = \tilde{\mu}_{ik} e^{\eta_{ijk}} + \varepsilon_{ijk}, \tag{9}$$

where $\tilde{\mu}_{ik}$ is the mean expression level, η_{ijk} is the multiplicative error, and ε_{ijk} is the additive error for gene i, array j, and treatment k, $i = 1, \ldots, G$ and $j = 1, \ldots, n_{ik}$, $k = T, C$.

For each gene, the two error components are assumed, respectively, to be independently and identically distributed respectively as bivariate normal distributions,

$$(\eta_{ijT}, \eta_{ijC})' \overset{\text{i.i.d.}}{\sim} N_2(\mathbf{0}, \mathbf{\Phi}_i) \quad \text{and} \quad (\varepsilon_{ijT}, \varepsilon_{ijC})' \overset{\text{i.i.d.}}{\sim} N_2(\mathbf{0}, \mathbf{\Sigma}_i)$$

where

$$\mathbf{\Phi}_i = \begin{bmatrix} \phi_{iT}^2 & \tau_i \phi_{iT} \phi_{iC} \\ \tau_i \phi_{iT} \phi_{iC} & \phi_{iC}^2 \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma}_i = \begin{bmatrix} \sigma_{iT}^2 & \rho_i \sigma_{iT} \sigma_{iC} \\ \rho_i \sigma_{iT} \sigma_{iC} & \sigma_{iC}^2 \end{bmatrix}, \quad i = 1, \ldots, G.$$

The second model (Model II) is for the logarithmic transformation of background-subtracted and normalized intensities, Y_{ijk}, which is assumed as

$$Y_{ijk} = \mu_{ik} + \eta_{ijk} + \varepsilon_{ijk}, \quad i = 1, \ldots, G; \quad j = 1, \ldots, n_{ik}; \quad k = T, C. \tag{10}$$

The distributional assumptions for the multiplicative error and the additive error are the same as the first model, respectively. Under the second model, the mean, variance, and covariance of Y_{ijk} are given, respectively, as

$$E(Y_{ijk}) = \mu_{ik}, \quad \text{Var}(Y_{ijk}) = \phi_{ik}^2 + \sigma_{ik}^2, \quad \text{and}$$
$$\text{Cov}(Y_{ijT}, Y_{ijC}) = \tau_i \phi_{iT} \phi_{iC} + \rho_i \sigma_{iT} \sigma_{iC}.$$

Because the biological samples of the tested and controlled conditions come from different populations, the tested and controlled samples are independent of each other. Consequently, in the simulation, we assume that $\tau_i = \tau = 0$, $\rho_i = \rho = 0$ for both models. In addition, we also assume that the homogeneity of variances and equal sample sizes such that $\phi_{iT}^2 = \phi_{iC}^2 = \phi^2$ and $\sigma_{iT}^2 = \sigma_{iC}^2 = \sigma^2$, and $n_{iT} = n_{iC} = n$. In the simulation we choose the number of genes (G) to be 500, 1000, and 2000; the number of samples (or replicated arrays) to be 4, 5, and 8; the difference in average expression levels between the tested and controlled samples to be 0, ± 0.5, ± 1.0, ± 2.0, and ± 3.0; the magnitudes of multiplicative error to be 0.1 and 0.3; and the values of additive error to be 0.5 and 1.0. Following Holy et al. (2002), the true average expression level for the controlled sample of each gene, μ_{iC}, is randomly drawn from a log-normal (base 2) distribution with a mean of 10 and a standard deviation of 1.2. Once μ_{iC} is generated, the true average expression level of the tested sample μ_{iT} is determined by adding the differences mentioned above. For each of 180 combinations, 1000 replicates of random samples are generated.

Since only a small portion of the genes in a microarray experiment are truly differentially expressed, to mimic the real situation, in the simulation, 90% of genes are not differentially expressed for which $\mu_{iT} - \mu_{iC} = 0$ and the other 10% of genes are differentially expressed such that $\mu_{iT} - \mu_{iC} \neq 0$. In addition, half of the differentially expressed genes are overexpressed ($\mu_{iT} - \mu_{iC} > 0$) and half are underexpressed ($\mu_{iT} - \mu_{iC} < 0$). Furthermore, $C = 1$ for the fixed and combined fold-change rules, and the TOST and PTOST procedures based on the interval hypothesis. Therefore, for the interval hypothesis as well as the fixed and combined fold-change rules, a gene is not expressed if $|\mu_{iT} - \mu_{iC}| \leq 1$. All methods used a nominal significance level of 5%. Three criteria are employed to examine the performance of different methods for identification of expressed genes. The overall type I error rate is defined as the proportion of at least one unexpressed gene being falsely identified as differentially expressed. Therefore, the overall type I error rate is the familywise type I error rate. The average type I error rate is defined as the average of the proportions of the number of falsely identified differentially expressed

genes among truly unexpressed genes over 1000 replicates. The average power is defined as the average of the proportions of the number of correctly identified differentially expressed genes among the truly expressed genes over 1000 replicates.

Due to a large number of combinations and similar performance between the two models, only the results for the number of genes being 1000 or 2000 and sample size being 5 or 8 under Model I are presented. The results for other combinations and those under Model II will be provided upon request. For the unpaired two-sample t-test and its Bonferroni adjustment, the parameter space of the null hypothesis consists only of a point at $\mu_{iT} - \mu_{iC} = 0$ while the parameter space of the null hypothesis for the fixed and combined fold change rules, and the TOST and PTOST procedures is an interval from -1 to 1. Therefore, for the unpaired two-sample t-test and its Bonferroni adjustment, the results of empirical overall and average type I error rates are provided at $\mu_{iT} - \mu_{iC} = 0$. On the other hand, for the fixed and combined fold-change rules, and the TOST and PTOST procedures, the results of empirical overall and average type I error rates are given at both $\mu_{iT} - \mu_{iC} = 0$ and ± 1.

Table 1 presents the empirical overall type I error rates under Model I. From Table 1, under Model I, the overall type I error rate for the unpaired two-sample t-test at $\mu_{iT} - \mu_{iC} = 0$ is almost 1 and that of its Bonferroni adjustment is around 0.05. This verifies that the unadjusted two-sample t-test fails to control familywise error rate. On the other hand, the empirical overall type I error rates of the fixed and combined change rules at $\mu_{iT} - \mu_{iC} = 0$ range from 0.0006 to 0.028 when the sample size is 8 and the variance of multiplicative error is 0.1. However, in the other situation, both methods cannot control the overall type I error rates, which range from 0.401 to 1. These results indicate that the performance of the fixed and combined change rules is very unstable at $\mu_{iT} - \mu_{iC} = 0$. They are very conservative when the sample size is large and the variance of multiplicative error is small. However, in other situations, they fail to control the overall type I error rate. On the other hand, the TOST procedure is very conservative at $\mu_{iT} - \mu_{iC} = 0$ with the overall type error being no larger than 0.02. However, for the PTOST, the empirical type I error rate ranges from 0.034 to 0.063 at $\mu_{iT} - \mu_{iC} = 0$. At $\mu_{iT} - \mu_{iC} = \pm 1$, the empirical overall type I error rates of the fixed and combined change rules, and the TOST and PTOST procedures, are all above 0.5. It appears that all four methods fail to control the familywise error rate at $\mu_{iT} - \mu_{iC} = \pm 1$. It should be noted that both TOST and PTOST procedures depend upon the minimal meaningful biological fold change, C. A smaller C will increase the familywise error rate.

Table 2 presents the empirical average type I error rates under Model I. From Table 2, at $\mu_{iT} - \mu_{iC} = 0$, the average error type I error rate of the unpaired two-sample t-test is around 0.05 whereas that of its Bonferroni adjustment is below 0.0002. For the fixed change and combined change rules, the average type I error rate at $\mu_{iT} - \mu_{iC} = 0$ ranges from 0.0005 to 0.0456 and from 0.0005 to 0.0229, respectively. It appears that the combined change rule seems to control the average type I error rate better than the fixed change rule at $\mu_{iT} - \mu_{iC} = 0$. On the other hand, the average type I error rate of the TOST and the PTOST procedures does not exceed 0.0008 at $\mu_{iT} - \mu_{iC} = 0$. Therefore the average type I error rate of the TOST and the PTOST procedures at $\mu_{iT} - \mu_{iC} = 0$ is comparable to that of the Bonferroni adjustment. At $\mu_{iT} - \mu_{iC} = \pm 1$, both the fixed and combined change rules have similar average type I error rates, ranging from 0.0498 to 0.909. The average type I

Table 1 Empirical overall type I error rates in identification of differentially expressed genes for Model I

					Methods					
G	n	ϕ^2	σ^2	$\mu_T - \mu_C$	1	2	3	4	5	6
1000	5	0.1	0.5	0	1.0000	0.0440	0.4080	0.4010	0.0000	0.0420
				±1	–	–	1.0000	1.0000	0.9050	1.0000
			1.0	0	1.0000	0.0570	0.3930	0.3850	0.0000	0.0410
				±1	–	–	1.0000	1.0000	0.9200	1.0000
		0.3	0.5	0	1.0000	0.0440	1.0000	1.0000	0.1400	0.0360
				±1	–	–	1.0000	1.0000	0.9340	0.5420
			1.0	0	1.0000	0.0500	1.0000	1.0000	0.1570	0.0400
				±1	–	–	1.0000	1.0000	0.9350	0.5530
	8	0.1	0.5	0	1.0000	0.0410	0.0006	0.0006	0.0000	0.0490
				±1	–	–	1.0000	1.0000	0.9180	1.0000
			1.0	0	1.0000	0.0570	0.0120	0.0120	0.0000	0.0540
				±1	–	–	1.0000	1.0000	0.9190	1.0000
		0.3	0.5	0	1.0000	0.0620	1.0000	1.0000	0.0170	0.0540
				±1	–	–	1.0000	1.0000	0.9260	0.9860
			1.0	0	1.0000	0.0540	1.0000	1.0000	0.0160	0.0630
				±1	–	–	1.0000	1.0000	0.9330	0.9830
2000	5	0.1	0.5	0	1.0000	0.0380	0.6350	0.6300	0.0010	0.0560
				±1	–	–	1.0000	1.0000	0.9940	1.0000
			1.0	0	1.0000	0.0380	0.6330	0.6250	0.0000	0.0340
				±1	–	–	1.0000	1.0000	0.9930	1.0000
		0.3	0.5	0	1.0000	0.0560	1.0000	1.0000	0.3090	0.0470
				±1	–	–	1.0000	1.0000	0.9940	0.5960
			1.0	0	1.0000	0.0440	1.0000	1.0000	0.2860	0.0420
				±1	–	–	1.0000	1.0000	0.9930	0.6250
	8	0.1	0.5	0	1.0000	0.0400	0.0280	0.0280	0.0000	0.0550
				±1	–	–	1.0000	1.0000	0.9940	1.0000
			1.0	0	1.0000	0.0420	0.0210	0.0210	0.0000	0.0460
				±1	–	–	1.0000	1.0000	0.9920	1.0000
		0.3	0.5	0	1.0000	0.0500	1.0000	1.0000	0.0150	0.0360
				±1	–	–	1.0000	1.0000	0.9950	0.9980
			1.0	0	1.0000	0.0510	1.0000	1.0000	0.0200	0.0350
				±1	–	–	1.0000	1.0000	0.9930	0.9950

G = the number of genes; n = number of samples; Methods: 1 = unpaired two-sample t-test; 2 = unpaired two-sample t-test with Bonferroni adjustment; 3 = fixed fold change; 4 = combined fold change; 5 = two one-sided tests procedure; 6 = permutation two one-sided test procedure.

error rate of the TOST and PTOST procedures is below 0.03. However, when the variance of the multiplicative error is small, the average type I error rate at $\mu_{iT} - \mu_{iC} = \pm 1$ of the TOST procedure is smaller than that of the PTOST procedure. On the other hand, the PTOST procedure appears to control the average type I error rate at $\mu_{iT} - \mu_{iC} = \pm 1$ better than the TOST procedure when the variance of the multiplicative error is large.

The above type I error rates are used for controlling no false positives. Sometimes, this type of criteria may be too stringent in identifying differentially expressed genes through a microarray experiment. Therefore, following the definition of the false discovery rate (FDR) by Benjamini and Hochberg (1995),

Table 2 Empirical average type I error rates in identification of differentially expressed genes for Model I

					Methods					
G	n	ϕ^2	σ^2	$\mu_T - \mu_C$	1	2	3	4	5	6
1000	5	0.1	0.5	0	0.0500	0.0000	0.0005	0.0005	0.0000	0.0000
				±1	–	–	0.0505	0.0500	0.0024	0.0106
			1.0	0	0.0504	0.0001	0.0005	0.0005	0.0000	0.0000
				±1	–	–	0.0504	0.0500	0.0025	0.0103
		0.3	0.5	0	0.0502	0.0000	0.0456	0.0229	0.0001	0.0000
				±1	–	–	0.0909	0.0548	0.0027	0.0008
			1.0	0	0.0497	0.0001	0.0454	0.0227	0.0002	0.0000
				±1	–	–	0.0903	0.0545	0.0026	0.0008
	8	0.1	0.5	0	0.0499	0.0000	0.0000	0.0000	0.0000	0.0000
				±1	–	–	0.0499	0.0499	0.0025	0.0295
			1.0	0	0.0501	0.0002	0.0000	0.0000	0.000	0.0001
				±1	–	–	0.0501	0.0501	0.0025	0.0289
		0.3	0.5	0	0.0499	0.0001	0.0113	0.0100	0.0000	0.0001
				±1	–	–	0.0604	0.0596	0.0024	0.0041
			1.0	0	0.0500	0.0001	0.0115	0.0102	0.0000	0.0000
				±1	–	–	0.0600	0.0566	0.0026	0.0041
2000	5	0.1	0.5	0	0.0499	0.0000	0.0005	0.0005	0.0000	0.0000
				±1	–	–	0.0503	0.0499	0.0025	0.0075
			1.0	0	0.0500	0.0000	0.0005	0.0005	0.0000	0.0000
				±1	–	–	0.0503	0.0498	0.0024	0.0075
		0.3	0.5	0	0.0501	0.0000	0.0455	0.0229	0.0002	0.0000
				±1	–	–	0.904	0.0544	0.0026	0.0004
			1.0	0	0.0501	0.0000	0.0455	0.0229	0.0002	0.0000
				±1	–	–	0.0909	0.0547	0.0027	0.0005
	8	0.1	0.5	0	0.0499	0.0000	0.0000	0.0000	0.0000	0.0000
				±1	–	–	0.0500	0.0500	0.0025	0.0221
			1.0	0	0.0500	0.0000	0.0000	0.0000	0.0000	0.0000
				±1	–	–	0.0500	0.0500	0.0026	0.0219
		0.3	0.5	0	0.0500	0.0000	0.0114	0.0101	0.0000	0.0000
				±1	–	–	0.0603	0.0568	0.0025	0.0027
			1.0	0	0.0500	0.0000	0.0114	0.0102	0.0000	0.0000
				±1	–	–	0.0604	0.0569	0.0025	0.0028

G = the number of genes; n = number of samples; Methods: 1 = unpaired two-sample t-test; 2 = unpaired two-sample t-test with Bonferroni adjustment; 3 = fixed fold change; 4 = combined fold change; 5 = two one-sided tests procedure; 6 = permutation two one-sided test procedure.

we also did a small simulation study for exploring the performance of the six procedures on FDR. Table 3 presents the results for comparison of FDR among six different methods. In general, the FDR of the PTOST procedure is smaller than those of the unpaired t-test, fixed fold-change rule, or combined fold-change rule but larger than those of the TOST procedure and the Bonferroni adjustment. Therefore, for identification of differentially expressed genes, our limited simulation demonstrates that the proposed PTOST procedure is neither too liberal nor too conservative in controlling the FDR.

Table 4 presents the empirical average power of the six procedures under Model I when $|\mu_{iT} - \mu_{iC}|$ is 2 or 3. In general, the empirical average power

Table 3 Comparison of false discovery rates for the number of genes being 1000 and the number of samples being 5

			Methods					
ϕ^2	σ^2	$\mu_T - \mu_C$	1	2	3	4	5	6
0.3	0.5	0	1.0000	0.0500	1.0000	1.0000	0.1490	0.7450
		±0.5	0.7596	0.0370	0.7184	0.6921	0.1188	0.4289
		±1.0	0.5134	0.0392	0.4469	0.3728	0.0557	0.1015
		±2.0	0.3227	0.0095	0.2927	0.1808	0.0037	0.0061
		±3.0	0.3084	0.0022	0.2892	0.1700	0.0020	0.0024
0.3	1.0	0	1.0000	0.0500	1.0000	1.0000	0.1770	0.7350
		±0.5	0.7602	0.0340	0.7190	0.6909	0.1130	0.4364
		±1.0	0.5174	0.0382	0.4490	0.3754	0.0526	0.1049
		±2.0	0.3235	0.0079	0.2940	0.1825	0.0037	0.0062
		±3.0	0.3084	0.0021	0.2883	0.1691	0.0013	0.0019

Methods: 1 = unpaired two-sample t-test; 2 = unpaired two-sample t-test with Bonferroni adjustment; 3 = fixed fold change; 4 = combined fold change; 5 = two one-sided tests procedure; 6 = permutation two one-sided test procedure.

increases as the sample size increases or as the total variability decreases. Although the Bonferroni adjustment can control both the overall and the average type I error rates, it cannot provide sufficient average power. The average powers of the unpaired two-sample t-test and fixed and combined fold-change rules are larger than those of the TOST and PTOST procedures. However, this false impression of superior powers of these methods comes with the steep price of highly inflated overall and average type I error rates. On the other hand, from Table 4, when the sample size is 5 and total variance is large, the TOST procedure is more powerful than the PTOST procedure. This phenomenon may be because there are only 252 possible permutations when the number of arrays is 5 for both groups. However, when the number of arrays increases, the power of the PTOST procedure becomes quite competitive.

Figure 1 provides the empirical average power curve as a function of $|\mu_{iT} - \mu_{iC}|$ under Model I for $G = 1000$, $n = 5$, $\phi^2 = 0.1$, and $\sigma^2 = 0.5$. From Figure 1, although the empirical average power for the unpaired two-sample t-test is greater than those of the TOST procedure and its multivariate permutation version, its power is for the rejection of the null hypothesis of no difference and not for testing a biologically meaningful change at a prespecified significance level. In addition, the empirical average power curves of the fixed fold-change and combined fold-change rules overlap each other. This indicates that a further screening by a nonstringent p-value employed by the combined fold-change rule does not alter that much on the results of the differentially expressed genes by the fixed fold-change rule. On the other hand, when $|\mu_{iT} - \mu_{iC}|$ exceeds 2, the empirical average power curve of our proposed methods is quite competitive to other procedures. Therefore from our simulation results, the PTOST procedure not only guarantees that the fold change is greater than the prespecified biologically meaningful threshold but also provides adequate average power when the sample size is 8.

Table 4 Empirical average powers in identification of differentially expressed genes for Model I

					Methods					
G	n	ϕ^2	σ^2	$\mu_T - \mu_C$	1	2	3	4	5	6
1000	5	0.1	0.5	±2	1.0000	0.3901	0.9997	0.9997	0.8578	0.9187
				±3	1.0000	0.8930	1.0000	1.0000	1.0000	1.0000
			1.0	±2	0.9361	0.0423	0.9774	0.9135	0.4216	0.1631
				±3	0.9993	0.2359	0.9999	0.9988	0.9364	0.7196
		0.3	0.5	±2	1.0000	0.3892	0.9997	0.9997	0.8586	0.9197
				±3	1.0000	0.8964	1.0000	1.0000	1.0000	1.0000
			1.0	±2	0.9378	0.0426	0.9775	0.9146	0.4221	0.1629
				±3	0.9994	0.2363	1.0000	0.9989	0.9380	0.7184
	8	0.1	0.5	±2	1.0000	0.9808	1.0000	1.0000	0.9824	0.9922
				±3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			1.0	±2	0.9968	0.3501	0.9945	0.9927	0.6554	0.5883
				±3	1.0000	0.9050	1.0000	1.0000	0.9968	0.9849
		0.3	0.5	±2	1.0000	0.9818	1.0000	1.0000	0.9827	0.9881
				±3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			1.0	±2	0.9967	0.3478	0.9942	0.9924	0.6526	0.5856
				±3	1.0000	0.9064	1.0000	1.0000	0.9972	0.9855
2000	5	0.1	0.5	±2	1.0000	0.2796	0.9997	0.9997	0.8588	0.8934
				±3	1.0000	0.8073	1.0000	1.0000	1.0000	1.0000
			1.0	±2	0.9360	0.0244	0.9773	0.9135	0.4215	0.1129
				±3	0.9994	0.1561	1.0000	0.9988	0.9367	0.6235
		0.3	0.5	±2	1.0000	0.2758	0.9997	0.9997	0.8571	0.8948
				±3	1.0000	0.8080	1.0000	1.0000	1.0000	1.0000
			1.0	±2	0.9370	0.0243	0.9772	0.9144	0.4215	0.1145
				±3	0.9994	0.1571	1.0000	0.9989	0.9368	0.6313
	8	0.1	0.5	±2	1.0000	0.9611	1.0000	1.0000	0.9826	0.9690
				±3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			1.0	±2	1.0000	0.2635	0.9943	0.9925	0.6544	0.4545
				±3	1.0000	0.8452	1.0000	1.0000	0.9969	0.9338
		0.3	0.5	±2	1.0000	0.9603	1.0000	1.0000	0.9827	0.9709
				±3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
			1.0	±2	0.9968	0.2631	0.9942	0.9923	0.6534	0.4503
				±3	1.0000	0.8430	1.0000	1.0000	0.9968	0.9331

G = the number of genes; n = number of samples; A = Average; Methods: 1 = unpaired two-sample t-test; 2 = unpaired two-sample t-test with Bonferroni adjustment; 3 = fixed fold change; 4 = combined fold change; 5 = two one-sided tests procedure; 6 = permutation two one-sided test procedure.

EXAMPLE

The data set of Luo et al. (2001) is used to illustrate the proposed methods. It consists of normalized gene expression ratios obtained from a collection of 25 prostate tissue samples, comprised of 16 prostate cancers and 9 benign prostatic hyperplasia (BPH) specimens. The data are downloaded from http://research.nhgri.nih.gov/microarray/prostate/supplement/images/6500GeneListw =CRs&QSs.xls. The array platform is a spotted cDNA array containing probes from 6500 human cDNAs (representing 6112 unique genes). A common reference design is used for this series of experiments. Quality scores taking values in the range of zero (unreliable) to one (reliable) are provided for each ratio measurement. For the

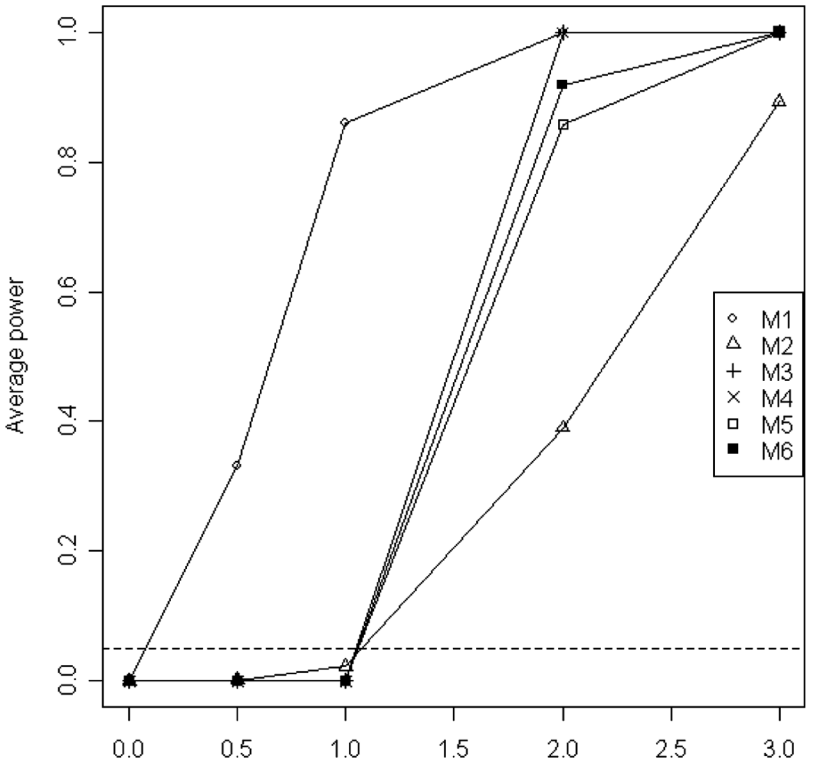

Figure 1 Average power curves of six methods in Model I for $G = 1000$, $n = 5$, $\psi^2 = 0.1$, $\sigma^2 = 0.5$. Methods: 1 = unpaired two-sample t-test; 2 = unpaired two-sample t-test with Bonferroni adjustment; 3 = fixed fold change; 4 = combined fold change; 5 = two one-sided tests procedure; 6 = permutation two one-sided test procedure.

example, only genes with quality scores greater than zero and occurring in at least 3 prostate cancer specimens and 3 BHP specimens are used. This reduces the number of analyzed clones to 5854. Log-transformation (base 2) is performed to the ratios for analysis.

For the purpose of illustration, we define that a gene is differentially expressed between the patients with prostatic cancer and with BPH if its true fold change is greater than 2. As a result, four current methods, the TOST procedures, and its multivariate permutation version with $C = 1$ ($\log 2$ scale) were applied to identify the genes that are differentially expressed between the prostate cancer and BPH samples. Table 5 provides a summary of results of the number of identified differentially expressed genes by the six procedures with the number of genes with a p-value smaller than 0.05 by the unpaired two-sample t-test or the number of genes with an observed fold changes greater than 2.

As pointed out by one of the guess editors, biologically meaningful fold changes are different from gene to gene. In addition, many important biologically significantly differentially expressed genes may have very subtle fold changes, which may be smaller than a two-fold change (Hughes et al., 2000). Therefore, it should be noted that a fold change of 2 used in this example is solely for the purpose of

Table 5 Summary of significant genes and genes with fold change greater than 2

Method	No. of identified genes	No. of genes with a p-value <0.05 by unpaired t-test	No. of genes with a fold change greater than 2
Unpaired t	1722	1722 (100.0%)	475 (27.5%)
Unpaired t (B)	4747 (100.0%)	31 (66.0%)	
Fixed fold change	597	475 (79.6%)	597 (100.0%)
Combined fold change	475	475 (100.0%)	475 (100.0%)
TOST	4747 (100.0%)	47 (100.0%)	
PTOST	181	180 (99.4%)	177 (97.8%)

TOST = Two one-sided tests procedure; PTOST = permutation two one-sided tests procedure.

illustration and does not imply that a gene with a fold change of less than 2 does not have biological meaning.

At the nominal significance level of 0.05, there is a total of 1722 genes (29.42%) declared as differentially expressed between the patients with prostatic cancer and the BPH patients by the unpaired two-sample t-test based on the hypothesis of equality. Among the 1722 declared differentially expressed genes, the observed fold changes of 1247 (72.4579%) genes are lower than 1 (log 2 scale). Because there are a total of 5854 genes, the Bonferroni adjustment is performed at 0.0000085 (0.05/5854) for each of 5854 genes. There are a total of 47 genes (0.8%) with the p-values from the unpaired two-sample t-test smaller than 0.0000085. However, even with an extremely stringent p-value of 0.0000085, among these 47 differentially expressed genes, the observed fold changes of 16 genes (34.04%) are still lower than 1. The observed fold changes of some identified differentially expressed gene can be as low as 0.5 by the Bonferroni adjustment.

The observed fold changes of 597 genes (10.20%) are larger than 2. Among these 597 genes, the p-values of 122 genes (20.44%) are larger than 0.05. In other words, although the observed fold changes of 597 genes exceed a minimal threshold of 1, due to variability, 20% of them do not reach statistical significance level at the 5% level. The results of the combined fold change method show that there is a total of 475 genes (8.49%) with p-values smaller than 0.05 and observed fold changes greater than 2. However, there are quite a few genes clustered in the neighborhood of the intersections between the vertical line of p-value being 0.05 and the horizontal line of the observed fold changes being ± 1. Most of these genes are not truly differentially expressed. This is because the p-values of the combined fold changes that are obtained from the unpaired two-sample t-test for testing the null hypothesis of no difference and the observed fold changes greater than 2 do not imply that the true parameter of fold change is greater than 2. Consequently, a considerable amount of genes are falsely identified as differentially expressed by the combined fold-change method.

The results of the TOST based on the interval hypothesis with a minimal threshold of 1 (log 2 base) demonstrate that at the 5% nominal significance level, there is a total of 47 genes (0.8%) with p-values calculated from the TOST procedure lower than 0.05. The minimal observed fold change with a p-value less than 0.05 obtained from the TOST procedure is above 2.83 (i.e., $1.5 \log 2$). A total of 181

genes were identified as differentially expressed by the PTOST procedure at the 5% significance level. Except for four genes, the minimal observed fold change with a p-value less than 0.05 obtained from the PTOST procedure is at least 2. There are several possible reasons why the observed fold changes of the 4 genes identified as differentially expressed are smaller than 2. One is that lots of expression levels of some genes are either missing or of poor quality in some array. Therefore, there is a total of 96 combinations of sample sizes for 5854 analyzed clones. Therefore, this makes the PTOST procedure very tedious to perform and the results of some combinations may become unreliable. Nonetheless, the 177 genes identified by the PTOST is much smaller than those of the unpaired t-test, fixed and combined fold-change rules, but is 4 times higher than those identified by the Bonferroni adjustment or by the TOST. Because the interval hypothesis directly takes into consideration the minimal threshold, these 177 differentially expressed genes identified by the PTOST not only possess fold changes with a magnitude of at least 2 but also take the variability of observed fold changes into account and reach statistical significance.

DISCUSSION

Correct identification of differentially expressed genes to characterize the molecular targets among different subpopulations of patients is the key to the success of individualized treatments (Dalton and Friend, 2006; Varmus, 2006). However, currently, different laboratories and different platforms employ various statistical methods to identify differentially expressed genes but yield inconsistent results (Tan et al., 2003). Only recently has this issue received its deserved attention (Dobbin et al., 2005; Irizarry et al., 2005; Larkin et al., 2005; Members of the Toxicogenomics Research Consortium, 2005). One of the objectives for the Microarray QC Metrics and Thresholds Project (MAQC, 2006; Shi et al., 2004) is to search for adequate and standard statistical methods in identification of gene signatures. Similar to any clinical or laboratory measurements such as levels of fasting glucose, sitting diastolic blood pressure, or aspartate aminotransferase (AST), a measurement of gene expression levels means nothing without biological implication and clinical interpretation. Therefore, if a gene is differentially expressed between two groups of patients, the difference in the average expression levels must exceed a minimal biologically meaningful threshold to have any clinical meaning and interpretations. Statistical methods for testing the null hypothesis of equality or the observed fold changes are the current approaches to identifying differentially expressed genes. Null hypothesis of equality completely ignores the biological meaning of gene expressions because one can always reach statistical significance by increasing the sample size. On the other hand, the method of observed fold changes overlooks the variability in estimation of fold changes and disregards the statistical significance for quantification of probability of errors in identification of differentially expressed genes. To resolve this dilemma, Liu and Chow (2008) proposed to directly formulate the minimal biologically meaningful fold change into the interval hypothesis. Similar to the equivalence or noninferiority margins in clinical trials the minimal biologically meaningful fold changes for the interval hypothesis in (3) should be determined jointly by biologists, clinicians, pharmacokineticists, biostatisticians, and other related

scientists. However, determination of the minimal biologically meaningful fold changes certainly deserves further research.

As mentioned before, the proposed PTOST procedure does not require the normality assumption and takes into account the correlation structure of gene expression levels among different genes. However, when missing gene expression levels occur differently in different genes at different arrays, the PTOST procedure becomes increasingly laborious to perform because permutations must be conducted separately for each missing pattern. In addition, if the sample size is small, the number of possible permutations is also small. This results in that the smallest possible p-value obtained from the reference distribution generated by permutation procedure is even larger than the nominal significance level. Consequently, sometimes, it is impossible or difficult to identify differentially expression of genes by the PTOST procedure if the sample size is smaller than or equal to 5. However, the behavior of the PTOST improved dramatically if sample size is greater than 5. However, when the sample size is large, the number of possible permutation becomes too large to perform the PTOST procedure. As a result, a random permutation becomes a feasibly attractive alternative. Our experience from simulation studies and examples indicate that a random permutation should have a size of at least 10,000 to generate a reliable reference distribution of the smallest p-value of G p-values under the null hypothesis. There are different definitions and methods for FDR. For example, Simon et al. (2003) and Korn et al. (2004) proposed the false discovery proportion and Li et al. (2005) and Reiner et al. (2003) suggested other FDR adjusted p-values. The possible extension of our proposed PTOST procedure to control the false discovery rate requires further research.

ACKNOWLEDGMENTS

We sincerely thank the two anonymous reviewers and one of the guest editors for their careful, thoughtful, and thorough review and comments, which greatly improved the content and presentation of our work. This research was partially supported by the Taiwan National Science Council Grant NSC95 2118-M-002-007-MY2 to Jen-pei Liu.

The views expressed in this article are personal opinions of the authors and may not represent the position of National Taiwan University and the National Health Research Institutes, Taiwan.

REFERENCES

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57:289–300.

Casciano, D. A., Woodcock, J. (2006). Empowering microarrays in the regulatory setting. *Nat. Biotechnol.* 24:1103.

Conover, W. J. (1980). *Practical Nonparametric Statistics* 2nd ed. New York: John Wiley.

Dalton, W. S., Friend, S. H. (2006). Cancer niomarkers—An invitation to the table. *Science* 312:1165–1168.

Dobbin, K. K., Beer, D.G., Meyerson, M., Yeatman, T. J., Gerald, W. L., Jacobson, J. W., Conley, B., Buetow, K. H., Heiskanen, M., Simon, R. M., Minna, J. D., Girard, L., Misek, D., Taylor, J. M. G., Hanash, S., Naoki, K., Hayes, D. N., Ladd-Acosta, C.,

Enkemann, S. A., Viale, A., Giordano, T. J. (2005). Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clin. Cancer Res*. 11:565–573.

Dudoit, S., Yang, Y. H., Callow, M. J., Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 12:111–139.

Hochberg, Y., Tamhane, A. C. (1987). *Multiple Comparison Procedures*. New York: Wiley.

Holy, D. C., Rattray, M., Jupp, R., Brass, A. (2002). Making sense of microarray data distributions. *Bioinformatics* 18:576–584.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell* 102:109–126.

Irizarry, K. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G. N., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S. C., Hoffman, E., Jedlicka, A. E., Kawasaki, E., Martínez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Qing Ye, S. Q., Yu, W. (2005). Multi-laboratory comparison of microarray platforms. *Nat. Methods* 2:345–349.

Korn, E. L., Troendle, J. F., McSahne, L. M., Simon, R. (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data. *Journal of Statistical Planning and Inference* 124:379–398.

Larkin, J. E., Frank, B. C., Gavras, H., Sultana, R., Quackenbush, J. (2005). Independence and reproducibility across microarray platforms. *Nat. Methods* 2:337–343.

Li, S. S., Bigler, J., Lampe, J. W., Potter, J. D., Feng, Z. (2005). FDA-controlling testing procedures and sampling determination for microarray. *Stat. Med.* 24:2267–2280.

Liu, J. P., Chow, S. C. (2008). Statistical issues on the diagnostic multivariate index assay for targeted clinical trials. *J. Biopharm. Stat.* 18:167–182.

Luo, J., Duggan, D. J., Chen, Y., Sauvageot, J., Ewing, C. M., Bittner, M. L., Trent, J. M., Isaaxs, W. B. (2001). Human prostate cancer and benign prostatic hyperplasia: Molecular dissection by gene expression profiling. *Cancer Res.* 61:4683–4688.

Maitournam, A., Simon, R. (2005). On the efficiency of targeted clinical trials. *Stat. Med*. 24:329–339.

MAQC Consortium. (2006). The MAQC project shows inter- and intra-platform reproducibility of gene expression measurements. *Nat. Biotechnol*. 24:1151–1161.

Members of the Toxicogenomics Research Consortium. (2005). Standardization of global gene expression analysis between laboratories and across platforms. *Nat. Methods* 2:351–356.

European Organization for Research and Treatment of Cancer (EORTC, 2006). MINDACT Design and MINDACT trial overview. http://www.breast internationalgroup.org/transbig.html (Last accessed June 5, 2006).

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., Hiller, W., Fisher, E. R., Wickerham, D. L., Bryant, J., Wolmark, N. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med*. 351:2817–2826.

Paik, S., Tang, G., Shak, S., Kim, C., Baker, J., Kim, W., Cronin, M., Baehner, F. L., Watson, D., Bryant, J., Costantino, J. P., Geyer Jr, E. C., Wickerham, D. L., Wolmark, N. (2006). Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J. Clin. Oncol*. 24:1–12.

Reiner, A., Yekutierli, D., Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19:368–375.

Shi, L., Tong, W., Goodsaid, F., Frueh, F. W., Fang, H. H., Tao, F. C., Casciano, D. A. (2004). QA/QC: Challenges and pitfalls facing the microarray community and regulatory agencies. *Expert Rev. Mol. Diagn.* 4:761–777.

Simon, R., Maitournam, A. (2004). Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin. Cancer Res.* 10:6759–6763.

Simon, R. M., Korn, E. L., McShane, L. M., Radmacher, M. D., Wright, G. W., Zhao, Y. (2003). *Design and Analysis of DNA Mircoarray Investigations*. New York: Springer.

Sprarano, J., Heyes, D., Dees, E., Olson, J., Perez, E., Pritchard, K., Oeyer, C. (2006). Phase III randomized study of adjuvant combination chemotherapy and hormonal therapy versus adjuvant hormonal therapy alone in women with previously resected axillary node-negative breast cancer with various levels of risk for recurrence (TAILORX Trial). http://www.cancer.gov/clinicaltrials/ECOG-PACCT-1. (Last accessed June 5, 2006).

Swain, S. M. (2006). A step in the right direction. *J. Clin. Oncol*. 24:1–2.

Tan, P. K., Downey, T. J., Spitznagel Jr, E. L., Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M., Cam, M. C. (2003). Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* 31:5676–5684.

Tsai, C. A, Chen, Y. J., Chen, J. J. (2003). Testing for differentially expressed genes with microarray data. *Nucl. Acids Res.* 31:e52.

Tusher, V. G., Tibshirani, R., Chu, G. (2001). Significance analysis of microarrays applied to ionizing radiation response. *Proc. Natl. Acad. Sci. USA* 98:5116–5121.

U.S. Food and Drug Administration. (2006). *Draft Guidance on In Vitro Diagnostic Multivariate Index Assays*. The U.S. Food and Drug Administration, Rockville, MD.

Van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voshuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., Berhards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* 347:1999–2009.

Van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415:530–536.

Varmus, H. (2006). The new era in cancer research. *Science* 312:1162–1165.

Wang, S., Ethier, S. (2004). A generalized likelihood ratio test to identify differentially expressed genes from microarray data. *Bioinformatics* 20:100–104.